Use of artificial intelligence (AI) in historical records transcription: Opportunities, challenges, and future directions

Yumeng Zhang

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of
Master of Science

Department of Geography
McGill University, Montreal
January 2023

Abstract

Historical weather documents are hidden treasures for understanding long-term climatic changes that were not fully exploited in the past due to limited resources. They are now attracting much attention from data rescue researchers who are interested in the valuable information the historical documents can bring. One obstacle of researching these historical documents is that the documents can only exist in hard copy format, making it challenging to access computational and geographically. The most common transcription approach is manual transcription. Manual transcription is labour-intensive and can be time-consuming when transcribing large numbers of documents, especially when they are written in cursive handwriting and in dense ledge sheets. If the transcription could be automated, considerable time and resources would be saved. The overarching question driving this research is what is the role and benefit of automation, especially artificial intelligence (AI)-augmented automation, in historical weather data rescue? To answer this question, I divide it into two complementary questions:

1. How do researchers and practitioners perceive the challenges and opportunities of using AI-augmented data rescue?

2. If AI-augmented data rescue is useful then what might an automated system look like?

I begin by reviewing the literature in historical weather data rescue and historical records transcription more generally. No systematic workflow exists to guide researchers in automating the transcription process step by step as AI augmented data transcription requires a sequence of models that demands different skills, and the researchers who would benefit the most usually lack relevant backgrounds. There also are risks and challenges; how researchers perceive these challenges are unknown. In Chapter 3, I conducted a survey on data rescue and citizen science researchers' attitude and opinion towards using AI-enhanced automated data rescue approaches instead of manual approaches. The result will provide an outlook of adapting AI in data rescue, and it also could help in overcoming the resistance along the way of automation. Respondents suggested a hybrid model where human and AI working together would be better than implementing a purely AI approach. In Chapter 4, I created and tested an AI-augmented data rescue workflow that can automate the transcription of a historical handwritten tabular dataset. The hope is this end-to-end workflow can achieve all steps to automation in one package and can

process raw tabular images, segment tabular cells, recognize the data entries, and rearrange the results into original formats. The proposed workflow is tested and evaluated on the Data Rescue: Archives and Weather (DRAW) dataset. The proposed workflow provides a guideline for researchers who want an end-to-end solution on automatic historical data rescue and it hopefully can be replicable for future projects that seek automated data rescue.

This thesis makes several contributions to the use of AI in historical weather data rescue and historical records transcription more generally. First, this study contributes by uncovering the opportunities and challenges of using AI in historical records transcription. Second, this research contributes a benchmark workflow for AI-augmented historical records transcription that can be customized and adapted to different types of historical records. Third, this study identifies and bridges the gap between the use of AI and the rescue and transcription of historical records, finding that addressing this multi-disciplinary problem requires efforts from different fields and communities. The results of this study will serve as a starting point for future studies who want to involve AI in their transcription process and as a reference for future attempts.

Résumé

Les documents météorologiques historiques sont des trésors cachés pour la compréhension des changements climatiques qui n'ont pas été pleinement exploités dans le passé en raison de ressources limitées. Ils attirent aujourd'hui l'attention des chercheurs qui s'intéressent aux informations précieuses que les documents historiques peuvent apporter. L'un des obstacles à la recherche de ces documents historiques est qu'ils n'existent que sous forme de copie papier, ce qui en rend l'accès difficile. La méthode de transcription la plus courante est la transcription manuelle. La transcription manuelle demande beaucoup de travail et peut prendre beaucoup de temps lorsqu'il s'agit de transcrire un grand nombre de documents. Si la transcription pouvait être automatisée, cela permettrait d'économiser beaucoup de temps et de ressources. La question primordiale qui sous-tend cette recherche est la suivante : quel est le rôle et les avantages de l'automatisation, en particulier de l'automatisation augmentée par l'intelligence artificielle (IA), dans le sauvetage des données météorologiques historiques ? Pour répondre à cette question, je la divise en deux questions complémentaires :

1. Comment les chercheurs et les praticiens perçoivent-ils les défis et les opportunités de l'utilisation de l'IA dans le sauvetage des données ?

2. Si le sauvetage des données par l'IA est utile, à quoi pourrait ressembler un système automatisé ?

Je commence par passer en revue la littérature sur le sauvetage des données météorologiques historiques et, plus généralement, sur la transcription des documents historiques. Il n'existe pas de flux de travail systématique pour guider les chercheurs dans l'automatisation du processus de transcription. Il existe également des risques et des défis ; la façon dont les chercheurs les perçoivent est inconnue. Au chapitre 3, j'ai mené une enquête sur l'attitude et l'opinion des chercheurs en sauvetage de données et en science citoyenne à l'égard de l'utilisation d'approches améliorées par l'IA plutôt que d'approches manuelles. Le résultat fournira une perspective de l'adaptation de l'IA dans le sauvetage des données. Les répondants ont suggéré qu'un modèle hybride où l'homme et l'IA travaillent ensemble serait préférable à une approche purement IA. Au chapitre 4, j'ai créé et testé un flux de travail de sauvetage de données enrichi par l'IA, qui peut automatiser la transcription d'un ensemble de données tabulaires manuscrites historiques.

L'espoir est que le flux de travail puisse réaliser toutes les étapes de l'automatisation en un seul paquet. Le flux de travail proposé est testé et évalué sur le Data Rescue : Archives and Weather (DRAW). Le flux de travail proposé fournit une ligne directrice pour une solution de bout en bout sur le sauvetage automatique des données historiques et, espérons-le, il peut être reproduit pour les projets qui cherchent une solution automatisée.

Cette thèse apporte plusieurs contributions à l'utilisation de l'IA dans le sauvetage des données météorologiques historiques et plus généralement dans la transcription des documents historiques. Premièrement, cette étude contribue à mettre en évidence les opportunités et les défis de l'utilisation de l'IA dans la transcription des documents historiques. Deuxièmement, cette recherche propose un flux de travail de référence pour la transcription de documents historiques enrichis par l'IA, qui peut être personnalisé et adapté à différents types de documents historiques. Troisièmement, cette étude identifie et comble le fossé entre l'utilisation de l'IA et le sauvetage et la transcription des documents historiques. Les résultats de cette étude serviront de point de départ pour les études futures qui souhaitent impliquer l'IA dans leur processus de transcription et de référence pour les tentatives futures.

Table of contents

List of Figures

List of Tables

Acknowledgements

I could not have undertaken this journey without the guidance of my supervisor, Dr. Renee Sieber, throughout my graduate study, and the encouragement and support she provided to me during these several years. I would like to express my sincere appreciation to my supervisory committee, Dr. Graham MacDonald, who has provided me invaluable ideas and feedback along the way. This endeavor would not have been possible without the generous support of Dr. Victoria Slonosky, Dr. Jin Xing, and everyone on the DRAW team who provided me with inspiration, ideas, feedback, and moral support. You make me feel like I am part of a loving family.

Special thanks to the 50 survey respondents who generously contributed to help me with my research and provided me with valuable answers and comments beyond what I could have asked for. I had the pleasure of working with everyone in the GeoThink lab and colleagues at McGill University, and I am grateful for their generous help and moral support. Thank you to all my friends for their encouragement and the precious memories we shared in Montreal.

Lastly, I would be remiss in not mentioning my family, especially my parents, who have supported and loved me unconditionally. I could not have gotten this far without their faith and encouragement.

Contribution of authors

Contributions to Chapter 1, 2, and 5:

I, Yumeng Zhang, wrote the introduction, literature review, and conclusion chapters. Dr. Renee Sieber provided comments, suggestions, and feedback during the idea conception phase and the writing process. Dr. Graham MacDonald contributed feedback and comments to the final version of the chapters.

Contribution to Chapter 3 and 4:

These chapters are peer-reviewed journal articles co-authored with my supervisor Dr. Renee Sieber. I am the primary author, and I designed and conducted the study, including designing the model and framework, implementing models, formulating questions, data collection, analysis, and writing. Dr. Renee Sieber helped shape the research and provided critical feedback throughout the research and writing process. My supervisory committee, Dr. Graham MacDonald, provided advice on the initial formulation of the research objectives and design and provided feedback on later drafts of this manuscript.

Chapter 1. Introduction

A vast amount of historical weather records is stored in libraries and archives around the world (Kwok, 2017). These records can be dated back to the nineteenth century or earlier. In recent decades, more attention has been paid to these records than ever before due to advances in technology. Thanks to new statistical models and advanced software, researchers are able to use these records to explore and achieve more results. For example, in climatology, Compo et al. (2011) and Slivinski et al. (2019) are able to take these individual records and combine them and reconstruct past weather patterns. These records exist in various formats and are vast in number. For example, some are dense ledger sheets; others may be diaries. The amount of existing historical records available far exceeds the number of experts in the field who may transcribe them for analysis. To analyze the records, they must be converted into a machine-encoded format so that they can be easily accessed and analyzed by a computer (World Meteorological Organization, 2016). Unfortunately, almost all of these historical records are stored in paper form, with a portion of them being scanned into images or photographs. This method of preservation makes it difficult for scientists to access or download these records remotely and, if it is stored in paper form in libraries and archives, they have to visit the physical location (Brunet & Jones, 2011). Even downloadable scanned copies of these records cannot be indexed, searched or analyzed without transcribing the individual observations. To address these difficulties, researchers started data rescue initiatives.

The concept of data rescue was first introduced in the 1990s and was originally termed “data archaeology and rescue” (Levitus, 1992, p. iii). It first appeared in the Global Oceanographic Data Archaeology and Rescue (GODAR) Project, where data archaeology stands for the recovery of obsolete computerized data into new media or formats, and rescue refers to efforts to rescue endangered data at risk of being lost of deterioration (Levitus, 1996). However, the term data rescue mentioned in this study is slightly different. It is an effort made by researchers to prepare historical records for analysis (Brönnimann et al., 2018). The phrase originates from individuals concerned that the paper documents may be too fragile to survive and that scanning alone would not transform the pages into analyzable data (Tan et al., 2004). While it also aims to preserve historical records from being lost or deteriorated, it focuses on the transcription and digitization of its contents.

Kwok (2017) pointed out the importance of data rescue to future research and how beneficial it would be in filling gaps of historical records. There are many past and ongoing data rescue-type initiatives in different fields (Fischer et al., 2014), although many of them are weather related (Brönnimann et al., 2019). The initiatives can expand beyond meteorology-related fields to a variety of transcription efforts. It is worth noting that the initiatives outside of meteorology-related fields are not considered rescues, but transcriptions. By investigating and improving weather data rescue, other fields may also benefit greatly on the transcription of historical records. I focus on historical weather data rescue (hereafter referred to as "data rescue").

The most common method of data rescue is manual keying, which involves hiring paid data entry personnel or recruiting volunteer transcriptionists (Brönnimann et al., 2006; World Meteorological Organization, 2016). Manual keying is a laborious process (Brunet et al., 2014; Camuffo & Bertolin, 2012); it is time consuming (e.g., Ashcroft et al., 2018) and can be expensive if the project hires paid transcriptionists (e.g., Dupigny-Giroux et al., 2007). As a consequence, researchers and practitioners are interested in automation, such as optical character recognition (OCR). Automation is valuable in our everyday lives as well as in data rescue. OCR has been used to handle license plates and tax receipts recognition and it is useful for climate change research if used to transcribe historical weather records (Chimani et al., 2021; Singh et al., 2012). If automation has benefits in our daily lives and in solving complex societal problems like climate change, it may also be useful in data rescue.

With the advancement of automation from artificial intelligence (AI) technology, questions have been raised whether AI could be a possible automated method for data rescue. The OCR used in past data rescue projects is proprietary software and is often not equipped with AI technology. The use of AI in data rescue is immature; it may bring new advantages, but equally, there may be many problems and conflicts (Wilkinson et al., 2019). Several studies have discussed the use of automation or AI in their projects and they believe that AI may be helpful and could be a future direction for data rescue (Chimani et al., 2021; World Meteorological Organization, 2016). However, researchers who have tested AI or just OCR in their projects report high error rates and difficulty in handling handwritten records and multiple formats (e.g., graphs, tables, text) (Brönnimann et al., 2006; Craig & Hawkins, 2020; Stickler, Alexander et al., 2014; Wilkinson et al., 2019; World Meteorological Organization, 2016). These failures are part of the reason why automated approaches have not been widely used for past data rescue projects. Concerns and

potential issues have been raised questioning whether AI would be helpful. Examples of concerns include the elimination of public participation for citizen science projects (i.e., volunteers may not be needed if AI can complete all tasks), funding availability, technical complexity, and accuracy (Blancq, 2010; Brönnimann et al., 2006; Craig & Hawkins, 2020; Stickler, Alexander et al., 2014; Wilkinson et al., 2019). One significant problem is the source documents, which may be of poor quality (e.g., degraded, improperly scanned) and may be handwritten in cursive text on which AI may not have been trained. These strengths and concerns prompted this study to investigate whether innovations in AI surmount these problems to improve data rescue.

The purpose of this study is to provide a guideline and direction for future data rescue or historical records transcription projects that wish to employ automation and to form a foundation to aid in the future transition from manual to automated. The overarching question driving this research is what is the role and benefit of automation, especially AI-augmented automation, in historical data rescue? To answer this question, I divided it into two complementary questions:

1. How do researchers and practitioners perceive the challenges and opportunities of using AI-augmented data rescue?

2. If AI-augmented data rescue is useful then what might an automated system look like?

The first question is examined in Chapter 3 by surveying 50 principal investigators and leading researchers of transcription-related data rescue projects. The AI-augmented data rescue here refers to an automated version of record transcription that utilizes AI techniques, rather than approaches that are done purely manually. I asked them about their attitudes, perceptions and opinions on transitioning their data rescue projects to AI-augmented automated approaches. Since these leaders and researchers are the ones who can determine the future direction of the transcription projects, asking their opinion is a good way to discover the challenges and opportunities of transitioning to automated transcription approaches. The challenges and opportunities for AI-augmented data rescue identified by survey respondents are relative to how manual approaches currently operate. It is best to address the concerns of end users so that barriers can be removed in advance and the transition can be smooth. Here, respondents were primarily interested in historical weather, but I also sought the opinions of citizen scientists who were interested in transcription more broadly.

The second question is addressed in Chapter 4 by establishing and testing an end-to-end historical records transcription workflow that utilizes AI techniques. The usefulness of transitioning to AI-augmented data rescue can be determined from Chapter 3 through respondents' view on aspects such as costs, hours required to set up, and accuracy. Building on the observations in Chapter 3, the automated data rescue workflow proposed in Chapter 4 can serve as a guideline for future research attempting to automate their transcription projects. This workflow does not focus on achieving state-of-the-art performance; instead, it aims to combine data rescue and AI techniques to provide a reference for future studies.

There is no existing literature that provides a systematic review of the use of AI technologies in historical weather data rescue and historical records transcription more generally. This leaves a gap to be filled. To investigate the application of AI in data rescue and transcription, this thesis is organized into five chapters. Chapter 1 introduces the rationale for this study and provides two research questions. Chapter 2 summarizes the evolution of data rescue and historical records transcription in past and present projects. It documents the improvements and modifications made by researchers to improve the performance of manual and OCR-based data rescue. It also mentions their attempts at implementing AI-augmented data rescue, citizen science projects, and transcription and how they recognized that AI can be a potential solution. Chapter 3 presents a survey study to help understand how data rescue community leaders perceive AI in data rescue, including their willingness, concerns, and possible trade-offs to AI use. Chapter 4 proposes a benchmark workflow for implementing AI technologies in data rescue. This workflow combines multiple AI techniques and is tested to automatically transcribe historical weather records. This process not only identifies potential future challenges, but also helps to provide recommendations for future steps. Chapter 5 concludes by identifying the challenges and opportunities for future AI-augmented data rescue and transcription research.

This thesis makes several contributions to the study of AI-augmented data rescue and historical records transcription. First, this research contributes a benchmark workflow for AI-augmented data rescue, which can be modified and customized to accommodate different kinds of historical records. Second, this research bridges the disconnect between AI technology and data rescue or historical records transcription in a broader sense. There have been calls for combining AI techniques with data rescue for improvement, but there are few studies that link AI with historical records transcription in general and address this multi-disciplinary transcription issue.

Third, this study is the starting point and cornerstone for future AI-augmented data rescue and historical records transcription research. It provides a guideline and reference for future data rescue and historical records transcription attempts that wish to benefit from AI technology.

References

Ashcroft, L., Coll, J. R., Gilabert, A., Domonkos, P., Brunet, M., Aguilar, E., Castella, M., Sigro, J., Harris, I., Unden, P., & Jones, P. (2018). A rescued dataset of sub-daily meteorological observations for Europe and the southern Mediterranean region, 1877–2012. *Earth System Science Data*, *10*(3), 1613–1635. https://doi.org/10.5194/essd-10-1613-2018

Blancq, F. L. (2010). Rescuing old meteorological data. *Weather*, *65*(10), 277–280. https://doi.org/10.1002/wea.510

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R., Brugnara, Y., Brunet, M., Brunetti, M., Chimani, B., Cornes, R., Domínguez-Castro, F., Filipiak, J., Founda, D., Herrera, R. G., Gergis, J., Grab, S., Hannak, L., Huhtamaa, H., … Wyszyński, P. (2019). Unlocking Pre-1850 Instrumental Meteorological Records: A Global Inventory. *Bulletin of the American Meteorological Society*, *100*(12), ES389–ES413. https://doi.org/10.1175/BAMS-D-19-0040.1

Brönnimann, S., Annis, J., Dann, W., Ewen, T., Grant, A. N., Griesser, T., Krähenmann, S., Mohr, C., Scherer, M., & Vogler, C. (2006). A guide for digitising manuscript climate data. *Climate of the Past*, *2*(2), 137–144. https://doi.org/10.5194/cp-2-137-2006

Brönnimann, S., Brugnara, Y., Allan, R. J., Brunet, M., Compo, G. P., Crouthamel, R. I., Jones, P. D., Jourdain, S., Luterbacher, J., Siegmund, P., Valente, M. A., & Wilkinson, C. W. (2018). A roadmap to climate data rescue services. *Geoscience Data Journal*, *5*(1), 28–39. https://doi.org/10.1002/gdj3.56

Brunet, M., Gilabert, A., Jones, P., & Efthymiadis, D. (2014). A historical surface climate dataset from station observations in Mediterranean North Africa and Middle East areas. *Geoscience Data Journal*, *1*(2), 121–128. https://doi.org/10.1002/gdj3.12

Brunet, M., & Jones, P. (2011). Data rescue initiatives: Bringing historical climate data into the 21st century. *Climate Research*, *47*(1), 29–40. https://doi.org/10.3354/cr00960

Camuffo, D., & Bertolin, C. (2012). The earliest temperature observations in the world: The Medici Network (1654–1670). *Climatic Change*, *111*(2), 335–363.

https://doi.org/10.1007/s10584-011-0142-5

Chimani, B., Auer, I., Prohom, M., Nadbath, M., Paul, A., & Rasol, D. (2021). Data rescue in selected countries in connection with the EUMETNET DARE activity. *Geoscience Data Journal*, *9*(1), 187–200. https://doi.org/10.1002/gdj3.128

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., … Worley, S. J. (2011). The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28. https://doi.org/10.1002/qj.776

Craig, P. M., & Hawkins, E. (2020). Digitizing observations from the Met Office Daily Weather Reports for 1900–1910 using citizen scientist volunteers. *Geoscience Data Journal*, *7*(2), 116–134. https://doi.org/10.1002/gdj3.93

Dupigny-Giroux, L.-A., Ross, T. F., Elms, J. D., Truesdell, R., & Doty, S. R. (2007). NOAA's Climate Database Modernization Program: Rescuing, Archiving, and Digitizing History. *Bulletin of the American Meteorological Society*, *88*(7), 1015–1017. https://doi.org/10.1175/BAMS-88-7-1015

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., & Ingold, R. (2014). *The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries* (pp. 91–106). https://doi.org/10.13140/2.1.2180.3526

Kwok, R. (2017). Historical data: Hidden in the past. *Nature*, *549*(7672), 419–421. https://doi.org/10.1038/nj7672-419

Levitus, S. (1992). *National Oceanographic Data Center Inventory of Physical Oceanographic Profiles: Global Distributions by Year for All Countries*. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service.

Levitus, S. (1996). Data Archaeology and Rescue of Historical Oceanographic Data: A Report on "The IOC/IODE GODAR Project." In *NOAA Technical Report NESDIS 87. Proceedings of The International Workshop on Oceanographic Biological and Chemical Data Management.*

Singh, A., Bacchuwar, K., & Bhasin, A. (2012). A Survey of OCR Applications. *International Journal of Machine Learning and Computing*, 314–318.

https://doi.org/10.7763/IJMLC.2012.V2.137

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R., Yin, X., Vose, R., Titchner, H., Kennedy, J., Spencer, L. J., Ashcroft, L., Brönnimann, S., Brunet, M., Camuffo, D., Cornes, R., Cram, T. A., Crouthamel, R., … Wyszyński, P. (2019). Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system. *Quarterly Journal of the Royal Meteorological Society*, *145*(724), 2876–2908. https://doi.org/10.1002/qj.3598

Stickler, Alexander, Brönnimann, Stefan, Jourdain, Sylvie, Roucaute, Eméline, Sterin, Alexander M, Nikolaev, Dmitrii, Valente, Maria Antónia, Wartenburger, Richard, Hersbach, Hans, Ramella Pralungo, Lorenzo, & Dee, Dick P. (2014). *ERA-CLIM Historical Upper-Air Data 1900-1972, supplement to: Stickler, Alexander; Brönnimann, Stefan; Jourdain, Sylvie; Roucaute, Eméline; Sterin, Alexander M; Nikolaev, Dmitrii; Valente, Maria Antónia; Wartenburger, Richard; Hersbach, Hans; Ramella Pralungo, Lorenzo; Dee, Dick P (2014): Description of the ERA-CLIM historical upper-air data. Earth System Science Data, 6(1), 29-48* [Application/zip]. 813 datasets. https://doi.org/10.1594/PANGAEA.821222

Tan, L. S., Burton, S., Crouthamel, R., van Engelen, A., Hutchinson, R., Nicodemus, L., Peterson, T. C., Rahimzadeh, F., Llansó, P., & Kontongomde, H. (2004). Guidelines on Climate Data Rescue. *World Meteorological Organization, WMO/TD No. 1210*. http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP-55.pdf

Wilkinson, C., Brönnimann, S., Jourdain, S., Roucaute, E., Crouthamel, R., Brohan, P., Valente, A., Brugnara, Y., Brunet, M., & Team, I. (2019). *Best Practice Guidelines for Climate Data Rescue*.

World Meteorological Organization. (2016). *Guidelines on Best Practices for Climate Data Rescue*. https://library.wmo.int/doc_num.php?explnum_id=3318

Chapter 2.Literature review

2.1 Introduction

The aim of this thesis is to investigate whether artificial intelligence (AI) can facilitate historical weather data rescue (hereafter referred to as “data rescue”). Historical climate data are fundamental building blocks for research to understand the past, present, and future climate scenarios. For example, it can be used to improve and validate weather and climate prediction models. However, many of these records are paper-based and are at risk of loss or deterioration. They are often stored in libraries and archives, and researchers cannot easily access and use these data digitally. Therefore, data rescue initiatives worldwide have been focused on preserving and transcribing these records into digital format to facilitate the creation of a global repository to access and preserve these records.

Data rescue is generally done by hand, which is labour-intensive and expensive. The question is whether this process can be automated. AI has been a rapidly developing field, and it has proven helpful in many different fields. Therefore, it is worthwhile to determine if AI is helpful in automating data rescue projects. Specifically, the goal is to determine the role of AI in data rescue and whether AI will bring opportunities or challenges to the field. The hypothesis is that AI will improve the automation of historical weather data rescue projects, as well as historical records transcription more generally. To assess this hypothesis, it is crucial to understand how data rescue projects have operated in the past and how automation has been carried out in past data rescue projects. Understanding how data rescue projects have worked in the past will provide information on whether automation or AI is necessary for future improvements. Examining past automation experience can also provide ideas on ways in which AI may help achieve a better version of automated data rescue.

In this literature review, I will first examine the definition of data rescue. Specifically, the questions are what constitutes the “data”, and how do we “rescue” them. This will provide an idea on the past and present operations of data rescue. Second, I will investigate the improvements and refinements that have been made to improve data rescue, including citizen science and optical character recognition (OCR). This will present the obstacles and barriers to current data rescue projects and show which are the remaining issues that still need

improvement. At last, I will examine the current stage of automated data rescue and the role that AI plays in automated data rescue projects. In other words, I will identify possible opportunities and challenges presented by past research on AI, and how these researchers perceive the role of AI in facilitating data rescue and historical records transcription projects.

2.2 Defining data rescue: how have people carried out data rescue?

Data rescue has been pursued in recent decades as a method to utilize historical data from past centuries. It was originally termed "data archaeology and rescue" by the U.S. National Oceanographic Data Center (NODC)/World Data Center A for Oceanography (WDC-A), referring to the recovery of records (e.g., scans, images, digital photograph) in obsolete media forms to more stable formats (Levitus, 1992, 1996). Nowadays, the action of data rescue is more focused on transcribing historical records (e.g., in paper, images, photographs) that are at risk of being lost due to deterioration or technical obsolescence into computer-encoded formats (World Meteorological Organization, 2016). These historical records are usually preserved on paper, and they need to be digitalized to be used by researchers for analysis. Much research over the past decades has used data rescue to obtain data (e.g., Ashcroft et al., 2014; Brunet et al., 2014; Kwok, 2017). For many researchers, the focus is on the content once it is rescued, for instance to be integrated into global circulation reconstruction models (e.g., the international Atmospheric Circulation Reconstructions over the Earth; Allan et al., 2011). As will be described, while some studies have documented detailed steps to "do" data rescue (e.g., Brönnimann et al., 2006; Slonosky et al., 2019), the majority has not focused on documenting the process. This prevents later studies from learning and improving from past experiences. The opacity of the process can hinder the potential for integrating AI and understanding the capacity of researchers to use AI in the data rescue process. In this section, I will infer from past research and summarize what "data" is and how researchers can "rescue" them.

2.2.1 What constitutes 'data' and which data should be rescued?

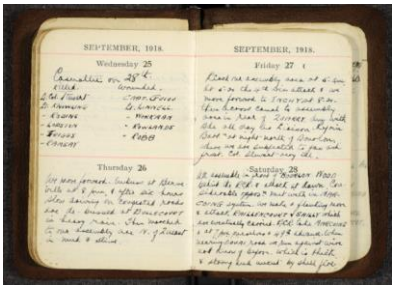
The public looks at the paper records and believes that they can be analyzed immediately. However, Sieber and Slonosky (2019) suggested that the source material only becomes analyzable if it is transformed into a machine-readable form. Some projects proposed that a repository that brings together information by storing fragmented data in a unified place is what data rescue aims for and that rescuing the metadata is sufficient (e.g., Veale et al., 2017).

However, a descriptive machine readable metadata of the meteorological records is not enough to comply with digital research methods; these records need to be transcribed into computer-encoded formats and then compiled into a digitally accessible repository for further analysis (Brönnimann et al., 2019). In other words, the content of the records as well as the metadata are key elements to be rescued. Traceability should be guaranteed while rescuing the data (World Meteorological Organization, 2016). The link to the original sources and context of the records must be retained (Veale et al., 2017). Ideally, it also should be a database searchable by geographic location, data, keywords (Black & Law, 2004; Veale et al., 2017). The value of data rescue lies in creating this digital repository.

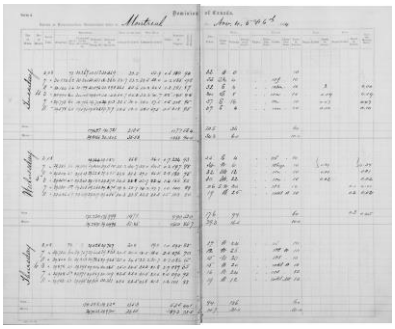
Researchers have discussed the resource constraints that limit their ability to collect all the data they would like to rescue. Researchers often need to make choices about which series of selected data should be prioritized for transcription in accordance with scientific needs and available resources. The prioritization can be based on relevant literature, applicable past experiences (Brönnimann et al., 2006), the age and rarity of the data, and the demands by cross-checking it with international databases (World Meteorological Organization, 2016). Guidelines like those from the WMO (2016) usually prioritize the minimum amount of data to be collected; specifically, they recommend prioritizing rainfall, pressure and wind data among other fields due to their importance. Other similar guidelines recommended that priority should be given to records that are at risk of being lost due to deterioration (e.g., Tan et al., 2004). In such cases, librarians and archivists may need to transform the content from paper to scanned images or microfilm prior to transcription. Data rescuers may need to deal with poorly scanned, noisy, skewed, and low resolution images without access to the paper records. Prioritization is important because many data rescue projects can be hampered by lack of funding and personnel if the project duration becomes too long (Brönnimann et al., 2018). Prioritization also reflects these limited resource issues, which automation may alleviate because less labour is needed to transcribe more records. With automation, prioritization may not be necessary. Historical records, regardless of their importance in being recognized by researchers, may have the opportunity to be transcribed.

It is also important to know what constitutes sources in the field of historical data rescue, as these sources can influence the collection procedure. Currently, three types of sources are recognized as potential sources of historical data: documentary evidence, instrumental information, and

natural proxies (see Figure 2.1 for examples) (Alcoforado et al., 2012). Instrumental data are predominantly used in data rescue studies, while fewer studies use documentary evidence and natural proxies (Kwok, 2017). Documentary evidence is non-instrumental evidence recorded in documentary archives. For instance, documentary evidence can be diaries, memories, daily weather reports and ship logbooks whose records are not measured with instruments. On the other hand, instrumental data are measurements made from standardized instruments, mostly found at weather stations and observatories. It can be stored in various media (e.g., paper logbooks, scanned images and digital files). Ship logbooks and many other kinds can also be instrumental data if the records are measured by standardized instruments (Brohan et al., 2010). Natural proxies, as the name suggests, are natural recorders such as tree rings, fossil, animal and plant remains, ice-cores and boreholes. Of these three types of data sources, priority is given to rescuing instrumental data, while the other two types can be compiled at a later stage (Brönnimann et al., 2019). These different types of historical records complicate the process of automating historical data rescue, as the techniques used for different types of records can be very different and it can be difficult to build a generic solution. Here, we will follow this recommendation and focus on instrumental data as the data source, although instrumental data also exhibits a degree of heterogeneity in its presentation.



(a) Documentary evidence. A page of Albert J. Kelly's weather diary. MG3054. McGill University Archives.



(b) Instrumental information. An example of a scanned leger sheet from McGill Observatory sub daily weather observations.

(c) Natural proxies. Tree rings that record the changes of the environment. Photo from the McGill University Redpath Museum.

Figure 2.1 - Three potential sources of historical data.

2.2.2 How to “rescue”?

Let us start with an explicit reference to the process of data rescue. Blancq (2010, p. 278) distilled the process of rescuing data into three steps: “First, old observations or data must be traced [obtained]; second comes the task of inputting or digitising [transcribing]; finally the data must undergo a thorough quality control check to detect and correct errors.” This data rescue workflow is instructive but may be insufficiently comprehensive to provide a guideline in the field. These three steps, while sounding less challenging if done by humans, can be very difficult and pose many challenges if we want to automate this process. Challenges include inconsistent source quality, source format, content, and many other factors, which will be mentioned in later sections.

The reference presented by Blancq (2010) is not sufficiently elaborate; two guidelines provide a more detailed workflow of data rescue (Brönnimann et al., 2006; World Meteorological Organization, 2016) and some studies provided a clear workflow for citizen science approaches (e.g., Slonosky et al., 2019; I will elaborate on the citizen science approach in a later section). Both guidelines contain clear recommendations for metadata collection, data preprocessing, and transcription. The first step of data rescue is metadata collection, which some call “data” collection. Before deciding to rescue observations, researchers and practitioners identify the source (e.g., official ship logs, newspapers, station ledger books, or weather diaries), geographic coverage/extent of the source material (e.g., the City of Montreal, southeastern Australia), and the data format (e.g., original, photocopy, image file or microfilm; table, graph or plain text; printed, handwritten or typewritten). The necessity of this step is also documented in the later work of Brönnimann et al. (2019). Metadata collection is not only about collecting metadata, but it also involves examining whether the data meet the project’s criteria. Sometimes, data rescue projects might only consider instrumental data sources (Brönnimann et al., 2019) or type written data. This information is also important for the selection of transcription methods. After passing the screening of this step, data can be proceeded for further processing. These steps are necessary, whether the data rescue is automated or transcribed manually.

The quality of the original documents is also important as noted in two guidelines (Brönnimann et al., 2006; World Meteorological Organization, 2016). According to the World Meteorological Organization (2016), it is important to ensure that the records are well organized and preserved,

and stored in acid-free archive boxes to ensure their quality. Brönnimann et al. (2016) are more concerned with the legibility and sometimes accuracy and precision of the records. Legibility issues can be diverse (Mateus et al., 2021). It includes blurred handwriting, faint ink, corrections over observations, missing decimal points, illegible symbols or values written in incorrect cells (Brönnimann et al., 2006; Mateus et al., 2021). Both guidelines are relevant because physical storage conditions can greatly affect the legibility of the records, and the quality of the records can greatly affect the transcription process as well. In other words, the quality of the source documents can also potentially affect the capabilities of automated data rescue.

Once the decision is made, researchers, librarians and archivists can start transforming paper copies to electronic images to facilitate transcription. Both guidelines discuss choices about photocopies (scanners) or photographs (digital cameras) as steps of imaging. Imaging can make the records more accessible, easier to handle, and sometimes more legible as well (Brönnimann et al., 2006; World Meteorological Organization, 2016). As part of preprocessing, good quality scanning is very important for optimal transcription results, especially if we apply AI automation as a transcription method.

Many data rescue studies use the word digitization to define the transcription of records into machine-encoded formats. This can be confusing, as archivists who scan the original documents often call this scanning process digitization. To avoid confusion about the meaning of digitization, this study will use the term transcription. In general, there are three types of transcription: key entry, OCR and speech recognition. According to Brönnimann et al. (2016), these three transcription methods have their own benefits and drawbacks. They recognized the role of automation in the transcription step of data rescue.

Key entry is usually accurate and efficient when the transcriptionist is a well-trained and fast typist; otherwise, manually keying the data can be very time-consuming and can also result in high error rates. One way to guarantee low error rates is through double or triple keying (i.e., having two or more transcriptionists transcribe the same text and compare the results to detect transcription errors), which has been suggested by both guidelines (Brönnimann et al., 2006; World Meteorological Organization, 2016). For example, many studies guarantee 99 percent or higher accuracy rate through double or triple keying (e.g., Brohan et al., 2009; Ryan et al., 2021). Budget constraints are the most common reason that makes this option unfeasible, given that double or triple keying would consume longer time and require more personnel (Ashcroft et al.,

2018). Another way to lower the error rate is to have professionals transcribe the data themselves or to help train transcriptionists on what the data means (Brönnimann et al., 2006; Capozzi et al., 2020; Stickler, Brönnimann, Jourdain, et al., 2014; Stickler, Brönnimann, Valente, et al., 2014). With some knowledge, transcriptionists will have a good idea of the range, format or units of the entry, and they will be able to decipher semi-illegible entries and flag suspicious values more often than those without the knowledge.

Speech recognition has similar advantages and drawbacks, as it is an alternative way to perform key entry. Speech recognition is done by the speaker dictating content to the program, which then converts it into text. Whether a project should choose key entry or speech recognition depends mainly on the transcriptionist (Brönnimann et al., 2006). One disadvantage that the key entry does not have is that it is common to find two words with the same pronunciation, a problem that varies from person to person. Brönnimann et al. (2016) mentioned that sore throats are also a common problem that may slow down the transcribing process. Considering that almost no current and past projects have used this approach, I will not discuss this approach further in this literature review.

OCR methods are many times faster than the two methods mentioned above, but the error rate is also much higher (Stickler, Brönnimann, Jourdain, et al., 2014). I will discuss OCR methods in far greater depth below. For now, higher error rates from automation can make the correction process time consuming and thus transcription can take more time than keying manually (Blancq, 2010; Brönnimann et al., 2006).

In deciding which transcription method is the best practice, both data rescue guidelines agreed that key entry is the most efficient method because it produces the fewest errors, so human interpretation and manual transcription remain the most trustworthy methods. Later studies have also shown that key entry methods are preferred (e.g., Allan et al., 2021; Ashcroft et al., 2018; Stickler, Brönnimann, Jourdain, et al., 2014). A quantitative evaluation showed that key entries produced the fewest errors among the three approaches (Brönnimann et al., 2006). For example, Brönnimann et al. (2006, p. 140) found that, while key entry was slightly slower than speech recognition, the error rate was “low” if compared to “middle” for speech recognition and “high” for OCR. Key entry is also more adaptable than other methods, as it works with most record types and formats; whereas OCR has proven too challenging for handwritten or unstructured records (Brönnimann et al., 2006; World Meteorological Organization, 2016). As of the date of

these two guides (2016), OCR, if successful, is the fastest way to transcribe printed and tabular materials; speech recognition is the next best thing for people who do not prefer key entry.

How one does transcription, especially key entry, is important since it points to the potential for automation. However, aside from these two guidelines, the majority of the literature worldwide has a very limited documentation of the 'how' (e.g., Alcoforado et al., 2012 for Portugal; Allan, Brohan, et al., 2011, Allan, Compo, et al., 2011 and Brunet & Jones, 2011 for global effort; Ashcroft et al., 2014 for southern Australia; Brohan et al., 2010 for Arctic; Camuffo et al., 2013 for Western Mediterranean; Cornes et al., 2012 for city of London; Domínguez-Castro et al., 2014 for Spain). For example, Alcoforado et al. (2012) only detailed the data source, data collection, and data analysis. Allan, Brohan, et al. (2011), Allan, Compo, et al. (2011) and Brunet and Jones (2011) focused on making data rescue a global and collective effort by facilitating interactions among countries, organizations and academia, but none of the articles address the processes of transcription. Further examples include Ashcroft et al. (2014) who focused primarily on quality control and data analysis; similarly, Brohan et al. (2010) and Cornes et al. (2012) focused mainly on data sources. These articles briefly mentioned transcription, but none of them documented the process. Both Camuffo et al. (2013) and Domínguez-Castro et al. (2014) discussed data sources in detail, but they also did not provide sufficient details about transcription either. This insufficient information about the transcription process makes it difficult to replicate and improve, let alone implementing automation.

In this section, I focus on investigating the key entry process, as other transcription methods will be discussed in the subsequent sections. I was able to gather a bit of information about the key entry process, although most data rescue articles give only a few sentences or barely a paragraph to describe their key entry process. Starting with the role of transcriptionists, the articles presented three options: hired contractors, project personnel, and volunteers. For example, Dupigny-Giroux et al. (2007) hired contractors to key over 100,000 meteorological observations from daily newspapers. On the other hand, Blancq (2010), Capozzi et al. (2020) and Stickler, Brönnimann, Valente, et al. (2014) transcribed the records by personnel involved in the study or projects. The slight difference here is that personnel in the study are people with expertise in the data context but not the keying process, while contractors are professionals who key the records but do not have the knowledge about the data context. Then, there is a third option of transcriptionists – volunteers. For example, Ryan et al. (2021) have final year Geography

undergraduates serve as volunteers to transcribe rainfall data in Ireland; Brohan (2012) used an citizen science - more on citizen science in the next section - website to help transcribe US weather records from ship logbooks; Ashcroft et al. (2016), Hawkins et al. (2019), Sieber and Slonosky (2019), Slonosky (2014) and Slonosky et al. (2019) created or customized their own website to attract volunteers help transcribing their historical records. Although these studies listed their transcribing method and transcriptionists, they did not provide detailed steps such as how they performed the key in procedures. This does provide some insight for future data rescue studies or projects, but it is still not detailed enough for future reference, especially if we want to include AI in data rescue.

I can only speculate on the process of data rescue by referring to a few studies (e.g., Ashcroft et al., 2018; Brunet et al., 2014; Camuffo & Bertolin, 2012; Stickler, Brönnimann, Valente, et al., 2014). In reviewing their workflow, these data rescue projects have several common characteristics. My conclusion is that data rescue projects are time consuming, require special care, and need to be well structured and managed.

It is time-consuming to complete a data rescue project when key entry is used. For example, eleven transcriptionists worked fifteen hours per week for two years to manually transcribe sub-daily meteorological observations in Europe (Ashcroft et al., 2018). Other articles provided similar observations. Even key entry is a time-consuming method, its advantage of being easy to apply and low cost make it by far the most preferred data rescue approach. Stickler, Brönnimann, Jourdain, et al. (2014, p. 39) suggested that “Even though manual digitization [transcription] is a simpler, but slower process, it was maximally optimised”. Similarly, Ashcroft et al. (2018) claimed key entry is by far the most cost-efficient method, although it is time-consuming. If the involvement of AI in data rescue can increase the speed of transcription while maintaining its simplicity and low cost, this will greatly benefit the rescue of historical records.

The data rescue process requires special care and expertise. By providing expert knowledge such as formats (e.g., numeric, string, boolean), ranges, and units, Stickler, Brönnimann, Jourdain, et al. (2014) improved accuracy of the data entry. Brunet et al. (2014) and Camuffo and Bertolin (2012) also emphasized the importance of expert knowledge, as most historical data are handwritten in faded ink and are difficult to recognize, so expertise can help transcriptionists identify the true value of the data. Microsoft Excel templates have been by far the most popular choice for recording transcribing results, especially for tabular data format. Stickler,

Brönnimann, Jourdain, et al. (2014, p. 34) embedded expert knowledge as info tabs in the Excel template to further facilitate transcriptionists and improve accuracy. As a result, the transcriptionist can transcribe faster and more accurately because they can understand the context of the data and less personnel training will be needed. However, training or otherwise embedding expert knowledge is also time consuming. If this expert knowledge can be embedded into the automation process, there would be no need for training to help transcriptionists understand the data context, and possibly more time can be saved.

Data rescue is also a project that needs well-structured management, especially when it comes to human transcriptionists. Proper organization and categorization of the records to be scanned, transcribed, and validated is important, and the training and handling of transcriptionists is critical. Stickler, Brönnimann, Jourdain, et al. (2014) developed a web interface to manage their transcriptionists and to catalog over 2.8 million station days of records. Similarly, Ashcroft et al. (2018) used a central server to track the progress of eleven transcriptionists and back up the transcriptions regularly. When it comes to human transcriptionists, extra effort is needed to properly manage the personnel; although this extra effort may not be needed if automation is used, experience can be gained from cataloging millions of records.

2.3 Engaging citizen science

Improvements have also been made to enhance the performance of key entry by adding in new concepts (Brönnimann et al., 2018; World Meteorological Organization, 2016). Given the large number of historical records that need to be transcribed, many data rescue projects seek citizen science as an alternative and improvement (Brönnimann et al., 2018). AI also brings unique implications for citizen science, which I will also discuss. I will first discuss the use of citizen science in historical records transcription, and then I will discuss AI use in citizen science.

2.3.1 Citizen science in rescuing historical records

There are nuances in the exact definition and interpretation of citizen science, but it is generally accepted that citizen science refers to the inclusion of public members in some aspects of scientific studies (Eitzel et al., 2017). It helps create large and stable data repositories, promotes data accessibility, increase citizen engagement in scientific research, and improves scientific literacy (de Sherbinin et al., 2021). There are two main branches of citizen science projects:

primary observations and secondary data collection (mainly transcription). Examples of primary observations include monitoring wildlife (Davis et al., 2012; Howard et al., 2010; Sullivan et al., 2009) and the environment (e.g., Rambonnet et al., 2019); examples of secondary data collection include transcription of historical records (e.g., Slonosky et al., 2019) and image classification (e.g., Fortson et al., 2011). Citizen science volunteers can participate in three major ways: contributory, in which participants collect the data; collaborative, in which participants are involved in the analysis phase; and co-created, in which participants are involved in all phases of the project (Follett & Strezov, 2015).

In data rescue, many of these citizen science projects used web-based tools to transcribe historical records, such as Data Rescue Archives and Weather,[1] Old Weather,[2] Rainfall Rescue,[3] Weather Rescue,[4] Weather Detectives (ended in 2017), and Weather Wizard.[5] These are all climate-related projects. In addition to climate-related projects, the Zooniverse has a large number of data transcription projects using citizen science methods in different disciplines, such as HMS NHS: The Nautical Health Service,[6] which transcribes hospital medical records in Greenwich from 1826 to 1930, Criminal Characters,[7] which transcribes Australian prison records from the 1850s to 1940, and Star Notes,[8] which transcribes over 2,500 notebooks produced by early 20th century women astronomers at the Harvard College Observatory. In recent years, many data rescue projects have been using citizen science approaches to help rescue historical records, and it has proved to be an improvement.

One of the main themes of citizen science is motivation. Primary research summarizes the motivations for participating in citizen science projects, including contributing to scientific knowledge, fulfilling interests, receiving recognition for contributions, and acquiring knowledge (Batson et al., 2002; Raddick et al., 2009; Rotman et al., 2014; Sieber et al., 2022). In data rescue transcription projects, volunteers may be motivated by their interests in history and the corresponding fields. Interestingly, Eveleigh et al. (2014) found that citizen science transcription

[1] https://draw.geog.mcgill.ca/
[2] https://oldweather.org
[3] https://zooniverse.org/projects/edh/rainfall-rescue
[4] https://zooniverse.org/projects/edh/weather-rescue
[5] https://weatherwizards.org
[6] https://zooniverse.org/projects/msalmon/hms-nhs-the-nautical-health-service
[7] https://zooniverse.org/projects/ajpiper/criminal-characters
[8] https://zooniverse.org/projects/projectphaedra/star-notes

projects may also motivate “dabblers” when the task soothes anxiety. It is important to understand keeping participants motivated is important to retention of volunteers in a project.

Citizen science approaches serve scientific goals as well as educational purposes (Bauer et al., 2016; Brossard et al., 2005). Citizen science can produce accurate data and it can enhance scientific literacy. It can be used in formal settings such as classrooms or informal settings such as museums, where students and the general public have an opportunity to learn about scientific research. Here are some examples of projects that have undergraduate or graduate students transcribing historical data as part of their assignments, with results reported to be as accurate as those transcribed by professionals (Mateus et al., 2021; Ryan et al., 2018, 2021). Ryan et al. (2018, 2021) had 142 undergraduate geography students from Maynooth University transcribe more than 1,300 stations years of daily rainfall observations. An examination showed that all transcriptions had an error rate of less than one percent (Ryan et al., 2021). Similarly, Mateus et al. (2021) had university and secondary school students transcribe daily air temperature records. The secondary school students achieved an accuracy rate of 95.2 percent, which was even slightly higher than the 95 percent accuracy rate given by the Weather Rescue project (Mateus et al., 2021). The results show that with proper training, instructor-guided students can be an excellent resource to accelerate the transcribing process compared to traditional approaches (Ryan et al., 2018). They identified implementation of citizen science as a new strategy with the potential to improve current data rescue strategies. However, as pointed out, there are too many records and too few people, so involving students may still not be enough for all the records that need to be transcribed. AI automation may be a way to break through this difficulty.

Studies have shown that the citizen science approach has made data rescue more accurate and efficient. For example, Hawkins et al. (2019) used the weather rescue website to transcribe more than 1.5 million weather observations in less than three months with the help of more than 3500 citizen science volunteers. Typically, transcribing this amount of records would be equivalent to six years of a full time job (Hawkins et al., 2019), but with the help of citizen science volunteers, the project was completed in three months. Not only was the project very efficient, but it was also of high quality. There were three replicates per entry and the accuracy of the transcription was over 95 percent (Hawkins et al., 2019). Craig and Hawkins (2020) later used the same websites to transcribe an additional 1.8 million weather observations with the help of 2,148 citizen science volunteers. These citizen science volunteers were able to transcribe weather

observations from 72 different locations over a seven-month period. The quality of this transcription project was also improved by having multiple volunteers typing the same data. Thus, Craig and Hawkins (2020) concluded that citizen science is a more efficient and accurate way to transcribe a large number of historical records than traditional data rescue methods where limited numbers of personnel work together. The improvements citizen science has made to data rescue projects are partially aimed at improving efficiency and accuracy, but it might not be sufficiently helpful given the vast amount of historical data that is not transcribed to date. In other words, even though citizen science methods are efficient, they may not be fast enough to rescue all the records. AI may be a possibility to further improve data rescue and make the transcription process more accurate and efficient.

A number of ongoing data rescue projects using citizen science approaches have shown that the citizen science approach promotes reusability of projects compared to traditional data rescue projects that are hard to replicate. It also has the advantage of positive feedback because citizen science volunteers can help improve the project. For example, in the case of DRAW, a volunteer suggested that the team add a "ruler" to virtually line up the weather diaries, which potentially avoided many errors (Sieber & Slonosky, 2019). DRAW is an ongoing citizen science project that transcribes the McGill Observatory's meteorology records from 1874 to 1953. The DRAW software is publicly shared via GitHub and anyone is welcomed to download the code and use it for their data rescue projects (Sieber & Slonosky, 2019; Slonosky et al., 2019). By doing so, the team hopes that other historical climatologists or historical researchers in other fields, such as geophysics, medicine and other complex forms, will use the code for their data rescue projects. These types of shared systems ensure the reusability and replicability of projects and it has the potential to help transcribe large amounts of historical data. With more people brainstorming improvements, citizen science projects will have more perspective and can outperform traditional data rescue projects. Knowledge sharing and reusability are things AI automation can learn to further improve data rescue projects, perhaps over and above the performance of citizen science projects.

2.3.2 Involving AI in citizen science

As a step forward, some of the citizen science projects have also involved AI. Although many studies have investigated the use of citizen science and AI separately, the integration of citizen

science and AI and the application of this integration is new and still in its infancy (Franzen et al., 2021; Green et al., 2020; Lotfian et al., 2021; Rafner et al., 2021). Unfortunately, as of this writing, there are no citizen science projects involving AI in historical records transcription. There are some examples of citizen science projects in environmental science, neuroscience, astronomy, and life science that involve AI in tasks that can be grouped into recognition (e.g., classification, object detection) and prediction (e.g., validation, learn from participants to improve performance) (Ceccaroni et al., 2019; Franzen et al., 2021; Lotfian et al., 2021; McClure et al., 2020; Ponti & Seredko, 2022). Results have shown that the integration of AI and citizen science can maintain the high accuracy of human work and reduce much more effort and time (e.g., Torney et al., 2019; Willi et al., 2019). For example, Willi et al. (2019) used the animal species classification results from citizen science volunteers to train a deep learning model and showed that the model achieved accuracy very close to that of humans while reducing human effort by 43 percent. Similarly, in another project to count wildlife in aerial survey images, a deep learning model trained with a dataset annotated by citizen science volunteers achieved accuracy comparable to humans and much faster than citizen science volunteers (Torney et al., 2019). Green et al. (2020) concluded that incorporating AI into citizen science projects to save time and resources is the best option when dealing with large amounts of material. Similarly, Franzen et al. (2021) recommended incorporating AI as the next step in citizen science, as this integration presents unimaginable opportunities and possibilities.

The idea of automated data transcription has been widely considered in the field of data rescue, especially for historical climate data, but the adoption and the corresponding results have not been promising. Most of the AI in citizen science-related data rescue is OCR. OCR originally referred to the extraction of text from an original source (e.g., images, photographs) and transcription of that text. However, sometimes the text extraction step is overlooked, and the automated transcription step is referred to as OCR. To avoid this confusion, OCR here will refer to the combination of text extraction and transcription. Also it should be noted that, similar to the term digitization, OCR can be a confusing term in this multidisciplinary setting, as researchers in different fields may refer to slightly different definitions when using the term OCR. Several data rescue projects have attempted transcription with OCR, and while they do mention that automated transcription would save time and money (Brönnimann et al., 2006; Wilkinson et al., 2019; World Meteorological Organization, 2016), they have also found some collateral problems. Researchers reported that the OCR techniques they used led to many errors, mainly

due to poor image quality of the transcribed sources (Blancq, 2010; Craig & Hawkins, 2020; Stickler, Brönnimann, Jourdain, et al., 2014). Other reported problems include OCR software being too sensitive to script (e.g., handwritten, printed, and typewritten text) (Blancq, 2010; Brönnimann et al., 2006; Craig & Hawkins, 2020; World Meteorological Organization, 2016) and immature in transcribing records in different formats (e.g., text, table, graph) (Stickler, Brönnimann, Jourdain, et al., 2014; Stickler, Brönnimann, Valente, et al., 2014; Wilkinson et al., 2019) and context (e.g., digits, characters, and alphanumeric) (Brönnimann et al., 2006; Stickler, Brönnimann, Jourdain, et al., 2014). At present, automated transcription approaches in data rescue are still under development and testing, but it is expected that these approaches will be more reliable and sufficiently mature in the future (World Meteorological Organization, 2016). Specifically, they pointed to machine learning techniques, which is a subset of AI, as a possible future direction for refining the transcription of historical records (Chimani et al., 2021; World Meteorological Organization, 2016).

There has been much discussion about integrating AI and humans as a hybrid approach to rescue historical records. Discussions on this hybrid approach have focused on the role of AI and humans and task allocation. Ceccaroni et al. (2019) proposed three ways of allocating the roles of AI and humans: (1) assisting or replacing humans in completing tasks, (2) influencing human behavior, and (3) improving insights. Rafner et al. (2021) agreed and suggested that the first role and the third role could be merged into assistance in solving the citizen science tasks. More specifically, Ponti and Seredko (2022) suggested that routine tasks that are well-defined (e.g., collecting data, counting objects, pattern recognition) are susceptible to AI and automation, whereas nonroutine tasks that involve problem solving, creativity and intuition (e.g., designing RNA sequences) should be done by human - citizen science volunteers. On the other hand, Green et al. (2020) and Torney et al. (2019) argued that humans (especially citizen science volunteers) have a clear role to play here, that is to participate in the project and help provide training datasets for AI models. They suggested that AI results could also be used as another vote to validate the human results. So far, in many tasks, AI is not capable of replacing humans, so deciding the role of AI will depend heavily on the goals of the project (e.g., public engagement, efficiency, productivity) (Ponti & Seredko, 2022). As previously discussed, this integration of AI and humans, while still respecting the role of the citizen scientists, is still new.

Integrating and involving AI in historical records transcription implies opportunity and possibilities, but also poses risks and challenges. A major risk is the ethics and data ownership when citizen science volunteers disengaged (Ceccaroni et al., 2019; Lotfian et al., 2021; McClure et al., 2020). They may not be comfortable to share their contribution to building training datasets, especially for commercial use. As a result, participants may disengage from the citizen science community. It is crucial that they understand how their contributions are being used and that a transparency approach, rather than a black box project, should be adopted so that participants can decide early on whether they want to participate (Ceccaroni et al., 2019; Lotfian et al., 2021; McClure et al., 2020). For example, eBird has made information available about their AI and human integrated species identification system (Sullivan et al., 2014). Another identified risk is the opacity of AI algorithms. This can be an issue when critical decisions need to be made and we need to interpret the model by understanding how the results were produced (Franzen et al., 2021). This can be critical when dealing with social science related projects, where the results can directly impact societal concerns. As AI develops, more risks may be identified along the way, and it is important to be prepared before things get complex.

2.4 AI usage for data rescue

In this section, I discussed the use of AI in data rescue. First, I reviewed and summarized the use of AI for automated transcription in past and present data rescue. Second, I discussed what the perception of using AI is in rescuing historical records and how it might improve future projects. Third, I reviewed layout analysis briefly. At last, I summarized the unsolved challenges remained in AI use for historical records transcription.

2.4.1 Automated transcription

OCR is a valuable technology for translating printed and handwritten text from historical documents (e.g., from scanned images, photographs) into datasets in a machine-readable and analyzable format (Memon et al., 2020). The earliest OCR system was developed eight decades ago, and due to technological advances, it has become more powerful over time as it can handle printed, typed, and handwritten documents. It has benefited many fields by transcribing materials such as license plates, invoices and legal documents and provides services to areas such as banking and healthcare (Singh et al., 2012). It is also widely used to preserve historical records in various fields for national repositories or institutional libraries (Chimani et al., 2021; Swindall

et al., 2021; Yasser et al., 2017). As OCR continues to evolve, it will continue to benefit many fields, so it is worthwhile to examine how the development of OCR can improve current methods in preserving historical records.

Studies have found that automated character recognition is especially challenging for historical handwritten documents (Alabau & Leiva, 2012; Firmani et al., 2017; Holley, 2009; Jander, 2016; Jenckel et al., 2016; Swindall et al., 2021). Alabau and Leiva (2012) believed that the wide variety of handwriting styles and degraded documents make recognizing historical handwriting a difficult and challenging task. This task is conditional on the quality and type of training data. The training data is data that has been discretized and is labeled. Most of the training datasets are done on printed handwriting (cf., Burrell, 2016). Firmani et al. (2017) added that, while impressive results have been achieved with historical printed document recognition, recognizing handwritten documents is still challenging and laborious. As a result, many researchers agreed that more effort needs to be put into recognizing documents, especially handwritten ones (Biondich et al., 2002; Gatos et al., 2014; Holley, 2009; Yasser et al., 2017). To understand the current status and future directions of handwritten character recognition, Memon et al. (2020) conducted a systematic literature review of 176 papers on character recognition of handwritten documents published between year 2000 and 2019. They found a dramatic increase in the number of publications from 2018 to 2019, specifically, 59 publications in the eight years from 2010 to 2017, and 55 new papers from 2018 to 2019. They also highlighted that automated transcription research has shifted from classical feature extraction approaches to deep learning approaches, especially CNNs (Convolutional Neural Network). This is not surprising, as we have seen advances in deep learning and computer vision in recent years. To advance Handwritten Text Recognition (HTR) to rescue historical documents, state-of-the-art approaches need to be investigated.

Training is an important step for automated transcription to achieve the desired performance (Terras, 2022). This means that the availability of training datasets and training models need to be considered when building an automated transcription system. However, the challenge is that it is often difficult to find matching training datasets, and building a training datasets is challenging and will cost a considerable amount of time and money (Dahl et al., 2021; Jenckel et al., 2016). The details will be discussed in the next section.

Researchers have investigated technological fixes that can improve the text recognition. Memon et al. (2020) summarized the most prevalent method used in automated transcription research from 2000 to 2019. They synthesized the method into five major types: Artificial Neural Networks (ANN) (e.g., CNN, Recurrent Neural Network {RNN}), kernel methods (e.g., Support Vector Machines {SVMs}), statistical methods (e.g., Logistic Regression {LR}), template matching techniques (e.g., deformable template matching), and structural pattern recognition (e.g., Chain Code Histogram {CCH}). Recent benchmark automated transcription models have made extensive use of deep learning methods, particularly CNNs (e.g., Firmani et al., 2017; Swindall et al., 2021; Yasser et al., 2017), RNNs (e.g., Fornés et al., 2017; Jenckel et al., 2016; Parthiban et al., 2020), and hybrids of CNNs and RNNs (e.g., Chamchong et al., 2019; Dahl et al., 2021; Lehenmeier et al., 2020; J. A. Sánchez et al., 2019). Memon et al. (2020) concluded that kernel methods were one of the most popular and robust methods for automated transcription before the emergence of deep learning methods, while statistical methods (J. A. Sánchez et al., 2013, 2014) were frequently used in the 2000s. Prior to that, structural pattern recognition was widely used by the automated transcription research community (Memon et al., 2020). There are also several other studies and benchmarks that have used open-source automated transcription software, such as Tesseract (Li et al., 2016; Neudecker et al., 2019; Odunayo et al., 2021; Rakshit et al., 2010) and OCRopus (Holley, 2009), as well as commercial OCR software, such as ABBYY FineReader (Holley, 2009; Odunayo et al., 2021) and Teleforms Elite (Biondich et al., 2002). Tesseract uses a deep learning based Long short-term memory (LSTM) OCR engine in its version 4.0. However, the underlying technology used by commercial software such as ABBYY FineReader remains unknown, probably because it is a commercial product (Tafti, 2016). To date, deep learning remains the most popular and advanced automated transcription method, enabling researchers to improve the accuracy of results, so it is necessary to find a robust automated transcription algorithm that can adapt to all the different documents.

Current automated transcription algorithms still have room for improvement and refinement. Automation is still in the early stage of development (Yasser et al., 2017), and the existing OCR techniques are “far from offering error-free solutions” (J. A. Sánchez et al., 2014, p. 112). One of the future improvements suggested by several studies is a robust system that can incorporate all possible character variations (Hanson & Simenstad, 2018; Jenckel et al., 2016; Memon et al., 2020; Neudecker et al., 2019). Hanson and Simenstad (2018) noted that while the current models have high accuracy on the specific documents, they do not generalize well on other documents,

for example, with different handwriting styles and languages. Therefore, there is a growing demand for automated transcription algorithms that can handle a variety of documents and still perform well (Jenckel et al., 2016; Neudecker et al., 2019). Robustness in automated transcription development would be something that future research should investigate.

Another suggestion for future improvements of the automated transcription models is the commercialization of historical records transcription research, for example by providing a user interface or customizing the model to something that is easy to install and does not require much prerequisite knowledge. As Shen et al. (2021, p. 2) commented in their study, "many researchers who would benefit the most from using these methods lack the technical background to implement them from scratch." Memon et al. (2020) suggested that this move will help create a low-cost system that could be used in a larger population. Jenckel et al. (2016) agreed, saying that there is a growing need for an OCR system that requires low effort and that researchers in other specialties should be able to apply it in their research. Meanwhile, these programs should be able to be customized for different styles of documents (Lehenmeier et al., 2020). Customization will increase the effective utilization of automated transcription in historical document research and provide a direction for future benchmarks to improve the process.

2.4.2 AI could be a solution

AI is a rapidly emerging field that offers many solutions for the transcription of historical data in data rescue. The automated nature of AI-augmented transcription could open a new chapter in analysis of historical records (Terras, 2022). Harnett (1985) pointed out that the OCR used in the past has limited ability to handle a variety of fonts and formats without the augmentation of AI. Harnett also believed that AI will be able to enhance OCR so that it would be able to recognize any type of record with proper training. Harnett's belief has been supported by many studies (Memon et al., 2020; Ströbel et al., 2022; Terras, 2022). Handwritten character recognition has undergone tremendous improvements and breakthroughs, partly due to the development of neural networks in recent years (Graves et al., 2009; Graves & Schmidhuber, 2008; Ströbel et al., 2022). Terras (2022) noted that machine learning techniques have been actively integrated with OCR to improve accuracy, and many benchmarks and approaches have been published or applied by various projects (J. A. Sánchez et al., 2019; Terras, 2022). Specifically, Memon et al. (2020), after reviewing 176 articles about handwritten OCR from 2000 to 2019, concluded that

the recent trend of transcription methods is shifting from manual approaches to automated approaches based mainly on deep neural networks. They stated that the most AI-augmented approaches have been published between 2017 and 2019, and they believed the advances in AI-enhanced character recognition will continue and increase in the future. The popularity of AI in enhancing OCR suggests that AI provides an opportunity for every researcher to explore historical handwritten character recognition solutions.

Automated approaches have become a popular research topic mainly due to their ability to save time. Fischer et al. (2014) implied that manual transcription is time-consuming and does not allow transcribing a large number of records in a reasonable amount of time, while automation allows researchers to do this. For example, Biondich et al. (2002) reported in their study that for a task that would have taken 36 minutes to complete manually would have taken only twelve minutes with their automated approach. Similarly, Fornés et al. (2017) found in their study that manual transcription was fifteen percent slower than the automated approach in their case. World Meteorological Organization (WMO) also implies that AI-augmented transcription is one of the solutions to make data rescue projects more efficient (World Meteorological Organization, 2016). In conclusion, it has been found that manual transcription is tedious and time-consuming, while automated approach can speed up the process (Chen et al., 2018). Therefore, it is crucial that we investigate AI-augmented methods to improve the current historical document transcription projects.

2.4.3 Layout analysis

Several studies have built a transcription workflow on top of existing layout analysis models to improve performance. For example, Odunayo et al. (2021) incorporated a CascadeTabNet pre-trained model fine-tuned by their dataset to predict and draw bounding boxes of their tabular regions. Similarly, Ziomek and Middleton (2021) used an improved version of CascadeTabNet by adding a post-processing step to detect table boundaries and individual cells. Both studies yielded good results, in particular, Ziomek and Middleton (2021) reported a significant improvement in the F1 score of their improved model. Building up on previous research, it is possible to improve the performance of layout analysis by fine tuning the model or adding post-processing steps. This knowledge can be used in future attempts to build a transcription workflow that includes layout analysis.

Other studies have also used some off-the-shelf software for layout analysis, especially for line segmentation. Ströbel et al. (2022) used the software Transkribus to segment their text lines for transcription. Similarly, Lehenmeier et al. (2020) have also used Transkribus to manually extract the text line data and export them into annotated format. Transkribus is a handwritten text recognition platform and software that provides AI-powered layout analysis and text recognition for historical handwritten documents in different languages.[9] In a survey that received 154 responses from Transkribus users, they reported that the most useful features were "automatic line detection, layout analysis, and segmentation" (Terras, 2022, p. 53). However, several studies have shown that these off-the-shelf software did not produce promising results. Odunayo et al. (2021) reported some rows and columns went missing sometimes while using ABBYY FineReader for text line recognition on tabular data. Lehenmeier et al. (2020) found that Transkribus was not able to automatically detect the table region and structure on their tabular weather data. It seems that the existing off-the-shelf software does not perform consistently across different documents and may introduce errors. Therefore, it is preferable to find a robust and accurate layout analysis model for the transcription workflow.

2.4.4 Unsolved challenges

Automatic transcription of historical documents is still a relatively new area of research that is still evolving and is always incorporating innovative ideas. Inevitably, therefore, there are many unsolved issues and challenges. One of the most mentioned challenges in many studies is the training process involved in automation. Research has shown that training is necessary and useful because it improves accuracy and makes the process more robust and faster. Studies in both layout analysis (Prasad et al., 2020; Shen et al., 2021) and automated transcription (Fornés et al., 2017; Holley, 2009; Terras, 2022) have pointed out that the training process is necessary for the model to achieve desired accuracy. In other words, they confirmed that training can improve the accuracy of the model. For example, Odunayo et al. (2021) and Rakshit et al. (2010) found in their study that there was a significant improvement in the accuracy after training the original Tesseract model with their custom training dataset. Another study, a deep learning approach using a combination of CNN and LSTM, also reported an increase in accuracy if the model was trained with a training set (Lehenmeier et al., 2020). It has also been concluded that

[9] https://readcoop.eu/transkribus/

training is unavoidable to make the model more robust. In a systematic literature review that synthesized 176 handwritten automated transcription papers, Memon et al. (2020) concluded that training would allow the model to adapt to a variety of documents. Another study also mentioned that training will make the transcription workflow faster (Jander, 2016). Researchers appear uniform in their belief that training can bring a lot of benefits to the workflow, although who will do it or how it will be customized is questionable.

Many studies have demonstrated in their experience that although training is a necessary step, in practice it is always hard to find a training dataset and most of the time such projects will lack a training dataset (Dahl et al., 2021; Fischer et al., 2014; J. A. Sánchez et al., 2019). Specifically, this can be particularly hard for small collections, as it is difficult to obtain enough training data (Lehenmeier et al., 2020). In addition, there is no easy way to create a training dataset from scratch with minimal knowledge (Shen et al., 2021). In terms of creating a training datasets, studies also found that it is time consuming, expensive and requires a lot of effort, especially for large scale datasets with various types of documents(Holley, 2009; Jenckel et al., 2016; Swindall et al., 2021; Yasser et al., 2017). Another challenge in creating and finding training datasets is that there are no one-dataset-fit-all cases (Nikolaidou et al., 2022). This means that different documents will likely require different training datasets to achieve desired outcomes. While the creation of training datasets remains a problem in automating historical documents transcription, future efforts in improving this workflow should consider finding a solution to this problem.

Recent studies have made some suggestions to alleviate the problem of training data. J. A. Sánchez et al. (2014) suggested that we should find a method that requires fewer training data to achieve a good result. Therefore, transfer learning has been proposed and used in several studies where only a small amount of data is needed to solve multiple tasks sharing similar domains (Prasad et al., 2020; Ströbel et al., 2022). For example, Ströbel et al. (2022) proposed a highly transferable method that was pre-trained on modern English manuscripts, which they successfully used to transcribe historical Latin documents with good results. The ability of this model to transcribe language and handwriting style it has never been trained on is potentially a partial solution to the lack of training data.

2.5 A combination of techniques may be needed to accommodate data rescue/citizen science

Many studies suggested that AI-augmented historical handwritten document transcription requires a combination of techniques and corresponding models. For example, to achieve good results, techniques such as layout analysis, text line segmentation, and deep learning augmented text recognition need to work together although much of them are still under development (J. A. Sánchez et al., 2013, 2019). The authors stated that the maturity of automatic handwritten text recognition largely depends on the progress of the combination of benchmark models. The theory of utilizing several techniques and models to solve the problem of historical handwritten text recognition is not created out of thin air. The Regensburg University Library found that no existing "off-the-shelf software" can automate the text recognition process of their collection of historical meteorological records from 18th and 19th centuries (Lehenmeier et al., 2020, p. 232). They found that the detection of tabular formats and the recognition of unconstrained layout by different observers hampered the attempt of automation and they concluded that a customized workflow is needed to handle such a job. Through AI, especially deep learning, these techniques have made great progress in recent years (Lehenmeier et al., 2020); therefore, it is possible that AI will revolutionize the field of automated text recognition.

An important part of OCR, in addition to character recognition, is layout analysis. Layout analysis is the detection of text blocks or regions and their processing into a format recognizable by the recognition algorithm (Neudecker et al., 2019). Many studies that transcribe historical documents have incorporated segmentation as their layout analysis. For example, Chamchong et al. (2019), Gatos et al. (2014), and Neudecker et al. (2019) have segmented their historical text documents into lines for further recognition. Aside from text documents, studies have also dealt with tabular documents. Dahl et al. (2021) and Lehenmeier et al. (2020) first detect the table and then segment the table cells. Deep learning is a widely used solution for layout analysis, such as TableNet (Paliwal et al., 2019), CascadeTabNet (Prasad et al., 2020), and LayoutParser (Shen et al., 2021).

Layout analysis is often overlooked, probably due to the all-in-one solutions offered by OCR packages. Similarly, current research divides the required techniques (e.g., text detection, text recognition, layout detection, layout analysis) into sub-problems and focuses on improving each

sub-problem rather than having them working together. For example, Dahl et al. (2021) stated that current studies focus on improving benchmarking algorithms separately, without considering that these models (e.g., layout analysis and text recognition) might eventually need to work in unison. They argued that these algorithms would need to work together at some point, and such perception would make the study of benchmarking algorithms less practical. Similarly, several studies demonstrated the need for layout analysis to work with text recognition algorithms as a pipeline. Shen et al. (2021) said in their study that character recognition engines such as Tesseract, easyOCR, and paddleOCR are not equipped to do the work of layout analysis. Similar to a preliminary study conducted by Lehenmeier et al. (2020), researchers found that transcription tools such as Transkribus, Tesseract 4, and OCRopy were unable to detect the table and locate the text in their sample pages. However, for text recognition algorithms, it is crucial to know the location of the text. Therefore, layout analysis is considered as a key step to locate texts and separate them from the page before text recognition (Fischer et al., 2014; Namboodiri & Jain, 2007). Both steps are important to automated text recognition.

Many studies have proposed workflows that incorporated the combination of layout analysis and text recognition. In a research of automatic transcription of ancient Thai handwritten manuscripts, researchers generated a workflow that first performed a line and character segmentation on each page and then fed the resulting text block/line images to the recognition algorithm (Chamchong et al., 2019). Likewise, a similar workflow was provided by researchers from the OCR-D project, a study that recognizes historical printed documents in German (Neudecker et al., 2019). Their workflow is also a pipeline that includes layout analysis, where they optimize the original images and segment them into text lines, deep learning-enhanced text recognition, and post-editing of transcribed text. Furthermore, to automate the handwritten tabular recording of historical weather data in Europe, Lehenmeier et al. (2020) proposed a similar deep learning-assisted workflow that includes preprocessing, layout analysis (i.e., text line segmentation), recognition, and postprocessing. It can be concluded that layout analysis (e.g., image segmentation) and recognition are two indispensable steps of end-to-end automated transcription, and that layout analysis is always done before recognition. Future implementation should consider this finding and refine their approaches.

2.6 Conclusions

As the application of AI in historical records transcription is still in its infancy, there remain many research gaps to be filled. The first step in addressing these gaps will be to understand how researchers who may apply this integration in future research perceive the risks and challenges as well as the opportunities of incorporating AI. Since they are at the forefront of AI integration, it would be beneficial to understand their willingness or reluctance and address this in future projects. Second, it would be beneficial to explore what an AI-augmented data transcription workflow looks like and how it performs on historical handwritten records. This will provide opportunities to test the effectiveness of integrating AI with historical records transcription and inform future attempts.

References

Alabau, V., & Leiva, L. (2012). Transcribing handwritten text images with a word soup game. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, 2273–2278. https://doi.org/10.1145/2212776.2223788

Alcoforado, M. J., Vaquero, J. M., Trigo, R. M., & Taborda, J. P. (2012). Early Portuguese meteorological measurements (18th century). *Climate of the Past*, *8*(1), 353–371. https://doi.org/10.5194/cp-8-353-2012

Allan, R., Brohan, P., Compo, G. P., Stone, R., Luterbacher, J., & Brönnimann, S. (2011). The International Atmospheric Circulation Reconstructions over the Earth (ACRE) Initiative. *Bulletin of the American Meteorological Society*, *92*(11), 1421–1425. https://doi.org/10.1175/2011BAMS3218.1

Allan, R., Compo, G., & Carton, J. (2011). Recovery of Global Surface Weather Observations for Historical Reanalyses and International Users. *Eos, Transactions American Geophysical Union*, *92*(18), 154–154. https://doi.org/10.1029/2011EO180008

Allan, R., Wood, K., Freeman, E., Wilkinson, C., Andersson, A., Lorrey, A., Brohan, P., Stendel, M., & Kennedy, J. (2021). Learning from the past to understand the future: Historical records of change in the ocean. *WMO Bull*, *70*, 36–42.

Ashcroft, L., Allan, R., Bridgman, H., Gergis, J., Pudmenzky, C., & Thornton, K. (2016). Current climate data rescue activities in Australia. *Advances in Atmospheric Sciences*, *33*(12), 1323–1324. https://doi.org/10.1007/s00376-016-6189-5

Ashcroft, L., Coll, J. R., Gilabert, A., Domonkos, P., Brunet, M., Aguilar, E., Castella, M., Sigro, J., Harris, I., Unden, P., & Jones, P. (2018). A rescued dataset of sub-daily

meteorological observations for Europe and the southern Mediterranean region, 1877–2012. *Earth System Science Data*, *10*(3), 1613–1635. https://doi.org/10.5194/essd-10-1613-2018

Ashcroft, L., Gergis, J., & Karoly, D. J. (2014). A historical climate dataset for southeastern Australia, 1788–1859. *Geoscience Data Journal*, *1*(2), 158–178. https://doi.org/10.1002/gdj3.19

Tafti, A. P., Baghaie, A., Assefi, M., Arabnia, H. R., Yu, Z., & Peissig, P. (2016). OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, & T. Isenberg (Eds.), *Advances in Visual Computing* (pp. 735–746). Springer International Publishing. https://doi.org/10.1007/978-3-319-50835-1_66

Batson, C. D., Ahmad, N., & Tsang, J.-A. (2002). Four Motives for Community Involvement. *Journal of Social Issues*, *58*(3), 429–445. https://doi.org/10.1111/1540-4560.00269

Bauer, M. W., Petkova, K., & Boyadjieva, P. (2016). Public Knowledge of and Attitudes to Science: Alternative Measures That May End the "Science War": *Science, Technology, & Human Values*. https://doi.org/10.1177/016224390002500102

Biondich, P. G., Overhage, J. M., Dexter, P. R., Downs, S. M., Lemmon, L., & McDonald, C. J. (2002). A modern optical character recognition system in a real world clinical setting: Some accuracy and feasibility observations. *Proceedings of the AMIA Symposium*, 56–60.

Black, A., & Law, F. (2004). Development and Utilization of a National Web-Based Chronology of Hydrological Events. *Hydrological Sciences Journal/Journal Des Sciences Hydrologiques*, *49*, 246. https://doi.org/10.1623/hysj.49.2.237.34835

Blancq, F. L. (2010). Rescuing old meteorological data. *Weather*, *65*(10), 277–280. https://doi.org/10.1002/wea.510

Brohan, P. (2012). oldWeather.org: Citizen Science for Climate Reconstruction. *AGU Fall Meeting Abstracts*, *2012*, ED53A-0922. https://ui.adsabs.harvard.edu/abs/2012AGUFMED53A0922B

Brohan, P. (2017, August 17). *RealClimate: Data rescue projects*. https://www.realclimate.org/index.php/archives/2017/08/data-rescue-projects/

Brohan, P., Allan, R., Freeman, J. E., Waple, A. M., Wheeler, D., Wilkinson, C., & Woodruff, S. (2009). Marine Observations of Old Weather. *Bulletin of the American Meteorological Society*, *90*(2), 219–230. https://doi.org/10.1175/2008BAMS2522.1

Brohan, P., Ward, C., Willetts, G., Wilkinson, C., Allan, R., & Wheeler, D. (2010). Arctic marine climate of the early nineteenth century. *Climate of the Past*, *6*(3), 315–324.

https://doi.org/10.5194/cp-6-315-2010

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R., Brugnara, Y., Brunet, M., Brunetti, M., Chimani, B., Cornes, R., Domínguez-Castro, F., Filipiak, J., Founda, D., Herrera, R. G., Gergis, J., Grab, S., Hannak, L., Huhtamaa, H., … Wyszyński, P. (2019). Unlocking Pre-1850 Instrumental Meteorological Records: A Global Inventory. *Bulletin of the American Meteorological Society*, *100*(12), ES389–ES413. https://doi.org/10.1175/BAMS-D-19-0040.1

Brönnimann, S., Annis, J., Dann, W., Ewen, T., Grant, A. N., Griesser, T., Krähenmann, S., Mohr, C., Scherer, M., & Vogler, C. (2006). A guide for digitising manuscript climate data. *Climate of the Past*, *2*(2), 137–144. https://doi.org/10.5194/cp-2-137-2006

Brönnimann, S., Brugnara, Y., Allan, R. J., Brunet, M., Compo, G. P., Crouthamel, R. I., Jones, P. D., Jourdain, S., Luterbacher, J., Siegmund, P., Valente, M. A., & Wilkinson, C. W. (2018). A roadmap to climate data rescue services. *Geoscience Data Journal*, *5*(1), 28–39. https://doi.org/10.1002/gdj3.56

Brossard, D., Lewenstein, B., & Bonney, R. (2005). Scientific knowledge and attitude change: The impact of a citizen science project. *International Journal of Science Education*, *27*(9), 1099–1121. https://doi.org/10.1080/09500690500069483

Brunet, M., Gilabert, A., Jones, P., & Efthymiadis, D. (2014). A historical surface climate dataset from station observations in Mediterranean North Africa and Middle East areas. *Geoscience Data Journal*, *1*(2), 121–128. https://doi.org/10.1002/gdj3.12

Brunet, M., & Jones, P. (2011). Data rescue initiatives: Bringing historical climate data into the 21st century. *Climate Research*, *47*(1), 29–40. https://doi.org/10.3354/cr00960

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Camuffo, D., & Bertolin, C. (2012). The earliest temperature observations in the world: The Medici Network (1654–1670). *Climatic Change*, *111*(2), 335–363. https://doi.org/10.1007/s10584-011-0142-5

Camuffo, D., Bertolin, C., Diodato, N., Cocheo, C., Barriendos, M., Dominguez-Castro, F., Garnier, E., Alcoforado, M. J., & Nunes, M. F. (2013). Western Mediterranean precipitation over the last 300 years from instrumental observations. *Climatic Change*, *117*(1), 85–101. https://doi.org/10.1007/s10584-012-0539-9

Capozzi, V., Cotroneo, Y., Castagno, P., De Vivo, C., & Budillon, G. (2020). Rescue and quality control of sub-daily meteorological data collected at Montevergine Observatory (Southern Apennines), 1884–1963. *Earth System Science Data*, *12*(2), 1467–1487. https://doi.org/10.5194/essd-12-1467-2020

Ceccaroni, L., Bibby, J., Roger, E., Flemons, P., Michael, K., Fagan, L., & Oliver, J. (2019). Opportunities and Risks for Citizen Science in the Age of Artificial Intelligence. *Citizen Science: Theory and Practice, 4*. https://doi.org/10.5334/cstp.241

Chamchong, R., Gao, W., & McDonnell, M. D. (2019). Thai Handwritten Recognition on Text Block-Based from Thai Archive Manuscripts. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1346–1351. https://doi.org/10.1109/ICDAR.2019.00217

Chen, J., Riba, P., Fornés, A., Mas, J., Lladós, J., & Pujadas-Mora, J. M. (2018). Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 528–533. https://doi.org/10.1109/ICFHR-2018.2018.00098

Chimani, B., Auer, I., Prohom, M., Nadbath, M., Paul, A., & Rasol, D. (2021). Data rescue in selected countries in connection with the EUMETNET DARE activity. *Geoscience Data Journal*, *9*(1), 187–200. https://doi.org/10.1002/gdj3.128

Cornes, R. C., Jones, P. D., Briffa, K. R., & Osborn, T. J. (2012). A daily series of mean sea-level pressure for London, 1692–2007. *International Journal of Climatology*, *32*(5), 641–656. https://doi.org/10.1002/joc.2301

Craig, P. M., & Hawkins, E. (2020). Digitizing observations from the Met Office Daily Weather Reports for 1900–1910 using citizen scientist volunteers. *Geoscience Data Journal*, *7*(2), 116–134. https://doi.org/10.1002/gdj3.93

Dahl, C. M., Johansen, T. S. D., Sørensen, E. N., Westermann, C. E., & Wittrock, S. F. (2021). Applications of Machine Learning in Document Digitisation. *ArXiv:2102.03239 [Cs, Econ, Stat]*. http://arxiv.org/abs/2102.03239

Davis, A. K., Nibbelink, N. P., & Howard, E. (2012). Identifying large-and small-scale habitat characteristics of monarch butterfly migratory roost sites with citizen science observations. *International Journal of Zoology*, *2012*, e149026. https://doi.org/10.1155/2012/149026

de Sherbinin, A., Bowser, A., Chuang, T.-R., Cooper, C., Danielsen, F., Edmunds, R., Elias, P., Faustman, E., Hultquist, C., Mondardini, R., Popescu, I., Shonowo, A., & Sivakumar, K. (2021). The Critical Importance of Citizen Science Data. *Frontiers in Climate*, *3*. https://doi.org/10.3389/fclim.2021.650760

Domínguez-Castro, F., Vaquero, J. M., Rodrigo, F. S., Farrona, A. M. M., Gallego, M. C., García-Herrera, R., Barriendos, M., & Sanchez-Lorenzo, A. (2014). Early Spanish meteorological records (1780–1850). *International Journal of Climatology*, *34*(3), 593–603. https://doi.org/10.1002/joc.3709

Dupigny-Giroux, L.-A., Ross, T. F., Elms, J. D., Truesdell, R., & Doty, S. R. (2007). NOAA's

Climate Database Modernization Program: Rescuing, Archiving, and Digitizing History. *Bulletin of the American Meteorological Society*, *88*(7), 1015–1017. https://doi.org/10.1175/BAMS-88-7-1015

Eitzel, M. V., Cappadonna, J. L., Santos-Lang, C., Duerr, R. E., Virapongse, A., West, S. E., Kyba, C. C. M., Bowser, A., Cooper, C. B., Sforzi, A., Metcalfe, A. N., Harris, E. S., Thiel, M., Haklay, M., Ponciano, L., Roche, J., Ceccaroni, L., Shilling, F. M., Dörler, D., … Jiang, Q. (2017). Citizen Science Terminology Matters: Exploring Key Terms. *Citizen Science: Theory and Practice*, *2*(1), Article 1. https://doi.org/10.5334/cstp.96

Eveleigh, A., Jennett, C., Blandford, A., Brohan, P., & Cox, A. L. (2014). Designing for dabblers and deterring drop-outs in citizen science. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2985–2994. https://doi.org/10.1145/2556288.2557262

Firmani, D., Merialdo, P., Nieddu, E., & Scardapane, S. (2017). In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks. *Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage (AI*CH 2017)*, 9–16.

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., & Ingold, R. (2014). *The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries* (pp. 91–106). https://doi.org/10.13140/2.1.2180.3526

Follett, R., & Strezov, V. (2015). An Analysis of Citizen Science Based Research: Usage and Publication Patterns. *PLOS ONE*, *10*(11), e0143687. https://doi.org/10.1371/journal.pone.0143687

Fornés, A., Megyesi, B., & Romeu, J. M. (2017). Transcription of Encoded Manuscripts with Image Processing Techniques. *DH*.

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, K., & Wallin, J. (2011). Galaxy Zoo: Morphological Classification and Citizen Science. *ArXiv:1104.5513 [Astro-Ph]*. http://arxiv.org/abs/1104.5513

Franzen, M., Kloetzer, L., Ponti, M., Trojan, J., & Vicens, J. (2021). Machine Learning in Citizen Science: Promises and Implications. In K. Vohland, A. Land-Zandstra, L. Ceccaroni, R. Lemmens, J. Perelló, M. Ponti, R. Samson, & K. Wagenknecht (Eds.), *The Science of Citizen Science* (pp. 183–198). Springer International Publishing. https://doi.org/10.1007/978-3-030-58278-4_10

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J. A., Toselli, A. H., & Vidal, E. (2014). Ground-Truth Production in the Transcriptorium Project. *2014 11th IAPR International Workshop on Document Analysis Systems*, 237–241. https://doi.org/10.1109/DAS.2014.23

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2009). A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(5), 855–868. https://doi.org/10.1109/TPAMI.2008.137

Graves, A., & Schmidhuber, J. (2008). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, *21*. https://proceedings.neurips.cc/paper/2008/hash/66368270ffd51418ec58bd793f2d9b1b-Abstract.html

Green, S. E., Rees, J. P., Stephens, P. A., Hill, R. A., & Giordano, A. J. (2020). Innovations in Camera Trapping Technology and Approaches: The Integration of Citizen Science and Artificial Intelligence. *Animals*, *10*(1), Article 1. https://doi.org/10.3390/ani10010132

Gura, T. (2013). Citizen science: Amateur experts. *Nature*, *496*(7444), Article 7444. https://doi.org/10.1038/nj7444-259a

Hanson, D., & Simenstad, A. (2018). Combining Human and Machine Transcriptions on the Zooniverse Platform. *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-Generated Text*, 215–216. https://doi.org/10.18653/v1/W18-6129

Harnett, J. (1985, November 14). Developments in OCR for automatic data entry. *Proceedings of Translating and the Computer 7*. TC 1985, London, UK. https://aclanthology.org/1985.tc-1.12

Hawkins, E., Burt, S., Brohan, P., Lockwood, M., Richardson, H., Roy, M., & Thomas, S. (2019). Hourly weather observations from the Scottish Highlands (1883–1904) rescued by volunteer citizen scientists. *Geoscience Data Journal*, *6*(2), 160–173. https://doi.org/10.1002/gdj3.79

Holley, R. (2009). How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, *15*(3/4). https://doi.org/10.1045/march2009-holley

Howard, E., Aschen, H., & Davis, A. K. (2010). Citizen Science Observations of Monarch Butterfly Overwintering in the Southern United States. *Psyche: A Journal of Entomology*, *2010*, e689301. https://doi.org/10.1155/2010/689301

Jander, M. (2016). Handwritten Text Recognition—Transkribus: A User Report. *ETRAP Research Group, Institute of Computer Science, University of G¨ Ottingen, Germany*, 3.

Jenckel, M., Bukhari, S. S., & Dengel, A. (2016). anyOCR: A sequence learning based OCR system for unlabeled historical documents. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 4035–4040. https://doi.org/10.1109/ICPR.2016.7900265

Kwok, R. (2017). Historical data: Hidden in the past. *Nature*, *549*(7672), 419–421.

https://doi.org/10.1038/nj7672-419

Lehenmeier, C., Burghardt, M., & Mischka, B. (2020). Layout Detection and Table Recognition – Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital Libraries for Open Knowledge* (pp. 229–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-54956-5_17

Levitus, S. (1992). *National Oceanographic Data Center Inventory of Physical Oceanographic Profiles: Global Distributions by Year for All Countries*. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service.

Levitus, S. (1996). Data Archaeology and Rescue of Historical Oceanographic Data: A Report on "The IOC/IODE GODAR Project." In *NOAA Technical Report NESDIS 87. Proceedings of The International Workshop on Oceanographic Biological and Chemical Data Management.*

Li, Q., An, W., Zhou, A., & Ma, L. (2016). Recognition of Offline Handwritten Chinese Characters Using the Tesseract Open Source OCR Engine. *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, *02*, 452–456. https://doi.org/10.1109/IHMSC.2016.239

Lotfian, M., Ingensand, J., & Brovelli, M. A. (2021). The Partnership of Citizen Science and Machine Learning: Benefits, Risks, and Future Challenges for Engagement, Data Collection, and Data Quality. *Sustainability*, *13*(14), Article 14. https://doi.org/10.3390/su13148087

Mateus, C., Potito, A., & Curley, M. (2021). Engaging secondary school students in climate data rescue through service-learning partnerships. *Weather*, *76*(4), 113–118. https://doi.org/10.1002/wea.3841

McClure, E. C., Sievers, M., Brown, C. J., Buelow, C. A., Ditria, E. M., Hayes, M. A., Pearson, R. M., Tulloch, V. J. D., Unsworth, R. K. F., & Connolly, R. M. (2020). Artificial Intelligence Meets Citizen Science to Supercharge Ecological Monitoring. *Patterns*, *1*(7), 100109. https://doi.org/10.1016/j.patter.2020.100109

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, *8*, 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

Namboodiri, A. M., & Jain, A. K. (2007). Document structure and layout analysis. In B. B. Chaudhuri (Ed.), *Digital Document Processing: Major Directions and Recent Advances* (pp. 29–48). Springer. https://doi.org/10.1007/978-1-84628-726-8_2

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., &

Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 53–58. https://doi.org/10.1145/3322905.3322917

Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A Survey of Historical Document Image Datasets. *ArXiv:2203.08504 [Cs]*. http://arxiv.org/abs/2203.08504

Odunayo, O., Sookoo, N. N., Bathla, G., Cavallin, A., Persaud, B. D., Szigeti, K., Van Cappellen, P., & Lin, J. (2021). Rescuing historical climate observations to support hydrological research: A case study of solar radiation data. *Proceedings of the 21st ACM Symposium on Document Engineering*, 1–4. https://doi.org/10.1145/3469096.3474929

Paliwal, S. S., D, V., Rahul, R., Sharma, M., & Vig, L. (2019). TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 128–133. https://doi.org/10.1109/ICDAR.2019.00029

Parthiban, R., Ezhilarasi, R., & Saravanan, D. (2020). Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–5. https://doi.org/10.1109/ICSCAN49426.2020.9262379

Ponti, M., & Seredko, A. (2022). Human-machine-learning integration and task allocation in citizen science. *Humanities and Social Sciences Communications*, *9*(1), Article 1. https://doi.org/10.1057/s41599-022-01049-z

Prasad, D., Gadpal, A., Kapadni, K., Visave, M., & Sultanpure, K. (2020). CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2439–2447. https://doi.org/10.1109/CVPRW50498.2020.00294

Raddick, M. J., Bracey, G., Gay, P. L., Lintott, C. J., Murray, P., Schawinski, K., Szalay, A. S., & Vandenberg, J. (2009). Galaxy zoo: Exploring the motivations of citizen science volunteers. *ArXiv Preprint ArXiv:0909.2925*. https://doi.org/10.48550/arXiv.0909.2925

Rafner, J., Gajdacz, M., Kragh, G., Hjorth, A., Gander, A., Palfi, B., Berditchevskaia, A., Grey, F., Gal, K., Segal, A., Walmsley, M., Miller, J. A., Dellerman, D., Haklay, M., Michelucci, P., & Sherson, J. (2021). *Revisiting Citizen Science Through the Lens of Hybrid Intelligence* (arXiv:2104.14961). arXiv. https://doi.org/10.48550/arXiv.2104.14961

Rakshit, S., Kundu, A., Maity, M., Mandal, S., Sarkar, S., & Basu, S. (2010). Recognition of handwritten Roman Numerals using Tesseract open source OCR engine. *ArXiv:1003.5898 [Cs]*. http://arxiv.org/abs/1003.5898

Rambonnet, L., Vink, S. C., Land-Zandstra, A. M., & Bosker, T. (2019). Making citizen science

count: Best practices and challenges of citizen science projects on plastics in aquatic environments. *Marine Pollution Bulletin*, *145*, 271–277. https://doi.org/10.1016/j.marpolbul.2019.05.056

Rotman, D., Hammock, J., Preece, J., Boston, C., Hansen, D., Bowser, A., & He, Y. (2014). *Does motivation in citizen science change with time and culture?* 229–232. https://doi.org/10.1145/2556420.2556492

Ryan, C., Duffy, C., Broderick, C., Thorne, P. W., Curley, M., Walsh, S., Daly, C., Treanor, M., & Murphy, C. (2018). Integrating Data Rescue into the Classroom. *Bulletin of the American Meteorological Society*, *99*(9), 1757–1764. https://doi.org/10.1175/BAMS-D-17-0147.1

Ryan, C., Murphy, C., McGovern, R., Curley, M., & Walsh, S. (2021). Ireland's pre-1940 daily rainfall records. *Geoscience Data Journal*, *8*(1), 11–23. https://doi.org/10.1002/gdj3.103

Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K., & de Does, J. (2014). Handwritten text recognition for historical documents in the transcriptorium project. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 111–117. https://doi.org/10.1145/2595188.2595193

Sánchez, J. A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R., Vidal, E., & de Does, J. (2013). *tranScriptorium: A european project on handwritten text recognition*. 227–228. https://doi.org/10.1145/2494266.2494294

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. (2019). A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition*, *94*, 122–134. https://doi.org/10.1016/j.patcog.2019.05.025

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *ArXiv:2103.15348 [Cs]*. http://arxiv.org/abs/2103.15348

Sieber, R., & Slonosky, V. (2019). Developing a Flexible Platform for Crowdsourcing Historical Weather Records. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *52*(3), 164–177. https://doi.org/10.1080/01615440.2018.1558138

Sieber, R., Slonosky, V., Ashcroft, L., & Pudmenzky, C. (2022). Formalizing Trust in Historical Weather Data. *Weather, Climate, and Society*, *14*(3), 993–1007. https://doi.org/10.1175/WCAS-D-21-0077.1

Singh, A., Bacchuwar, K., & Bhasin, A. (2012). A Survey of OCR Applications. *International Journal of Machine Learning and Computing*, 314–318. https://doi.org/10.7763/IJMLC.2012.V2.137

Slonosky, V. (2014). Historical climate observations in Canada: 18th and 19th century daily temperature from the St. Lawrence Valley, Quebec. *Geoscience Data Journal*, *1*(2), 103–

120. https://doi.org/10.1002/gdj3.11

Slonosky, V., Sieber, R., Burr, G., Podolsky, L., Smith, R., Bartlett, M., Park, E., Cullen, J., & Fabry, F. (2019). From books to bytes: A new data rescue tool. *Geoscience Data Journal*, *6*(1), 58–73. https://doi.org/10.1002/gdj3.62

Stickler, A., Brönnimann, S., Jourdain, S., Roucaute, Eméline, Sterin, Alexander M, Nikolaev, Dmitrii, Valente, Maria Antónia, Wartenburger, Richard, Hersbach, Hans, Ramella Pralungo, Lorenzo, & Dee, Dick P. (2014). *ERA-CLIM Historical Upper-Air Data 1900-1972, supplement to: Stickler, Alexander; Brönnimann, Stefan; Jourdain, Sylvie; Roucaute, Eméline; Sterin, Alexander M; Nikolaev, Dmitrii; Valente, Maria Antónia; Wartenburger, Richard; Hersbach, Hans; Ramella Pralungo, Lorenzo; Dee, Dick P (2014): Description of the ERA-CLIM historical upper-air data. Earth System Science Data, 6(1), 29-48* [Application/zip]. 813 datasets. https://doi.org/10.1594/PANGAEA.821222

Stickler, A., Brönnimann, S., Valente, M. A., Bethke, J., Sterin, A., Jourdain, S., Roucaute, E., Vasquez, M. V., Reyes, D. A., Allan, R., & Dee, D. (2014). ERA-CLIM: Historical Surface and Upper-Air Data for Future Reanalyses. *Bulletin of the American Meteorological Society*, *95*(9), 1419–1430. https://doi.org/10.1175/BAMS-D-13-00147.1

Ströbel, P. B., Clematide, S., Volk, M., & Hodel, T. (2022). Transformer-based HTR for Historical Documents. *ArXiv:2203.11008 [Cs]*. http://arxiv.org/abs/2203.11008

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., Farnsworth, A., Fink, D., Fitzpatrick, J. W., Fredericks, T., Gerbracht, J., Gomes, C., Hochachka, W. M., Iliff, M. J., Lagoze, C., La Sorte, F. A., … Kelling, S. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. https://doi.org/10.1016/j.biocon.2009.05.006

Swindall, M. I., Croisdale, G., Hunter, C. C., Keener, B., Williams, A. C., Brusuelas, J. H., Krevans, N., Sellew, M., Fortson, L., & Wallin, J. F. (2021). Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data. *2021 IEEE 17th International Conference on EScience (EScience)*, 128–137. https://doi.org/10.1109/eScience51609.2021.00023

Tan, L. S., Burton, S., Crouthamel, R., van Engelen, A., Hutchinson, R., Nicodemus, L., Peterson, T. C., Rahimzadeh, F., Llansó, P., & Kontongomde, H. (2004). Guidelines on Climate Data Rescue. *World Meteorological Organization, WMO/TD No. 1210*.

http://www.wmo.int/pages/prog/wcp/wcdmp/documents/WCDMP-55.pdf

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections*, 179–204. https://doi.org/10.1515/9783839455845-008

Torney, C. J., Lloyd-Jones, D. J., Chevallier, M., Moyer, D. C., Maliti, H. T., Mwita, M., Kohi, E. M., & Hopcraft, G. C. (2019). A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, *10*(6), 779–787. https://doi.org/10.1111/2041-210X.13165

Veale, L., Endfield, G., Davies, S., Macdonald, N., Naylor, S., Royer, M.-J., Bowen, J., Tyler-Jones, R., & Jones, C. (2017). Dealing with the deluge of historical weather data: The example of the TEMPEST database. *Geo: Geography and Environment*, *4*(2), e00039. https://doi.org/10.1002/geo2.39

Wilkinson, C., Brönnimann, S., Jourdain, S., Roucaute, E., Crouthamel, R., Brohan, P., Valente, A., Brugnara, Y., Brunet, M., & Team, I. (2019). *Best Practice Guidelines for Climate Data Rescue*.

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

World Meteorological Organization. (2016). *Guidelines on Best Practices for Climate Data Rescue*. https://library.wmo.int/doc_num.php?explnum_id=3318

Yasser, A. M., Clawson, K., & Bowerman, C. (2017). *Saving Cultural Heritage with Digital Make-Believe: Machine Learning and Digital Techniques to the Rescue*. https://doi.org/10.14236/ewic/HCI2017.97

Ziomek, J., & Middleton, S. E. (2021). GloSAT Historical Measurement Table Dataset: Enhanced Table Structure Recognition Annotation for Downstream Historical Data Rescue. *The 6th International Workshop on Historical Document Imaging and Processing*, 49–54. https://doi.org/10.1145/3476887.3476890

Preface to Chapter 3

Chapter 3 presents a survey study to address the first research question: “How do researchers and practitioners perceive the challenges and opportunities of using AI-augmented data rescue?” This chapter was motivated by the literature review presented in Chapter 2, which revealed research gaps in understanding the opportunities and challenges of involving AI technology in the data rescue process.

This chapter was co-authored with my supervisor, Dr. Renee Sieber, as a peer-reviewed journal article. We plan to submit the manuscript to *Citizen Science: Challenges and Opportunities*.

Chapter 3. A survey on attitude and perception of AI-augmented data rescue among leaders of data rescue community

Abstract

Artificial intelligence (AI) has been applied in many fields, making once-impossible studies possible. Data rescue is one of the fields that is experimenting with AI to enhance transcription of historical records, and it has the potential to benefit greatly from the use of AI. However, AI-enhanced transcription is a relatively new approach to data rescue, so its advantages and disadvantages have not been thoroughly investigated. It would help future attempts at AI-enhanced data rescue if we know what are the opinions and expectations of researchers. There is minimal research on the attitude and the opinion researchers have towards using AI-augmented approach in historical data rescue. Therefore, a survey is conducted among the data rescue community, including citizen science researchers who are also doing the transcription work. In my study, I find that the majority of respondents are willing to try automation, but they hold concerns such as output accuracy and funding availability. These concerns about AI are related to the issue of handwriting materials. The average time respondents are willing to invest in building an AI-enhanced project is 2 ½ months; in other words, if it takes longer, respondents may be reluctant to try. They feel that there is a trade-off between public participation and opportunities for more data. Rather than pure automation, most are in favour of a hybrid model where humans supervise the AI. The result of the survey could form a guideline to future attempts on automation in data rescue, and it can also assist a smoother shift from manual transcription to automation.

3.1 Introduction

With increasing demands of higher quality and continuous climate data, data rescue has attracted more attention and is given a higher priority in climate research over the past decade (Brönnimann et al., 2018). The increasing demand of past climate data are partially due to the new techniques which enables the reconstruction of past weather patterns that was not previously

possible. Blancq (2010, p.280) described this rescue as "adding small pieces to the great jigsaw of our past weather and climate". Data rescue is also driven by the changing focus of climate research where past knowledge can be used to qualify the ongoing climate changes (Blancq, 2010). Researchers are starting to realize that past works are of great value, as Andrew Trant, an ecologist at the University of Waterloo in Canada, comments: "[t]here are so many stories that are locked away in historical data" (Kwok, 2017, p. 420). However, researchers cannot use historical records as it is. The obstacle they are constantly facing is that much of the historical records remain in manuscript form and usually pre-date the age of electronic data acquisition (Brönnimann et al., 2018). This means digitization as part of the data rescue process is necessary before researchers put historical records into use (World Meteorological Organization, 2016).

Great effort has been made worldwide in rescuing historical climate records by initiatives such as the International Atmospheric Circulation Reconstructions Over the Earth (ACRE) and the International Data Rescue Portal (I-DARE). Recent data rescue efforts incorporated citizen scientists as volunteers to transcribe historical climate records through web-based tools such as Data Rescue: Archives & Weather (DRAW), Old Weather, and Rainfall Rescue. University and post-secondary students have also been engaged in data rescue as part of their assignment (Mateus et al., 2021; Ryan et al., 2018). Not only in climate, historical records and data rescue are also important in many other fields such as ecology and astronomy. Astroinformatics, a field involved historical astronomy records, has emerged; Astrid Ogilvie, a climate historian, asks "[w]hy wouldn't we use every scrap of information we can get" (Kwok, 2017, p. 421)? However, those efforts have been facing the same problem. As mentioned above, digitization is an important and necessary step of data rescue, and this step can be a time-consuming and labor-intensive task where majority of the records are manually keyed in. This means transcribers need to be hired or volunteers need to be recruited most of the time to transcribe historical records. Besides, manually keying in requires a lot of person-hours. Therefore, researchers have been finding a way to fix this problem.

Data rescue is beyond historical weather records. People are transcribing poems, diaries of traders, civil and other wars (e.g., Anzac records[10]). So data rescue can be broadly thought of as

[10] https://discoveringanzacs.naa.gov.au/

digitization of historical records. The applications can be very different, but the challenges can be quite similar, for example, handling cursive handwritings, unstructured documents, and faded inks.

Automation incorporated with AI (Artificial Intelligence) is one solution that the community is experimenting with to optimize the time and person-hour spent on manual data rescue. AI has received growing attention in different fields and has provided great opportunities to multiple disciplines (Collins et al., 2021). The great availability and easy accessibility have contributed to the great advancement in AI methods since most of the algorithms have been made open source. Cloud-based services with extensive computational power have also made AI methods affordable to everyone. Therefore, there are emerging attempts of adapting AI in the field of data rescue (World Meteorological Organization, 2016). The most frequently cited source defined AI as something that "enables the machine to exhibit human intelligence, including the ability to perceive, reason, learn, and interact, etc." (Russell & Norvig, 2020). In the case of data rescue, researchers have been testing to see if AI can be used to learn and transcribe historical records with similar performance to human transcribers. If AI can be successfully adapted in data rescue projects, the great number of person-hours spent would no longer be a problem, and those person-hours can be used more effectively in some other steps.

If properly installed, AI-augmented data rescue can ideally be cheaper and more efficient considering it is open source and automated (Wilkinson et al., 2019). It is not here to replace the role of hired transcribers or volunteer citizen scientists, but it could alter the way their work, transcription, is done. AI could be an opportunity for improvement in data rescue, and it can be very useful as well. Before adapting AI in data rescue, it is crucial to know researchers' attitudes towards AI, and how well are they responding to the impact of new technology. By attending to researchers' thoughts, AI-augmented data rescue can be widely accepted, improved, and promoted.

However, people tend to hold different opinions towards AI, and their attitude of adapting AI are different as well. One survey on America's public perception of AI showed that they expressed mixed support for AI development (B. Zhang & Dafoe, 2019). Globally, different regions also expressed different opinions about AI adoption (Neudert et al., 2020). While more than half respondents in East Asia believe AI would be mostly helpful, only 26% of respondents in Latin America and Caribbean believe in its benefits. The opinion of AI has also been widely

researched among medical fields. A survey showed that radiology students are convinced that AI will be greatly involved but they do not think it will eventually replace them (Pinto dos Santos et al., 2019). Similarly, dental students agreed that AI will revolutionize dental practice but will not replace themselves in the near future (Yüzbaşıoğlu, 2021). A majority of radiology and dental students are willing to improve their knowledge in AI, and three-fourth of dental students are excited about using AI in dentistry. However, other discussions have shown some concerns about radiologists' future career being replaced by AI (Pinto dos Santos et al., 2019; Yüzbaşıoğlu, 2021). The above research showed a comprehensive knowledge of respondents' attitudes and opinion towards AI, and I believe it is important to have a thorough understanding before applying AI in any field. There is minimal research or survey that has been done previously on the attitude and opinion of applying AI in rescuing historical data.

The purpose of this study is to evaluate researchers' attitude and opinion on AI-augmented data rescue and their concerns on adapting AI in this field. I will first conclude the past experiences and thoughts on climate data rescue. Then I describe my method, which is an online survey to assess data rescue professionals' opinion on AI in data rescue and evaluate their thoughts of applying AI as part of the solution to manual data rescue. We hope that this study will serve as a guideline for future attempts of automation and document steps for improvement. This guideline should enable us to make a smoother transition from manual to automated data rescue, where obstacles are minimized. It can also help us avoid or overcome the resistance that we may encounter along the way of automation. Ultimately, we hope that the result of this study will facilitate automation in the future by addressing the opinions of the researchers, and therefore the end users.

3.2 Literature review

Researchers have deployed automation as part of their efforts in data rescue. However, it's not easy to determine what opinion they hold towards adapting AI in future projects (what should be addressed as improvements and promotion for future attempts at automated data rescue) without having a thorough review of what has been done in current automated data rescue projects. This section reviews researchers' experience and findings on past data rescue projects that engaged automation.

3.2.1 Experience and comments on automated data rescue

Automated data rescue is not a new concept. It has been tested or adapted in several data rescue projects during recent years. The conclusions and experience of each project using Optical Character Recognition (OCR) are various from case to case as well. Therefore, it is important to review people's experience on past attempts of automation before any suggestion is made.

Time is one measure that researchers talked about a lot while testing the automated data rescue approach. Some researchers considered automated approach as time-consuming (Ashcroft et al., 2018; Blancq, 2010; Craig & Hawkins, 2020); whereas some others considered this approach as fast enough (Brönnimann et al., 2006; Wilkinson et al., 2019). Brönnimann et al. (2006) and Wilkinson et al. (2019) argue that automated approach (e.g., OCR) is fast enough at this point because the machine transcribing speed is many times faster than manually keying in. This argument is made when we disregard all other factors, such as pre-processing, that might be involved while adapting automation. It is true that if we only talk about the transcribing speed, the automated approach saves way more time than the manual approach. However, transcribing is just one of the many steps in automated data rescue. Ashcroft et al. (2018) argued that the huge amount of training time should also be considered when evaluating the time consumption of automated approaches. Besides, a lot of the historical records are handwritten on forms, which is not suitable for OCR without being properly handled. Therefore, Blancq (2010) claimed that it will take a great amount of time to scan and pre-process the documents for the OCR process. It is also suggested that the overall time spent on an OCR project can be considerable when the post-processing quality control step is taken into consideration (Brönnimann et al., 2006). In conclusion, it is arguably true that the OCR approach is many times faster than traditional ways of data rescue (e.g., manually keying in) if we only focus on the transcription process itself. Yet the many steps, such as pre-processing and post-processing, that come with the OCR process can be very time-consuming, which makes the overall speed turn out to be slower.

Cost is another well discussed measure. In terms of the cost associated with automated approaches, the opinion is also twofold. One argued that the OCR approach can be costly when compared to other approaches (Ashcroft et al., 2018) while others suggested that it can be rather cost-efficient (Wilkinson et al., 2019; World Meteorological Organization, 2016). The opinions are opposite because the perspectives are different in terms of how they determine the related

cost. Ashcroft et al. (2018) focused on the development cost while Wilkinson et al. (2019) and World Meteorological Organization (2016) looked at the personnel cost. It is true that, as Ashcroft et al. (2018, p.1618) argued, the "diverse nature of each task' and the training effort needed for each data source make the automation option expensive. It is also true that, as Wilkinson et al. (2019) and World Meteorological Organization (2016) suggested, automated approaches can save a lot of funds on hiring manual transcribers since machines will do the transcription work. When looked at from different perspectives, how researchers determine the cost are also different. As a result, the cost of the automation depends on the balance between money spent on development effort and the money saved from hiring less personnel. It is hard to determine if automation approaches are more cost-efficient, and the situation would be different from project to project.

Yet, another factor of interest to researchers is the accuracy of results. They are concerned about using an automated approach because of its higher error rates. Studies have found that automated approaches, specifically OCR, have higher error rates on average than manual approach, especially on handwritten records (Brönnimann et al., 2006; Craig & Hawkins, 2020; Stickler, Alexander et al., 2014; Wilkinson et al., 2019). It has also been found that the accuracy of automation can vary from page to page, and from source to source (Stickler, Alexander et al., 2014; Wilkinson et al., 2019). Sources or pages that include more complicated data (e.g., measurements that include special characters, data that are not in tabular format) tend to have higher error rates than those containing simpler data. A more complicated layout can also lower the accuracy of automated approach, but it might be greatly improved by giving modest effort on image pre-processing (Wilkinson et al., 2019). Common errors found are typographical errors such as confused digits. There are technical ways already used to improve accuracy. Galaxyzoo, for example, may recommend that transcriptions (here, interpretation of images) be done 10 times (Fortson et al., 2012). If the transcription is done manually, the accuracy can be improved by double entry where all records will be transcribed twice. Brönnimann et al. (2006) argues that you cannot simply employ an algorithm since running the algorithm multiple times would produce the same output. This may be seen as a benefit to data rescue. However, switching to algorithms loses a proven way to improve accuracy. It has been suggested that the higher the quality of data source (i.e., legibility, image resolution, pixel noise, and layout), the higher accuracy of the automation output (Mateus et al., 2021; Odunayo et al., 2021; Stickler, Alexander et al., 2014). Overall accuracy might be improved if data quality can be improved by

high quality scan, pre-processing and layout analysis. rescanning is obviously a lot of work and in some cases impossible because the original documents have been destroyed. It is possible that, with proper implementation, AI could offer a solution to the problem of low accuracy in automated approach.

There is this sense that AI can offer a universal solution and that there is an automated character recognition that will work everywhere. However, experience with automated data rescue points to the fact that automation cannot be applied universally. This approach is sensitive to type of writings and is picky about the properties of data records. Most historical records are presented in three writings: handwriting, typewriting, and printing. Several studies have considered OCR, as one type of automated approach, impossible or not mature enough to handle handwritten records. Brönnimann et al. (2006), Ashcroft et al. (2018) and Craig & Hawkins (2020) found in their studies that OCR is not suitable for handwritten text because of the higher error rate, and they suggested manual approach would be the optimum solution for handwritten records. However, research has also suggested that OCR works relatively well for typewritten and printed records (Stickler, Alexander et al., 2014; World Meteorological Organization, 2016). Therefore, it can be concluded from these studies that OCR are sensitive to the source writing types, and researchers tend to not use automation because OCR brings uncertainty when it comes to handwritten records.

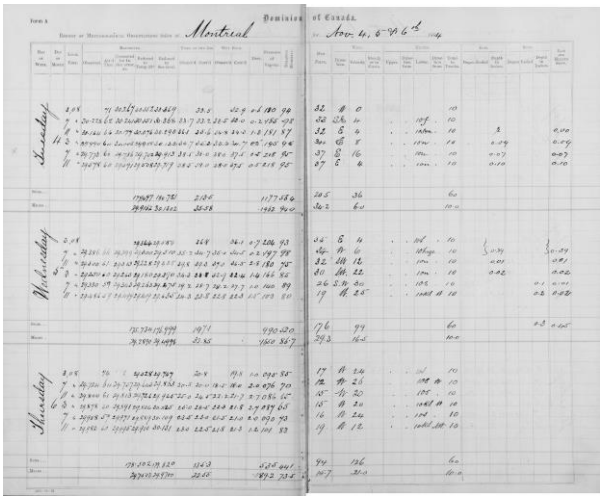
It is important, at this point, to describe what OCR is and isn't because the method is so prevalent and often used as a shorthand for any automation of these kinds of records. The automation should be a workflow that takes in the scanned records and outputs the transcribed documents in their original form. OCR is only one, but essential part of this process, which turns text images into machine-readable characters. It is training dataset dependent, but the training dataset is very difficult to build (Sánchez et al., 2019; Terras, 2022). Part of the reason is that there are lots of languages and different handwriting styles in the world, which means a huge training dataset. This certainly makes transcription of historical records with various handwriting styles difficult.

Two weather-related studies have shown that OCR is almost as good at reading handwritten records as typewritten or printed records. In the ERA-CLIM project, researchers successfully used OCR to transcribe handwritten, typewritten and printed records (Stickler, Alexander et al., 2014)). It is shown that OCR are capable of transcribing handwritten records, and the accuracy

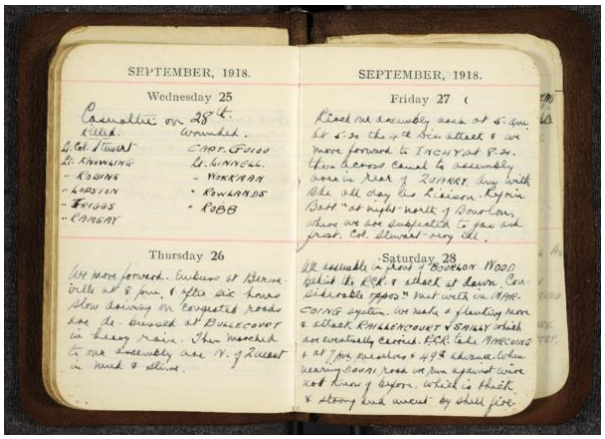
of the output depends much on the quality of the image source and its layout. Another trial in 2018 found a 50 percent successful rate in using OCR to transcribe barometric pressure observations in different writings (Wilkinson et al., 2019). Particularly, the study has found that OCR handles handwritten material almost as good as printed material, and its performance on handwritten records varies largely in one single source from page to page. The author suggested the difference in output quality might be related to the efforts in image pre-processing (i.e., converting to grayscale, removing noise). These two studies suggested that it is possible to use OCR on handwritten material and gain an adequate result, but effort must be made to optimum OCR's performance. However, the effort needed to make OCR a viable alternative to manual transcription may be enormous. In the examples, Stickler, Alexander et al. (2014) dealt only with printed sources and Wilkinson et al. (2019) processed only digits, and the results are still not comparable to manual transcription. Therefore, there may be a long way to go before OCR can handle cursive handwritten material. Newer technology can be a solution. *Guidelines on Best Practice for Climate Data Rescue* has suggested that ICR (Intelligent Character Recognition) and IWR (Intelligent Word Recognition) are able to work with handwritten materials. Started seven years ago, the project is still under development (World Meteorological Organization, 2016). Both methods use machine learning techniques, so they can automatically transcribe historical records more reliably than OCR if adequately trained in handwriting styles and patterns. With the rapid development of technology, there is concern that these projects will be obsolete by the time they are completed. Nevertheless, while current automation is not reliable enough, AI augmented data rescue may be the solution to automation reliability as future efforts are engaged. It does, however, point to another resource issue: can the data rescuers keep up with the pace of innovations?

Researchers have also argued that automated approaches are not reliable to use in all sources. In other words, this approach is "only appropriate for certain sources' (Ashcroft et al., 2018, p. 1631). The type of sources here includes the condition (e.g., faded ink, poor resolution), layout (e.g., tabular data, charts) and the content (e.g., special character, ink dots). Several studies have concluded that the poor condition of the records is one of the problems that automation is not preferred in data rescue. Sources that have faded ink (Camuffo & Bertolin, 2012), poor resolution (Stickler, Alexander et al., 2014) and mediocre print quality (Blancq, 2010) have been reported to be unsuitable for automated approach. Another obstacle to automated approach is the layout of the source. It is important that the source data have a clear layout, where the locations

of data values can be easily found by algorithm (Wilkinson et al., 2019). On the other hand, for those with vague layout and dense writing styles (diaries and tables and everything in between), it is difficult to automate them (Figure 3.1). What is visually obvious to a human has to be spelled out in an excruciating number of steps for the computer. Experience from image cleaning suggests that people are aware that AI may be hard to implement. Examples of a good layout are records that are in a very regular, tabular format (Stickler et al., 2014).



(a) An example of a scanned leger sheet (for November 4-6, 1884) that has faded table lines.



(b) Albert J. Kelly's weather diary page for September 25-28, 1918. MG3054. McGill University Archives. This is an example containing dense handwriting.

Figure 3.1 - Examples of vague layout and dense writing styles.

There are researchers holding positive ideas towards automated data rescue who proposed some ideas to cope with the limitations I discussed above. Image pre-processing and layout analysis are two processes that can optimize the automation results. Image pre-processing steps such as "deskewing, contrast enhancement, and despeckling' can remove pixel noise, aligned the characters, and improve the quality of an image, which will thus optimize the automation process and improve the quality of the output (Odunayo et al., 2021; Wilkinson et al., 2019, p. 25). As mentioned above, it is hard and important for the automated approach to quickly find the values on the page to be transcribed; layout analysis is an essential step that can address limitations. It could be done by using "automatic techniques to split images of large pages of data into different segments" (Craig & Hawkins, 2020, p. 133). The algorithm can identify the text in smaller sliced images easier than that in the whole page. This layout analysis process has already been

manually tested in the ERA-CLIM project and has proved to be successful (Stickler, Alexander et al., 2014). Therefore, automating the layout analysis is very essential in automated data rescue, and it can potentially improve the current attempt of automation as well.

Overall, it can be inferred from past experience that researchers are concerned about the adoption of automation in three main aspects: time, cost, and accuracy. While some researchers are excited about the evolution that automation may bring to historical data rescue, others point out the possible obstacles and offer their potential solutions. Although automation sounds appealing and there are few attempts, it has been realized that automation still needs to be developed and improved. AI augmented automation is a new field that has rarely been experimented and now there is an increasing amount of effort and attention being put into using AI as a helper for transcribing historical records. It is important to know in advance what researchers think about using AI augmented data rescue in research to aid the transcription process, and what suggestions and concerns there are. A connection needs to be made between historical data scientists and the idea of AI deployment so that future attempts of automation can be guided. To find out what they think, we conducted an online survey.

3.3 Methods

An online survey is conducted to assess automation as an approach for rescuing historical weather observations. The hope is that it will improve understanding of the perceived losses and gains of changing from manual data rescue to automated data rescue, the perceptions of the automated data's fitness of use, and the association of respondents' view towards automation with their background, skills, and roles. This survey provided a valuable reference for implementing machine learning automation as an alternative for manual data rescue and will identify the merits of automated data rescue by comparing it with manual data rescue. This could be a reliable result for future studies to facilitate the use of machine learning in data rescue, and it could be a possible basis for developing a better approach for data rescue.

Ways of understanding people's minds include surveys (paper or online) and interviews (in person, video or audio). Similarly, Bush et al. (2019) and Ryan et al. (2018) have also used surveys to investigate students' opinions on rescuing historical data in the classroom. There are advantages that a survey is conducted instead of an interview in this research. Studies have shown that interviews are way more time consuming to set up and conduct (Bowling, 2014;

Williams, 2003). Knowing that I have a large population of potential participants, a survey instead of an interview will make it possible for me to reach 66 participants within a relatively short time. Survey also gives participants the freedom to decide when to complete the survey, so different time zones would no longer be a problem. The participants that I'm trying to recruit covered a wide geographic area. The flexibility of the survey will work better than the interview especially when more than half my potential survey participants are in a different time zone with me. Privacy is also important for participants when conducting a survey interview. A survey will provide a better sense of anonymity than face-to-face survey either online or in-person.

3.3.1 The population and the sample

I sought two groups of people who are researchers and professionals who are involved in the historical handwritten data rescue process. One group is working on data rescue projects, and the other group is working on citizen science projects. These researchers are chosen because they have been involved in historical data transcription projects and have perspectives or experiences with automation. They are also multidisciplinary researchers who can guide and shape the future of historical record transcription. I identified approximately 248 potential participants who fit the description.

I selected a sample of 20 percent. However, this is not a random sample, but a purposive sample. The 20 percent are the people who are leaders or initiators of the projects. Part of the participants are data rescue project leaders who are working on rescuing historical handwritten datasets. The other part are citizen science project initiators that are working on transcribing handwritten materials. Here, the leaders and initiators of these projects typically have more experience and control over the deployment of AI augmented automation in future projects. They have demonstrated their interest in automation in their lectures and publications, and their views on automation can significantly shape the future of data transcription. Some of the participants were identified by colleagues that I'm currently working with. We have an extensive network of individuals working in the field we are researching. I also got in touch with citizen science and data rescue scientists who have a previous connection with the DRAW project and asked if other associate researchers they know are willing to participate.

I focused on surveying the initiative leader or leading scientists in the citizen science and data rescue community. These professionals are usually chairman, executive director or associate

members of the organizations, groups, or initiatives. 66 potential participants are contacted, and some of them have referred me to other interested researchers. Eventually, 50 responses were received, and the response rate is 76%.

The participants were recruited through email. Each individual received a maximum of two emails: one initial recruitment email and one follow-up email if there is no response after two weeks have passed. Once the individual consents to participate in the research, follow up correspondence via email will be used to send out a link to the survey questions.

Some of them respond to me with why they did not take the survey. Several people said they have not been in the field for a while or they have retired. Some other people respond that they have not been engaged with the transcription part of their project, and forwarded the survey to members that have more direct involvement.

3.3.2 Survey design

An online survey was conducted based on the participant's availability. The survey was made using McGill licensed Microsoft Form. Microsoft Forms is an online survey creator, part of Office 365 offered by McGill. It has been approved by McGill to use for survey purposes.

The survey begins with a consent form. If the respondent declines, then the survey will end, and the respondent will be thanked. If they agree, then they will proceed to the survey. The expected length of this is around 15 to 20 minutes according to the pilot test. It includes mostly multiple-choice questions, several open-ended questions, and one set of Likert scale questions. The survey contains four sections of questions. It includes the participants' background knowledge, their experience of transcription data rescue projects, their original goal of data rescue projects, and their perceptions and opinions about automated data rescue. The survey can be found in Appendix A.

Specifically, this survey included both closed questions and open-ended questions. On the one hand, closed questions such as multiple choice and Likert scale questions are quicker to complete and easier for coding and analysis; whereas open-ended questions take longer time for the respondent to answer and longer for the researcher to extract themes (McColl et al., 2001). On the other hand, open-ended questions get detailed and personalized responses by giving participants an opportunity to show their own views while closed questions' responses are

categorical and standardized (Bowling, 2014). Therefore, a mixture of closed and open-ended questions is included in this survey.

3.3.3 Pilot test

Two rounds of pilot tests, also called pretests, were done before the survey was officially launched. Studies have shown that pilot tests will improve the reliability and validity of the survey, and it is good practice to pretest your survey before you launch it (Bowling, 2014; Roopa & Rani, 2012; Williams, 2003). Since I am surveying two groups of people, it also is important to ensure that there is a common understanding of wording and concepts.

The first round of pretest is to test out the reliability of the survey. It was done within a group of native speakers--I am not a native speaker of English and this survey was conducted in English--who do not have professional knowledge about the survey topic but is working in the relevant fields. I am checking to see if people are answering in the same general way I expected. This pretest is to make sure the result is both consistent internally and under different circumstances (Roopa & Rani, 2012; Williams, 2003). This means, I'm testing to see if participants' responses are consistent within the survey and are consistent at different times. I also used this pretest to make sure the questions can be understood by participants regardless of their professions and to ensure the questions are placed with the best order possible to avoid any confusion.

The second round of pretest is to test out the validity of the survey. This pretest is done within a group of professionals who are working with me in the same projects and have professional knowledge about the content of the survey. This validity test is to make sure that this survey measures what it intends to measure (Black et al., 1998; Roopa & Rani, 2012; Williams, 2003). Specifically, this step makes sure the wordings of the questions achieve desired results, and no misunderstanding will appear across all respondents. For example, it helps me to eliminate misunderstandings in the definition, such as what I mean by automation. It will also make sure that the choice of response for closed questions are inclusive, and all of the relevant issues are discussed in the survey (Bowling, 2014).

3.3.4 Efforts to increase response rate

Several efforts are made to increase the response rate of this survey. The invitation email is mostly sent on Monday, when the workday starts and people check their emails regularly. I also

personalized the email and included familiar names of receivers in the subject line, so people will tend to pay more attention and the email will less likely be missed (Linsky, 1975; McColl et al., 2001). These two actions increased the chance that participants open and read the invitation email, thus increasing the response rate.

In case of participants forgetting about participating in the survey, one or two reminders were sent two weeks after any previous correspondence according to different cases. Studies have shown that follow-up reminders are very effective in stimulating responses (Linsky, 1975; McColl et al., 2001; Nulty, 2008), and the results showed that reminders indeed helped increase the response rate of this survey greatly. Another thing that helped arouse participants' interest is to include part of my preliminary results in the correspondence email as an incentive (McColl et al., 2001). The detailed and specific information can quickly arouse participants' interests, thus they tend to respond to the survey. I also included the survey URL in the invitation email, and guaranteed anonymity in both the email and the survey, which has been proven by studies to have a positive effect on increasing response rate (Nulty, 2008).

To further increase the response rate, the design of the survey is very important as well (McColl et al., 2001; Williams, 2003). One tip that I followed Larsen et al. (1987) is to keep the questions simple, short and specific. Another tip I followed is to present general questions first and then more specific ones (Black et al., 1998).

3.3.5 Acquiring ethics for human subjects, also anonymization

An REB (Research Ethics Board) approval was acquired prior to conducting the survey from Research Ethics Board (REB File # 21-05-005, Appendix B). The initial anonymity of the participation will not be guaranteed, since I directly contacted the participants for the survey. The participants were asked about their names and organizations at the beginning of the survey only for if they agree to include their names in the publication. As a result, there will be no identifier attached to the finished responses, unless the individual wishes to get a copy of the results or consent to include their names in the research. Those email addresses and names were separated from the responses and stored in a separate document. The key that linked the number identifiers to the names and other identifying information were stored in another password protected document. All results in the reported finding will remain anonymous unless the participants consented to include their names.

This online survey is conducted based on the participant's availability using McGill licensed Microsoft Forms. Responses to the survey are downloadable through Microsoft Forms platform into spreadsheet format and are coded for analysis. The information collected is coded to examine the bivariate correlation of their response with different factors (e.g., backgrounds, skills, and roles).

3.3.6 Data analysis

I have two types of responses. One type includes content that is amenable to univariate and bivariate analysis. These responses were analyzed using descriptive statistics (e.g., means, averages) and were supplemented with tables and graphs. A bivariate analysis of these responses was also conducted, and due to the small sample size, Fisher's exact test was used. A p-value of < 0.05 was considered as statistically significant.

The other types are open-ended questions. Themes were extracted from responses, and synthesized and coded on spreadsheets. There was triangulation in the extraction of themes to enhance the validity and credibility of the results. Several colleagues from different fields were consulted on the results of theme extraction.

3.4 Results and discussion

This section presents the survey results, cross-theme synthesis, and discussion of the results. It will begin with an introduction and discussion of the respondents' background, as it provides some indication of skills and knowledge that can be applied to AI-augmented automation. It will then talk about past and present operations of rescuing historical records. This will describe how much AI-related work has already been applied and will indicate how much work may be needed if we want to transition to AI-augmented automation. Lastly, it will discuss respondents' perception of AI-enhanced automation, including their previous experience with AI, their willingness to use AI, and their concerns about applying AI. We hope this result and discussion can serve as a guide to avoid unnecessary obstacles if AI-augmented automation will be applied to future projects.

3.4.1 Respondents' background

I am curious about possible distinctions between the responses of data rescues and citizen scientists, as data rescuers may prefer automation because they want transcribed output, while citizen scientists may be less interested in automation because they are more interested in motivating non-experts to participate in scientific processes. Fifty-two percent of the participants (n = 26) were self-identified as both data rescuers and citizen scientists; 48% were either one of data rescuers or citizen scientists (n = 24). Among the participants that were self-identified with only one role, 75% of them (n = 18) were self-identified as data rescuers; 25% were citizen scientists (n = 6). In later sections, these identities are used to investigate possible differences between groups.

The average time that the participants worked in data rescue or citizen science was 12.37 years, compared to a median of 10.5 years, as one respondent had worked for 50 years. Fifty percent of participants (n = 25) have been working in this field for less than 10 years; 24% (n = 17) have been working for 10 to 20 years; 16% (n = 8) have been working for more than 20 years (Figure 3.2). The assumption is that if one works longer, one may have more resources to devote to AI, but one may also be stuck in a particular way of transcribing. However, the result did not show statistical significance differences between different age groups. Most participants have some experience, either professional or amateur, in climate or related fields; among them 76% of the participants (n = 38) have been working in climate or related projects as a profession. I hypothesized that people interested in the results, such as climatologists, would be more inclined to use automation, although I did not find statistically significant differences in comparing climatologists and non-climatologists in terms of their motivations for adopting AI.

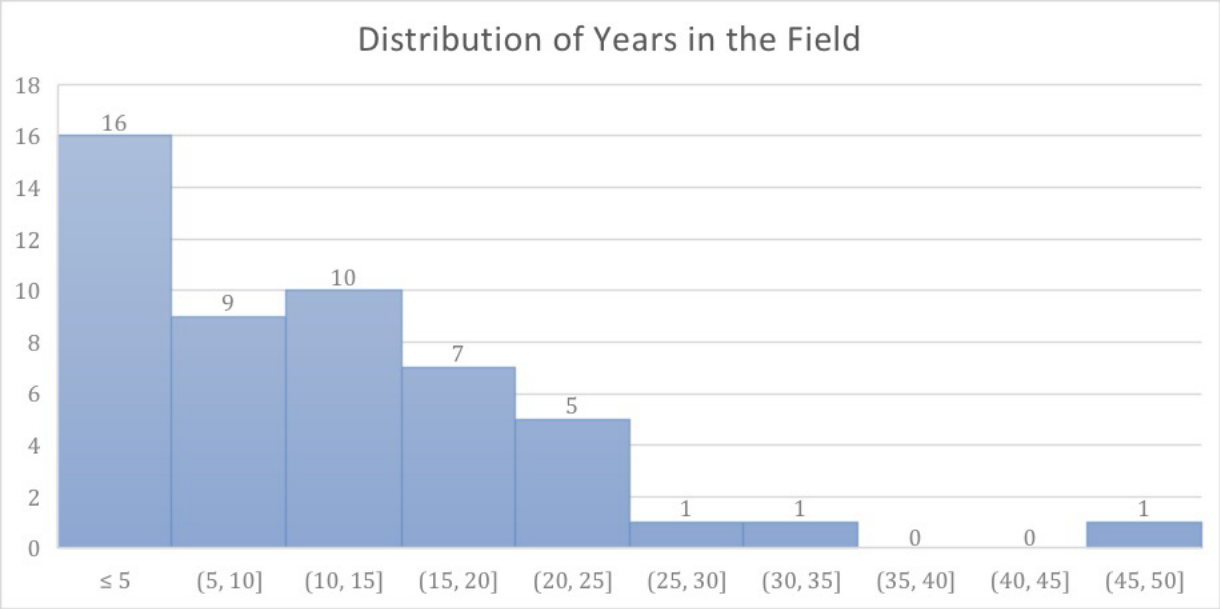


Figure 3.2 - Distribution of number of years in the field.

I endeavored to ensure there was no geographic bias. The participants are involved in diverse projects that help rescuing historical data worldwide. The origin of the projects is from all six WMO regions while more than half are from Europe. These are large international organizations, so they have the capacity to use automation if preferred. The participants include collaborators of ACRE and I-DARE who are working with existing international data rescue initiatives, and Copernicus Climate Change Service Data Rescue Service (C3S DRS) regional projects (e.g., European Reanalysis of Global Climate Observations (ERA-CLIM1, ERA-CLIM2), Early Meteorological Records from Latin-America and Caribbean (EMERLAC), C3S 311a Lot 1 Data Rescue Service and many more). The emphasis on Europe is confirmed by Brönnimann et al. (2018). The participants also include principal investigators or researchers from data rescuing citizen science projects that involved middle school or university students or volunteer citizen scientists (e.g., OldWeather, Weather Rescue at Sea, Meteororum ad Extremum Terrae, UK Tides, Addressing Health, Scarlets and Blues, Davy Notebooks Project and many more). These would be the main actors in adopting AI or diffusing its importance to other data rescuers and citizen scientists.

When asked about the roles they play in the historical data rescue project, 46% of participants (n = 23) identified themselves as project manager or principal investigator, and another 46% (n = 23) identified themselves as researchers of the project. Historical data rescue projects tend to be

small scale, so it is very common for one person to have several roles. Therefore, in many cases, project managers and principal investigators are the leaders of the projects, and, at the same time, they are also the researchers of the project. From the result, we can conclude that the majority of the participants are considered core members of historical data rescue pages. In other words, they are the leaders who make important decisions, such as whether to adopt AI. if they adopt it, then it will be adopted in the project.

3.4.2 Past and present operation of rescuing historical records

This survey asked the participants about the source of material to be transcribed, use of information and communications technology for transcription (e.g., in collection or storage), funding source for development of the transcription platform, and the goal of transcription efforts, currently as well as in the past. Understanding the past and current operation of these projects can possibly serve as a predictor of AI usage. For example, computational ability can be inferred from the transcription method, and the source material may be an indication of the acceptance for AI usage. The majority of participants (86%; n = 43) reported their data source as logbooks, including notebooks, minutes books, and data sheets. Some participants have also been working with diaries (40%; n = 20) and newspapers (24%; n= 12). It is also indicated in the response that the logbooks and diaries are predominantly handwritten. These materials have been studied and confirmed that they are very difficult to work with and can bring problems to the transcription process.

What method is used, especially if it involves technology, also could be a predictor for the acceptance of AI adoption. To transcribe the historical data, computer-related technologies and software have been widely used to capture and collect the information. Spreadsheet editing software (e.g., Microsoft Excel) has been widely used among data transcription projects (64%; n = 32). Typically, the data or project managers create a transcription template to guide users in keying in content (cf., one respondent used Microsoft Word). This confirms the respondents also have reported using Zooniverse (30%; n = 15). Zooniverse is a web portal that facilitates and hosts citizen science projects through customized tools and builders, where most participants build their citizen science transcribing projects (Fortson et al., 2011). This confirms the finding that manual transcription remains the most popular method of historical data transcription, in which expert researchers or hired transcribers key in the records, although there is a notable shift

to citizen science approaches where student transcribers in a classroom setting or volunteer citizen scientists are engaged in the transcribing process (Wilkinson et al., 2019; World Meteorological Organization, 2016). Others report using a database system (26%; n= 13) as part of their transcription method. Details are shown in Figure 3.3.

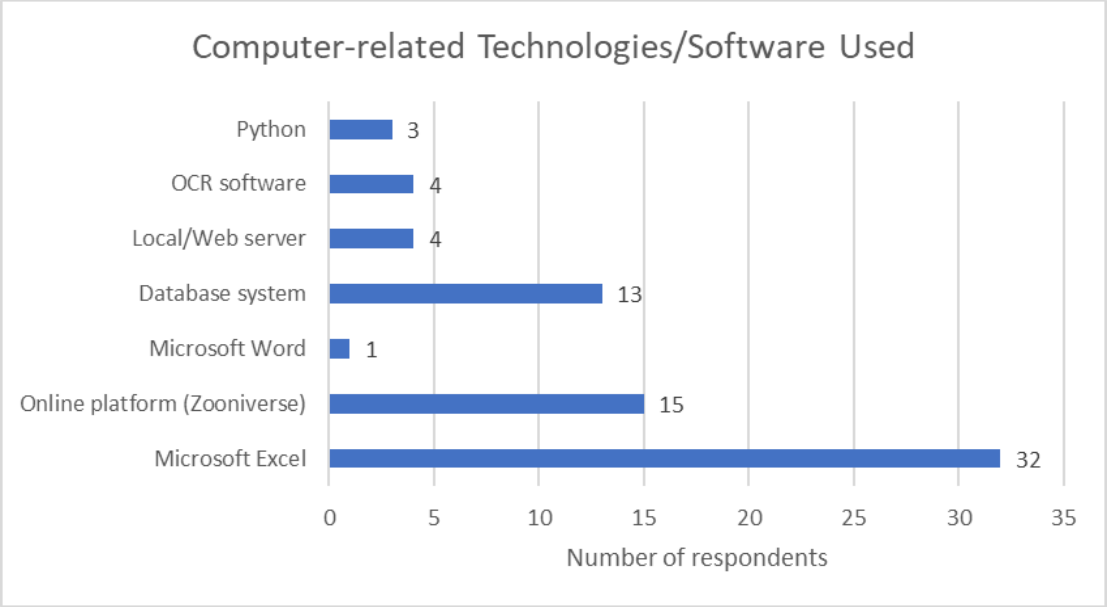


Figure 3.3 - Respondents' choice of computer-related technology/software for transcription.

Use of computer software to store data is not the same as use of computers to automate the data collection. There is a distinction between software like Microsoft Excel and OCR. Automation of transcription involves the original capture, the storage of the data, various decisions made about post-processing (e.g., validating transcriptions by selecting the interpretation with the greater agreement), and ways to enable downloading of the data. As the frequently used technologies and software by respondents, while Microsoft Excel and Zooniverse may involve some computation, database systems, on the other hand, are mainly used to store and manage transcribed data and do not involve computation. Yet, the OCR or automated methods are the least popular methods for the transcription process, perhaps because it requires the most computational ability. A small number of participants reported using OCR software (8%; n = 4), local/web server (8%; n = 4), and python script (6%; n = 3) as part of their transcribing method. Shen et al. (2021) stated expert knowledge and technical background are required prior to taking advantage of advanced AI-related approaches. This implies that experience in softwares such as

Excel is not enough to easily transition to AI, while experience in OCR and python may be closer.

Funding is an important basis of any citizen science or data rescue project, especially since data rescue activities are often under-resourced (Ashcroft et al., 2018; Brönnimann et al., 2006). Figure 3.4 shows the funding reported by respondents. More than half of participants stated that their projects are funded by research grants or contracts they applied (54%; n = 27). Less than one-fifths of participants (16%; n = 8) were rescuing the records as part of their job and getting paid by their employers. Very few participants (4%; n = 2) funded their projects from their own pocket money. It is worth noticing that 14% of the participants (n = 7) said their projects are funded by a mix of above-mentioned funding sources. In conclusion, the major funding source of data rescue projects is research grants, and a substantial number of projects are funded by various sources. This brings concerns to the implementation of AI-related methods. As research grants are not a stable source of funding, it is difficult to support staff and experts to maintain it over time, even if we ignore the funding needed for development and consider it open source.

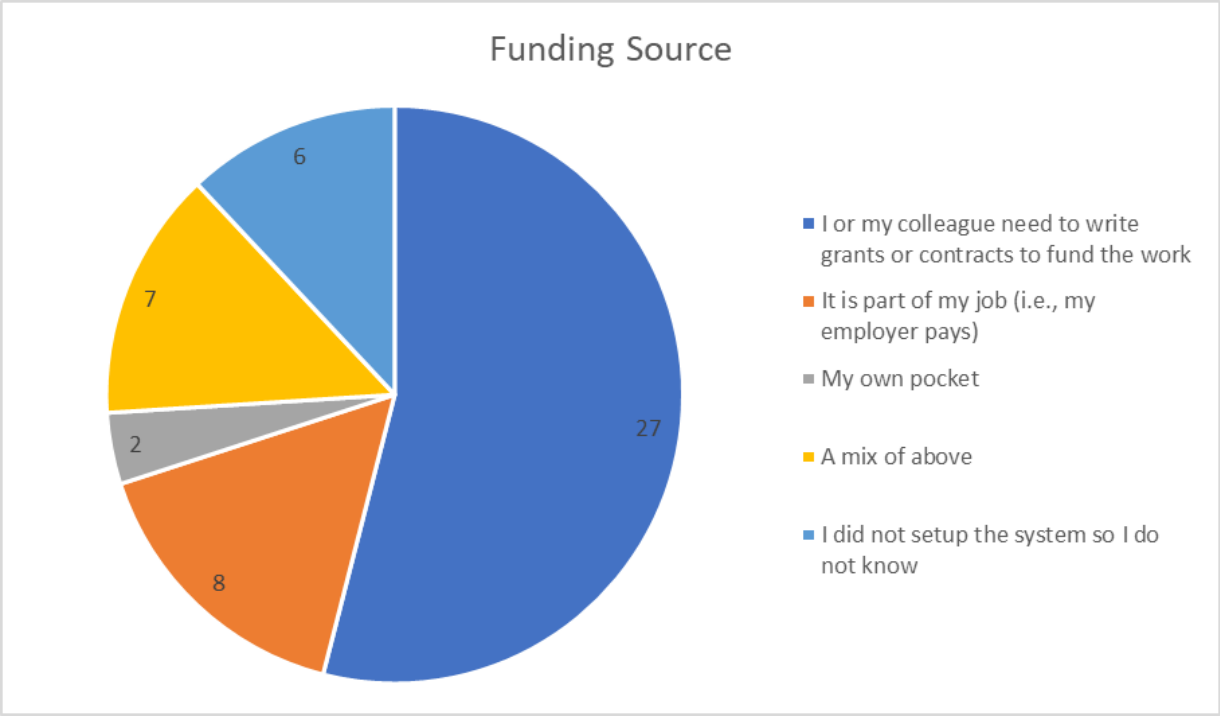
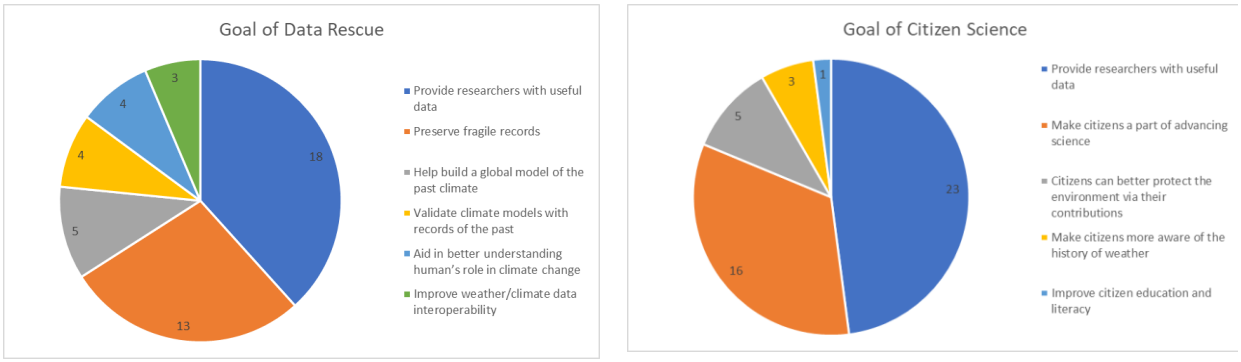


Figure 3.4 - Source of funding as reported by respondents.

This survey also asked the participants about the most important goal of climate data rescue (Figure 3.5a). The two most agreed goals are “provide researchers with useful data” (38%; n =

18) and “preserve fragile records” (28%; n = 13). Three respondents expressed that all the goals mentioned are equally important, and one of the respondents commented that: “All [goals are] very important in [their] own right, but… without preservation none of the rest is possible.” When asked about the most important goal of climate data related citizen science, there are also two most agreed options (Figure 3.5b). Nearly half of the participants (48%; n = 23) agreed the goal is to “provide researchers with useful data”, and one-third of the participants (33%; n = 16) believed the goal is to “make citizens a part of advancing science”. Two participants expressed that both providing data and public engagement are very important, and one commented that: “[T]here isn’t the most important’ [goal]… There are trade-offs but you can’t define one as the most important.” The most agreed goal of climate data rescue and climate citizen science is both providing useful data. Yet, the second most agreed goal is very different. Therefore, a conclusion can be made according to the responses. While both projects focus on providing data, citizen science projects value public engagement, and data rescue projects pay attention to record preservation. One could assume that data rescuers would be more amenable to automation, which removes the need for citizen participation.



(a) (b)

Figure 3.5 - Most important goal of (a) data rescue and (b) citizen science identified by respondents.

From the response, the status and operation of past and current data rescue projects can be concluded. Overall, it can be referred from the response that the majority of the data source in the past and current projects are handwritten on paper (although the paper may have been scanned). A few exceptions that include some typewritten and printed data sources. To transcribe these records, the most used method is manual transcription, essentially data entry, regardless of

traditional data rescue projects or citizen science aided data rescue projects. The sheer variety of handwritten data sources might be the reason why manual transcription is widely applied. As will be seen in the next section, many respondents have commented that manual transcription is the most reliable transcribing approach so far for handwritten records, although it is time-consuming. The technology used and funding sources in past and present operations hint at resistance regarding the acceptance of a transition to an AI-augmented approach. The difficulty of acquiring technical skills and supporting long term maintenance are some of the foreseeable obstacles to overcome if AI-augmented automation is to be applied.

3.4.3 Perceptions of automation

An automated approach might be one option to improve data rescue projects, and it is important to understand researchers' experience and opinion on this approach. The following section will first cover any respondents' previous experience on AI and their confidence to take on AI-related tasks. I will then examine their opinion and comments on using an automated approach in the future.

3.4.3.1 Previous experience of AI and/or training on AI

Automated approaches are mostly implemented using AI. Although AI is a widely known concept, the application of AI is not widely familiarized. In other words, because AI is a vast field, people may know what AI is, but that does not mean they can install a system utilizing it. If automation is going to be employed to data rescue in the future, researchers will need some knowledge of AI to run the project. Therefore, it is important to understand what researchers "know" about AI and their confidence in doing AI related tasks.

The participants were first asked about their experience in employing AI in the past (Figure 3.6). 72% of the participants (n = 35) said they have no experience using AI, and only 12% of participants (n = 6) are fluent in AI. The rest of the 16% said they had some experience in using AI. When asked about specific tasks, nearly half the participants stated: "I know what machine learning is, although I've never used it" (49%), yet very few participants have coded some algorithms (6%) or have used it extensively in their research (6%).

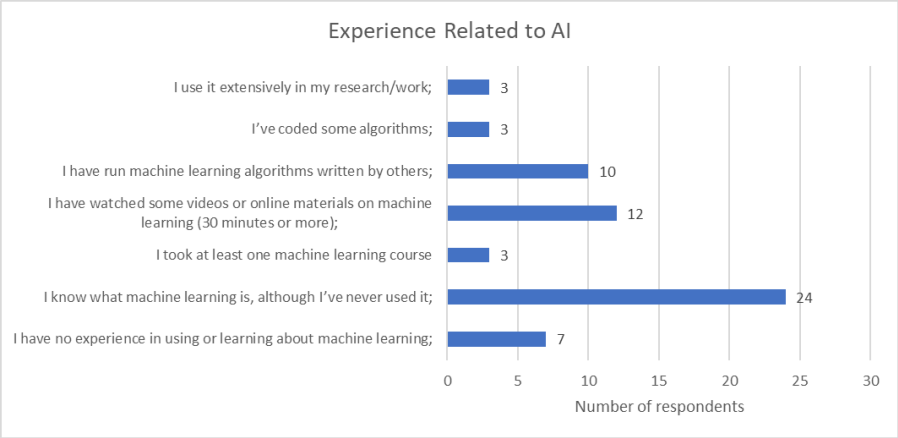


Figure 3.6 - Respondents' experience related to AI; they can select multiple choices.

The participants were then asked about their level of confidence in doing technical tasks that are related to automation employment (Figure 3.7). A group of Likert scale questions were asked to evaluate their level of confidence. Overall, more participants choose confident than not confident in all the tasks. Specifically, participants are the most confident at running an existing OCR software to transcribe records. This presumes a package with a well-established user interface, such as tax software. However, OCR for transcribing historical records is usually open source and does not have a user-friendly interface. That is why Kevin Wood from the National Oceanic and Atmospheric Administration (NOAA) suggests: "There should be a UI/UX that would make operations as simple as a typical software product developed for wide distribution."[11] Respondents also express moderate confidence in using SQL, programming language (e.g., Python, C, Java), and database systems. In contrast, they are the least confident in running deep learning algorithms (e.g., Convolutional Neural Network). Similarly, they are also not confident in using machine learning algorithms (e.g., Linear Regression, Random Forest) or libraries (e.g., Tensorflow, Pytorch). I hypothesize that some comfort with programming languages will allow them to access AI software libraries. Thus, even if they know little about AI development itself, they may possess the skills to install the software, troubleshoot it, and maintain it over time.

In conclusion, the majority of participants are not familiar with employing AI technology, yet, they have basic knowledge about AI. They are more familiar with technologies that have already been widely used or experimented in the community. Precisely, participants express confidence

[11] Any respondent mentioned by name agreed to be on the record.

in using OCR, database system and SQL because these technologies are the approaches that they have been using to transcribe the records. On the other hand, machine learning and deep learning are relatively new and are still under development. It has not been widely applied until recently. The assumption here is that participants have the relevant skills to "upgrade" to AI on their own or they have sufficient knowledge to collaborate with data scientists to exploit AI.

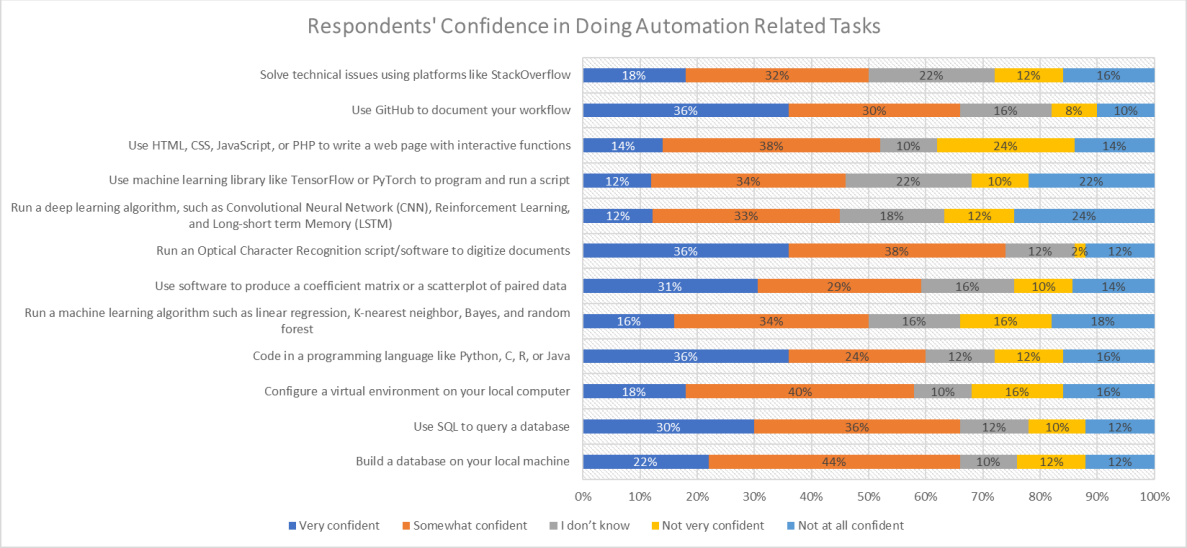


Figure 3.7 - Respondents' level of confidence in doing technical tasks that are related to automation employment.

3.4.3.2 Will of using AI in their projects

Participants may or may not possess familiarity with AI; it is possible that their perceptions of utility will influence their use of AI. The survey asked the participants about their opinion of using AI and what are the resistance that researchers need to overcome while applying AI. Specifically, respondents were asked about whether they would use automation if the transcription portion of their project could be automated; an overwhelming percentage (84%, n = 42) would consider using it. One respondent commented: "I think if [automation] is successful, it would be very useful." Only 10% of the participants (n = 5) were against. One participant commented: "I'm very against it. People are much better at transcription than machine." Others are worried about the output accuracy: "I've seen very low-quality results in the past with ORC [sic]." They are not confident in using automation because it is difficult to recognize those records even for experts. The following section discusses in detail about possible concerns that

might affect respondents' willingness to use AI-enhanced automation. Examples include output accuracy and the amount of time and resources required to set up the workflow.

Although opinions may be nuanced, respondents were asked about reasons that make them reluctant to automate the transcribing process (Table 3.1). “If an automation would provide you with a pre-trained algorithm with a customized user interface where you can adjust the parameters according to your project, what might be reasons not to automate your citizen science/data rescue process.” We tried to write the question so that people wouldn’t think about the user interface but focus on the automation process itself. The most common concern is the inability to guarantee output accuracy (54%). The next two biggest obstacles of automation are identified as lack of funding (34%) and insufficient expertise to maintain the software (34%). As has been discussed above, funding is very limited in data rescue projects, especially for non-profit organizations (Ashcroft et al., 2018; Brönnimann et al., 2006). A lack of funding also will make it difficult to hire software engineers or system administrators to maintain an automated system. One-fifth of the participants (22%) are worried that the benefits to involve citizen scientists will be reduced if transcription is automated. Specifically, one participant commented that, without citizen scientists or human transcribers, there is “potential to miss comments in data”. It is also worth noting that two participants said they could think of no reason not to automate the transcribing process. The respondents elaborated their detailed opinions in subsequent questions, which are analyzed in the next section.

Table 3.1 - Reasons that discouraged respondents from automating the transcription process.

If an automation would provide you with a pre-trained algorithm with a customized user interface where you can adjust the parameters according to your project, what might be reasons not to automate your citizen science/data rescue	
Response	**Number of respondent (%)**
The performance of automation is not guaranteed	27 (54)
Lack of funding	17 (34)
Do not have experts to maintain the automation	17 (34)
Have tried automation already, but the result is not promising	10 (20)
There are benefits to involving citizens that would be reduced if transcription was automated	11 (22)
Too much work and effort needed to shift the process	9 (18)
There are no obvious benefits to do so	2 (4)
Automation algorithms are just too “black box”, opaque to trust	2 (4)

Overall, the majority of respondents are willing to try automation although they also have some concerns, predominantly around accuracy. The next section of questions allowed respondents to express their reactions at length.

3.4.3.3 Concerns of applying AI

The literature says output accuracy and time are two important factors when contemplating the use of automation to transcribe environmental content. This section will first discuss respondents' thoughts and comments on the output accuracy. Then I will discuss the length of time respondents considered acceptable to set up an automation workflow. As time is an important factor in any research, it is necessary to know how much time participants are willing to spend on setting up the automation. At last, I ask respondents what are the advantages and disadvantages in switching to a purely automated approach in the future as well any other comments they might have. This led to some interesting findings about the optimum option for future data rescue.

Output Accuracy

In the survey, I asked respondents about accuracy in terms of how confident they are that automation will accurately transcribe the data (Figure 3.8). It is split between somewhat confident and not confident (36% each) of respondents. This lukewarm response matches the previous research, in which researchers showed concern about OCR output accuracy especially in transcribing handwritten records (Brönnimann et al., 2006; Craig & Hawkins, 2020; Stickler, Alexander et al., 2014; Wilkinson et al., 2019). Since accuracy is such an important issue, I was curious about any association between concerns over accuracy and any other attributes. So I tested the statistical significance of the confidence of accuracy in different groups of participants (e.g., identity, willingness to use an automated approach, years of working, expertise in climate research, confidence of doing technical tasks related to AI, AI experience). None of the results proved significant. The p-value can be found in Appendix C. This means there is no association between participants' choice of confidence in output accuracy and, for example, their years in the field (resistance to change), the groups they are in, etc.

Anticipating that resistance might be due to other factors, I asked them to explain. Possible reasons are that participants have concerns about accuracy for different reasons. The most identified concern about accuracy is the various data sources. Previous literature confirmed that

automated approaches are not reliable on all data sources, and the condition of the original material determines the output accuracy to a large extent (Holley, 2009; Klijn, 2008). Data source here includes quality, script type, data arrangement format (e.g., table, text), and other relevant variables. Even respondents who expressed confidence about automation accuracy have some concerns about the data source. They are particularly concerned about the 'script' of the data source, like those found in handwritten materials. Ciara Ryan, a climate data rescue researcher at Met Éireann commented:

> Having worked with historical meteorological records it can be difficult to decipher handwritten values. Unlike documents containing text where a word may be inferred from the context of the sentence/document, it may be difficult for automated programmes to accurately transcribe numerical tabular data where each value is independent.

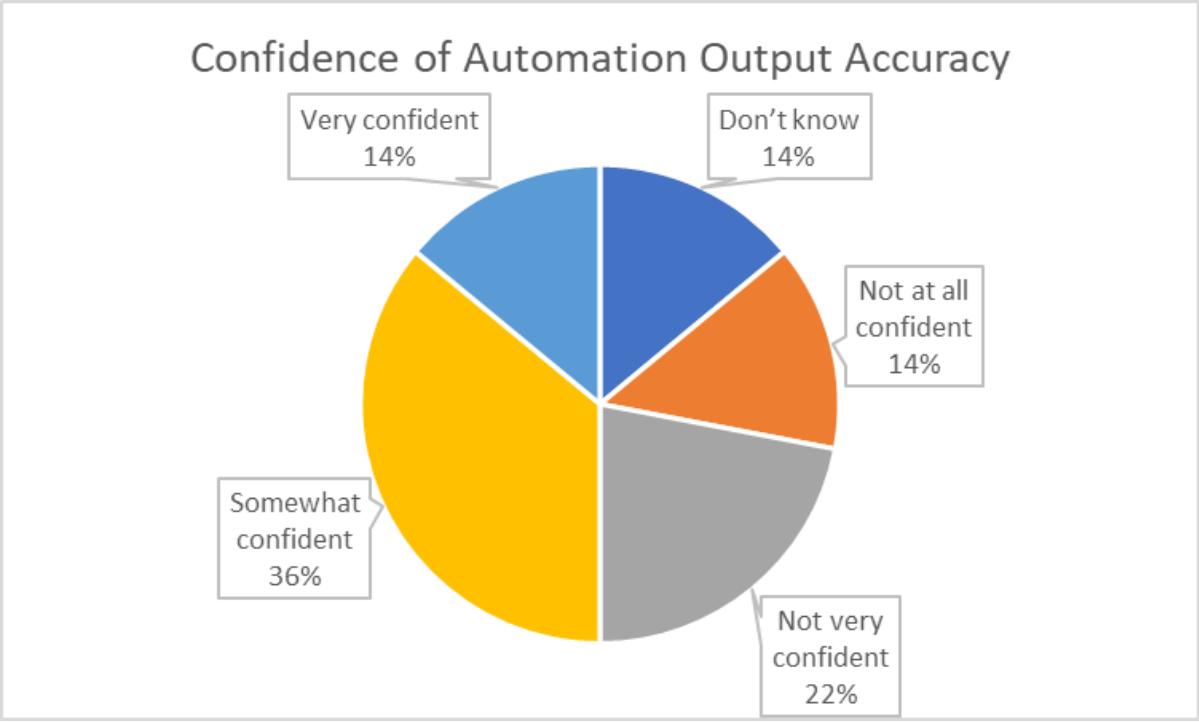


Figure 3.8 - Respondents' confidence in accurately transcribing records using AI-augmented approaches.

Conversely, respondents were more confident when the material was in typewritten or printed records. Gilbert P. Compo, a weather and climate scientist at National Oceanic and Atmospheric

Administration (NOAA), expressed concerns for using automated transcription approaches on handwritten records but was confident about type written records:

> I have been part of 3 different auto transcription projects. The most advanced used the same software used by the US postal service to read handwriting on US mail. [The auto transcription software] could not predict when it would make a successful transcription and when it would fail [for handwritten records], making the results useless. In contrast, attempts to transcribe type written tables were very promising ….

Respondents reported that accuracy depended on source format (i.e., data arrangement) as well. Some (60%) said automation had yet to demonstrate any progress on content arranged in tabular format (e.g., ledger sheets); whereas the rest (40%) were very confident about tabular records. In conclusion, the exact nature of the tabular format could determine the accuracy. In complicated cases where the original documents contain handwritten entries misaligned within entry boxes, it is possible for automation to miss some entries, reducing accuracy. In simple cases where all entries are properly formatted, it is possible to have better accuracy.

Respondents also mentioned other accuracy issues, related to automation-induced errors, overhead of verification, and threshold accuracy. These accuracy issues are gaps that have not been addressed in past studies and may be worth considering in future projects. A couple (n = 2) respondents were concerned that automation might introduce systematic errors new to the data rescue process (e.g., wrong image processing/transcription model parameters, training dataset for alphanumerics). Three were worried the time required for verifying the results would be more than having human transcribers. It would not be worthwhile to use automated approaches if there is no net time saving and more errors to correct. In terms of how respondents might respond to possible errors, a couple (n = 2) said they realized the process would not be perfect, but they were "somewhat at peace with it." It is hoped that these new findings will make it easier for future research to consider and evaluate the transition to automated approaches.

Time and Resources in Setting Up the Automation

I asked specifically if respondents would deploy automation if an acceptable short amount of time was needed (I left it to the respondent to determine what would be acceptable short). Approximately two-thirds of the respondents (68%, n = 34) said they would use it, and only eight percent of participants (n = 4) said they would not use it. Respondents were then asked

about their ideal set up time (Figure 3.9). While the answers ranged from three hours to one year, I calculated the average acceptable set up time as 2.5 months (75 days). This average value of set up time could be a guideline for future automation projects in estimating their installation and set up time which would be acceptable to both software engineers and researchers. A time indicator also could imply how many additional supporting staff a project needs to automate their process. The values suggest patience on the part of the respondents. One respondent commented:

> If the set-up time ultimately allows you to increase your data collection and provide useful research outputs for study, it seems like any amount of time would be justified.

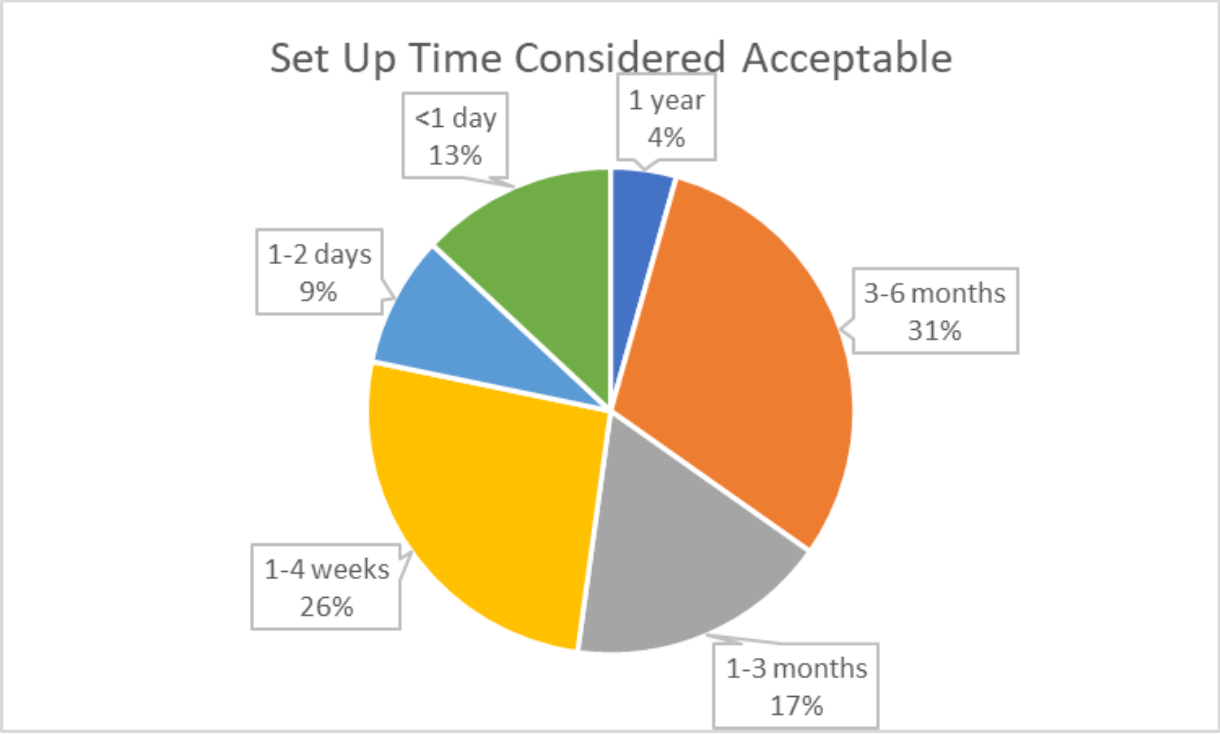


Figure 3.9 - The amount of time respondents considered acceptable to set up the project.

Respondents mentioned that the length of time they are willing to invest may change depending on a number of factors. They listed the following factors in their response. The two most mentioned factors are the output accuracy and funding availability. Six respondents said they would invest more time if automation would be able to yield accurate results. Another six respondents mentioned funding availability and expert assistance as a factor. This affirms Brönnimann et al. (2018) who mentioned that funding is usually limited for data rescue projects, which also can limit the development of the project. Therefore, a stable and continuous funding,

including for a team of experts, is crucial for data rescue projects to reach a long-term endeavor. Bernard Ogden, a citizen science researcher and software engineer at The UK National Archives, said that although the Archives' current funding model allows them to fund additional staff with automation expertise to help with tasks such as set up; he still finds it difficult to have experts maintain the system overtime. Another respondent also said they would love to use new technologies but only with expert assistance, and it was difficult to establish a team of experts with their limited funding. Overall, the six respondents said they would invest more time if funding and additional support staff were available.

Relatively fewer but enough respondents also said the amount of time they would like to invest depends on the scale (n = 4) and transferability (n = 2) of the project. They would invest their time if the scale of the project was large, so the setting up time will only be a small proportion of the whole. Two respondents thought that time spent setting up automation would be worth the effort if it was applicable to other documents. Ogden shared in the comment:

> One thing we might consider is that setting up automation for one set of records may open up automation of more records – if we can read the handwriting in early 20th century minute books, perhaps we can read other documents of that period.

There were respondents who would not choose automation (n = 4). They preferred human to machine digitization because they expressed the sentiment that humans would always outperform machines in many ways. Specifically, respondents stated that humans can learn through "diving into the data"; humans can add context. By contrast, machines are limited in their ability to really learn. They are also concerned that using automation will eventually make the transcription process take more time than human transcription due to the immature development of automated transcription. These concerns are based on respondents' experience with automation. People in the historical weather community have similar experiences with immature OCR (Brönnimann et al., 2006). Perhaps, their opinion can easily be changed if the machines perform like humans or better.

Overall, respondents appear open to the idea of implementing automated transcription if time allows but express concerns in terms of the amount of time they would invest. Respondents gave various factors that may influence their decision, such as accuracy, funding, scale of projects, and replicability. It is crucial to consider and optimize these factors before applying automated

approaches in the future. The main takeaway is to find a balance between these factors and the time needed to implement and maintain an automated approach.

Loss and Gain from the Shift from Human transcription to Automation

Early in the survey, respondents began talking about advantages and disadvantages in a shift to automated transcription. Respondents were overwhelmingly uniform in their responses to an open-ended question, "What, in your opinion, might we gain or lose if we started to use automation instead of humans (e.g., citizen science or paid transcribers)?"

I code this into categories: lose categories and gain categories. Some of the respondents talked only about gains, and similarly, some respondents talked only about losses. We'll start with the gains, although only half mentioned gains. One respondent wrote that "If automation sped up the digitisation process, I can't seen [sic] any losses." Another one wrote that "In my view the benefits far outweigh the downsides."

Almost half (42%, n = 21) said the most obvious advantage would be the ability to get more data in a shorter time with cheaper price. In other words, respondents think automated approaches would be time-efficient and cost-efficient. PiP Willcox, the head of research at The National Archives, said:

> …[W]e could gain a cost-efficient way of making large amounts of data computer-readable.

Another respondent said:

> We could gain more data of good quality, available to provide climate science, in a shorter period of time.

Another reason, as succinctly put by one respondent, is that automation will allow us to "skip the boring bits". Linden Ashcroft, a climate scientist and data rescue researcher from The University of Melbourne, said: "We would gain a lot of really useful data, and citizens who are involved in data rescue projects could focus on possibly more interesting aspects of climate history." Similarly, Ogden wrote that "We can potentially automate at much higher volume and we [can] free up a lot of person time to do something else." This clearly reflects the differences between data rescue, which is outcome focused, and citizen science, which is process and community

focused (Bonney et al., 2014; Brönnimann et al., 2018). In citizen science, the "boring bits" can be the interesting part of engagement.

Peter Siegmund, a climate researcher at Royal Netherlands Meteorological Institute, expressed this as well, “For some people citizen science is a main sense for life. These people will lose if automation is used. [However,] Science and society as a whole will gain.” This quote suggests how gains and losses can be intertwined. For it can be a matter of transcriptional results and public participation.

Thus, when discussing losses of using automation, respondents (32%, n = 16) agreed on the potential loss of public engagement and citizen interests if automation is used in the transcription process. Ashcroft commented that “We would lose an opportunity to engage people with the weather and climate of the past.” We would also lose “a very effective way of getting member of the public interacting with scientists”, as Andrew Matthews, a data rescue researcher at National Oceanography Centre, commented. Public engagement is quite important for citizen science projects as it has been pointed out by the respondents to be the second most important goal of citizen science transcription projects. Without the involvement, the public could lose a lot of learning opportunities. One respondent said:

> If the community is not involved, a good opportunity to make people concern and understand about the environmental problem will be lose [sic].

Another respondent also mentioned the loss of “educational and cultural importance of having people involved”. As such, it is a disadvantage if using automation will result in losing a way to educate the interested public.

With less public engagement, respondents (26%, n = 13) think we would not only lose opportunities to educate the interested public, but also their meaningful insights. If automation replaced human transcribers, respondents suggested we could lose meaningful information that is not classified as “observation”, such as interesting notes, anecdotal notes, and metadata notes. Stefan Brönnimann, a data rescue researcher at University of Bern, said we could lose valuable metadata that are extremely important for source inspection because it might not be recognized, and we cannot afford it to not be recognized by automated transcription. Similarly, Harry Smith, a social and economic historian at King’s College London, suggested we could lose detailed, hands-on knowledge of the records which often contain additional information that are hard to be

automatically transcribed but are often noted by volunteer transcribers. We could also lose the ability to handle uncertainties, questionable values, and odd edge cases. These valuable pieces of information could be lost because respondents think humans are better at dealing with uncertainties. For example, Ogden suggested we would lose the benefit of having human capacity to deal with unexpected inputs and the unexpected discoveries that a human might flag. This concern is empirically confirmed by the fact that valuable information may be lost in the process if the OCR cannot process the material properly (Brönnimann et al., 2006; Stickler, Alexander et al., 2014).

In conclusion, the respondents have a clear view of what we might gain and lose from shifting to automation. They are excited about the benefits automation would bring, but they are concerned about the loss of public engagement and their meaningful insights as well. The finding is that losses and gains seem to be a trade-off that needs to be carefully balanced when introducing automated transcription. The question is: does it have to be a trade-off?

Joaquin Osvaldo Rodriguez, a software engineer who has been dedicated to automating the data rescue process, has expressed in his response that we would lose nothing to try automation, not even public engagement. He suggested humans will always be engaged in the process, and the public can learn through the new process as well. This is a theory that has not been proposed before, and it may be interesting for future research to explore the trade-offs between loss and gains from the transition to automation.

3.4.4 Rejection of using AI

There are several respondents (n = 4) who firmly refuse using automation for future transcription under any circumstances. It is very interesting to find out that they have very similar concerns. The refusals are mostly coming from their experience. They know how automation works, and without exception they all have unsatisfied experience with automation in the past. They are worried that they will need to spend extra time and effort in employing the automated system compared to their current operation but ending up getting less accurate results. That is to say, these results will end up requiring more human-hours and efforts for quality control, and it is referred to as “fixing the mass” by respondent 26. Admittedly, many past attempts at automation have been unsatisfactory, so manual keying remains the most popular approach to transcribe

historical records (Brönnimann et al., 2018; World Meteorological Organization, 2016). It is only natural to reject automation at the moment and wait for a satisfactory solution.

3.5 Discussion: calls for a hybrid model

In the survey, many respondents (32%, n = 16) unpromptedly said that instead of having a computer fully handle the transcription, a human needed to be in-the-loop on this. In other words, they believe AI working under human supervision would be a better solution. The respondents used the word “hybrid” to describe this idea. This section will discuss this new concept, hybrid model, proposed by the respondents and its possible implementation in future projects.

As discussed, the majority of respondents are looking forward to applying automation in their projects. Researchers like the respondents urgently need more digitized data (Brönnimann et al., 2018), and automation has the potential to accelerate this process. Stephen G. Penny, a research scientist at NOAA, wrote:

> I am very supportive of a dedicated effort to automate the transcription process with AI/ML. This is a problem that is well suited for OCR and could have a big impact on climate science by improving climate reconstructions during a time where data were extremely limited.

He is one of the many researchers who is very supportive of automated transcription because it opens the opportunity of more transcribed records that were not available digitally. Similarly, Eric Freeman, a data rescue coordinator at NOAA, expressed a similar opinion in his comments. He said automation could be very useful in transcribing the historical records he is managing, and he was intrigued in automation being more efficient in rescuing data.

While most respondents are interested in automation because it can provide more useful data, only a fraction felt an immediate shift in transcription is needed. They believe that automation is the ultimate goal and that eventually we will need to look to automation for help. Specifically, some respondents are satisfied with their current approach (e.g., citizen science, paid transcriber) and are considering automation as an alternate option, while others feel the need to move to automated approaches because they believe it is impossible for humans to handle the vast number of historical records. Compo wrote in the comment:

> We would access … billions of weather records that are already imaged but will not be transcribed for years or decades because the number of citizen scientists and paid transcribers is much too small to handle the volume.

Praveen Teleti, a data rescue researcher at Cambridge University, also wrote:

> Automation is the answer, no amount of citizen-science projects can transcribe available logbooks.

As they wrote, the number of records that needed to be transcribed is too much for humans, so it is necessary for automation to step in. The vast amount of data that automation can provide would give a step change in data reanalysis and modeling. As respondent 23 commented, automation is strongly needed and is the future of data rescue.

However, as discussed in the previous section, there may be a trade-off in terms of what we gain or lose if we apply automation. By gaining more data with shorter time and cheaper price, we would lose the public engagement and non-experts' insightful comments. However, more than one-third of the respondents (36%) provided an aspect saying it does not have to be a trade-off. Kate Willet, a climate data and monitoring scientist at the Met Office, and Peter Thorne, a climate scientist at Maynooth University, wrote in their response that it should not be "either/or" when choosing between current transcription approach and automation. Respondents think a hybrid model - a "people with machine" approach would be the most appropriate and beneficial for current data rescue operations. They think automation will not replace humans; rather, there will always be a role for humans, as machines and humans will coexist in hybrid models.

To address the coexistence, respondents described what constitutes a hybrid model. Matthews wrote:

> I can't see machine learning fully replacing the capabilities of humans in transcribing data…. In the future, larger projects would probably work best using a hybrid approach where humans either train an algorithm, or deal with the cases where it fails, if they can be easily identified.

Similarly, Penny wrote:

> I would like to see an AI/ML solution that enhances the citizen science experience by providing AI/ML guided tools that allow the citizen scientist to transcribe an order of magnitude more records in the same amount of time. While a fully automated solution may be attractive, there is still benefit to human oversight and inclusion of the amateur community in this effort.

Both respondents emphasized that automation will not completely replace humans; instead, they will be working with humans as an enhancement. Even if humans are to be removed at the transcription stage, they will be added back in the previous or later step.

Without exception, respondents think it is important to add human oversight in the quality control step. They agreed that it is necessary to have humans inspect the automation results by reviewing and verifying the transcription since these records are produced by humans and would be best interpreted by humans. Joanne Williams, a data rescue researcher at National Oceanography Centre, said:

> We would probably need a hybrid model with the transcription checked by volunteers, as there are often problems beyond simple transcription (e.g., entries misaligned with boxes, corrections, explanatory notes).

Kevin Wood, a climatologist and data rescue researcher at NOAA, also wrote:

> Even with good accuracy I can imagine a role for [citizen scientists] in some [quality control] operations via a hybrid interface.

Some respondents think this hybrid model will be difficult; some think it will be easy. I believe it is difficult to build an automated transcription workflow from scratch because there are various constraints, such as time, funding, resources, etc. However, it will also be relatively easy because it is a hybrid model and human supervision can be very beneficial at any step to ease the complexity. Specifically, a hybrid model with human supervision will make it easier to detect common errors during and after the transcription process. It can also reduce the quality control time which some respondents concern that it might take too much time.

I think that participants found a middle point between current approach and automation where we would have the benefits of both approaches. We would be able to get more data faster and cheaper by applying automation in transcription, and we would also have public insights by

involving humans in the process. A hybrid model seems to be the best way of applying AI-enhanced transcription approaches. These are promises, but reality can be different. In a recent attempt to implement an automated transcription workflow, it was shown that a significant amount of effort was needed to build the automation from scratch (Y. Zhang & Sieber, 2022). This study shows that not only does it take a lot of effort to set up the workflow, but also the document preparation (e.g., image preprocessing, layout analysis, OCR). In addition, you may need to have experience in computer science to supervise the workflow, which implies that only a small percentage of researchers in the data rescue community are able to do this. Furthermore, Y. Zhang and Sieber (2022) said there is a lot of customization and parameterization as you build the workflow, so it can take more time and resources than manually keying in if the dataset is relatively small or you do not have the expert support. It may be a long road from manual keying to full automation, but it is hoped that the findings of this study and the suggestion of a hybrid model will make the process easier.

3.6 Conclusions

From this survey, I see how automation could bring a step change in historical data rescue with the help of willing researchers. The responses helped me understand the current status of data rescue projects, researchers' various opinions towards automation and, most importantly, the resistance and difficulties of shifting from current approach to AI-enhanced automation. This study is to examine what are researchers' attitudes to automation, and what are the resistances to the shift. This information will be a reference to interesting studies, and it will form a guideline for future automation projects to make the shift from traditional data rescue to AI-enhanced data rescue smoother and easier. Here I will draw conclusions to respondents' attitudes and opinions, delve into their implications, and then explore the future directions for applying AI-enhanced automation.

Firstly, most of the respondents have knowledge about AI or AI-enhanced data rescue, but they barely have experience in AI. The respondents, who are core members of data rescue projects, are predominantly using manual transcription paid or unpaid. Very few of them are using OCR or other automated transcription approaches for their research.

Secondly, the majority of the respondents said they are willing to try AI-enhanced transcription approaches, but there are resistances. The main resistance is the concern about

output accuracy of handwritten data due to the vast number of handwritten records and unsatisfactory experience in the past. Another resistance is the funding availability, and its ability to consistently support additional staff with AI expertise to build and maintain the automation system.

Thirdly, respondents reported a threshold number of hours for automated transcription (2.5 hours). If it takes more time, then it is not worth it. Researchers might reconsider applying automated approaches if the setting up time is longer. Although, participants said they might consider investing more time if the output accuracy is satisfactory and funding is sufficient. Currently, it is known that building up an automated approach takes more than 2.5 hours and requires more targeted expertise. Although the output accuracy of automation is slowly improving, there remains a huge gap between what promises and what it can currently deliver. We probably need a collective initiative like Zooniverse to make it happen.

Finally, respondents are uniform in what they expect to gain and lose if the shift to automation took place. They believe that automation would bring opportunities to have much higher volumes of transcribed data with less time and cost, and, correspondingly, researchers would lose the public engagement and their meaningful insights and comments.

The result of this study determines how researchers think of automation and provides a reference for future projects that are interested in applying automation. It also indicates that a hybrid approach might be the optimal solution instead of a pure automation approach. AI will work with humans, both researchers and citizen scientists, as an enhancement tool to provide better, faster, and cheaper transcribed data. These findings can be valuable to future data rescue projects and could possibly provide more than enough data for historical research.

Reference

Ashcroft, L., Coll, J. R., Gilabert, A., Domonkos, P., Brunet, M., Aguilar, E., Castella, M., Sigro, J., Harris, I., Unden, P., & Jones, P. (2018). A rescued dataset of sub-daily meteorological observations for Europe and the southern Mediterranean region,

1877–2012. *Earth System Science Data*, *10*(3), 1613–1635. https://doi.org/10.5194/essd-10-1613-2018

Black, N., Brazier, J., Fitzpatrick, R., & Reeves, B. (1998). Designing and using patient and staff questionnaires. *Black N, Brazier J, Fitzpatrick R, Reeves B. Health Services Research Methods-a Guide to Best Practice. London: BMJ Books*.

Blancq, F. L. (2010). Rescuing old meteorological data. *Weather*, *65*(10), 277–280. https://doi.org/10.1002/wea.510

Bonney, R., Shirk, J., Phillips, T., Wiggins, A., Ballard, H., Miller-Rushing, A., & Parrish, J. (2014). Next Steps for Citizen Science. *Science (New York, N.Y.)*, *343*, 1436–1437. https://doi.org/10.1126/science.1251554

Bowling, A. (2014). *Research methods in health: Investigating health and health services* (Fourth edition). Open University Press.

Brönnimann, S., Annis, J., Dann, W., Ewen, T., Grant, A. N., Griesser, T., Krähenmann, S., Mohr, C., Scherer, M., & Vogler, C. (2006). A guide for digitising manuscript climate data. *Climate of the Past*, *2*(2), 137–144. https://doi.org/10.5194/cp-2-137-2006

Brönnimann, S., Brugnara, Y., Allan, R. J., Brunet, M., Compo, G. P., Crouthamel, R. I., Jones, P. D., Jourdain, S., Luterbacher, J., Siegmund, P., Valente, M. A., & Wilkinson, C. W. (2018). A roadmap to climate data rescue services. *Geoscience Data Journal*, *5*(1), 28–39. https://doi.org/10.1002/gdj3.56

Bush, D., Sieber, R., Seiler, G., Chandler, M., & Chmura, G. L. (2019). Bringing climate scientist's tools into classrooms to improve conceptual understandings. *Journal of Environmental Studies and Sciences*, *9*(1), 25–34. https://doi.org/10.1007/s13412-018-0525-2

Camuffo, D., & Bertolin, C. (2012). The earliest temperature observations in the world: The Medici Network (1654–1670). *Climatic Change*, *111*(2), 335–363. https://doi.org/10.1007/s10584-011-0142-5

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Craig, P. M., & Hawkins, E. (2020). Digitizing observations from the Met Office Daily Weather Reports for 1900–1910 using citizen scientist volunteers. *Geoscience Data Journal*, *7*(2),

116–134. https://doi.org/10.1002/gdj3.93

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, K., & Wallin, J. (2011). Galaxy Zoo: Morphological Classification and Citizen Science. *ArXiv:1104.5513 [Astro-Ph]*. http://arxiv.org/abs/1104.5513

Fortson, L., Masters, K., Nichol, R., Edmondson, E. M., Lintott, C., Raddick, J., & Wallin, J. (2012). Galaxy zoo. *Advances in Machine Learning and Data Mining for Astronomy*, *2012*, 213–236.

Holley, R. (2009). How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, *15*(3/4). https://doi.org/10.1045/march2009-holley

Klijn, E. (2008). The Current State-of-art in Newspaper Digitization: A Market Perspective. *D-Lib Magazine*, *14*(1/2). https://doi.org/10.1045/january2008-klijn

Kwok, R. (2017). Historical data: Hidden in the past. *Nature*, *549*(7672), 419–421. https://doi.org/10.1038/nj7672-419

Larsen, J. D., Mascharka, C., & Toronski, C. (1987). Does the Wording of the Question Change the Numer of Headaches People Report on a Health Questionnaire? *The Psychological Record*, *37*(3), 423.

Linsky, A. S. (1975). Stimulating Responses to Mailed Questionnaires: A Review. *Public Opinion Quarterly*, *39*(1), 82. https://doi.org/10.1086/268201

Mateus, C., Potito, A., & Curley, M. (2021). Engaging secondary school students in climate data rescue through service-learning partnerships. *Weather*, *76*(4), 113–118. https://doi.org/10.1002/wea.3841

McColl, E., Jacoby, A., Thomas, L., Soutter, J., Bamford, C., Steen, N., Thomas, R., Harvey, E., Garratt, A., & Bond, J. (2001). Design and use of questionnaires: A review of best practice applicable to surveys of health service staff and patients. *Health Technology Assessment (Winchester, England)*, *5*(31), 1–256. https://doi.org/10.3310/hta5310

Neudert, L.-M., Knuutila, A., & Howard, P. N. (2020). *Global Attitudes Towards AI, Machine Learning & Automated Decision Making* (p. 10). Working paper 2020.10, Oxford Commission on AI & Good Governance.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education*, *33*(3), 301–314. https://doi.org/10.1080/02602930701293231

Odunayo, O., Sookoo, N. N., Bathla, G., Cavallin, A., Persaud, B. D., Szigeti, K., Van Cappellen, P., & Lin, J. (2021). Rescuing historical climate observations to support hydrological research: A case study of solar radiation data. *Proceedings of the 21st ACM Symposium on Document Engineering*, 1–4. https://doi.org/10.1145/3469096.3474929

Pinto dos Santos, D., Giese, D., Brodehl, S., Chon, S. H., Staab, W., Kleinert, R., Maintz, D., & Baeßler, B. (2019). Medical students' attitude towards artificial intelligence: A multicentre survey. *European Radiology*, *29*(4), 1640–1646. https://doi.org/10.1007/s00330-018-5601-1

Roopa, S., & Rani, M. (2012). Questionnaire Designing for a Survey. *Journal of Indian Orthodontic Society*, *46*(4_suppl1), 273–277. https://doi.org/10.5005/jp-journals-10021-1104

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Prentice Hall, N.J.

Ryan, C., Duffy, C., Broderick, C., Thorne, P. W., Curley, M., Walsh, S., Daly, C., Treanor, M., & Murphy, C. (2018). Integrating Data Rescue into the Classroom. *Bulletin of the American Meteorological Society*, *99*(9), 1757–1764. https://doi.org/10.1175/BAMS-D-17-0147.1

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. (2019). A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition*, *94*, 122–134. https://doi.org/10.1016/j.patcog.2019.05.025

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *ArXiv:2103.15348 [Cs]*. http://arxiv.org/abs/2103.15348

Stickler, A., Brönnimann, S., Valente, M. A., Bethke, J., Sterin, A., Jourdain, S., Roucaute, E., Vasquez, M. V., Reyes, D. A., Allan, R., & Dee, D. (2014). ERA-CLIM: Historical Surface and Upper-Air Data for Future Reanalyses. *Bulletin of the American Meteorological Society*, *95*(9), 1419–1430. https://doi.org/10.1175/BAMS-D-13-00147.1

Stickler, Alexander, Brönnimann, Stefan, Jourdain, Sylvie, Roucaute, Eméline, Sterin, Alexander M, Nikolaev, Dmitrii, Valente, Maria Antónia, Wartenburger, Richard, Hersbach, Hans, Ramella Pralungo, Lorenzo, & Dee, Dick P. (2014). *ERA-CLIM Historical Upper-Air Data 1900-1972, supplement to: Stickler, Alexander; Brönnimann, Stefan; Jourdain, Sylvie; Roucaute, Eméline; Sterin, Alexander M; Nikolaev, Dmitrii; Valente, Maria*

Antónia; Wartenburger, Richard; Hersbach, Hans; Ramella Pralungo, Lorenzo; Dee, Dick P (2014): Description of the ERA-CLIM historical upper-air data. Earth System Science Data, 6(1), 29-48 [Application/zip]. 813 datasets. https://doi.org/10.1594/PANGAEA.821222

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections*, 179–204. https://doi.org/10.1515/9783839455845-008

Wilkinson, C., Brönnimann, S., Jourdain, S., Roucaute, E., Crouthamel, R., Brohan, P., Valente, A., Brugnara, Y., Brunet, M., & Team, I. (2019). *Best Practice Guidelines for Climate Data Rescue*.

Williams, A. (2003). How to … Write and Analyse a Questionnaire. *Journal of Orthodontics*, *30*(3), 245–252. https://doi.org/10.1093/ortho.30.3.245

World Meteorological Organization. (2016). *Guidelines on Best Practices for Climate Data Rescue*. https://library.wmo.int/doc_num.php?explnum_id=3318

Yüzbaşıoğlu, E. (2021). Attitudes and perceptions of dental students towards artificial intelligence. *Journal of Dental Education*, *85*(1), 60–68. https://doi.org/10.1002/jdd.12385

Zhang, B., & Dafoe, A. (2019). Artificial Intelligence: American Attitudes and Trends. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3312874

Zhang, Y., & Sieber, R. (2022). *A guideline for AI-augmented end-to-end historical handwritten tabular data (digits) recognition workflow tested by DRAW dataset: Rescuing historical climate observations* [Manuscript in preparation]. McGill University.

Preface to Chapter 4

Chapter 4 presents an AI-augmented historical records transcription workflow and its testing process to address the research question: “If AI-augmented data rescue is useful then what might an automated system look like?” This chapter builds on the result of the previous chapter to further reveal the opportunities and challenges of building an AI-augmented workflow. The creation and testing of this AI-augmented workflow allows researchers to bridge the gaps between multiple disciplines and work together to rescue more historical records.

This chapter was co-authored with my supervisor, Dr. Renee Sieber, as a peer-reviewed journal article. We plan to submit the manuscript to *Geoscience Data Journal*.

Chapter 4. A guideline for AI-augmented end-to-end historical handwritten tabular data (digits) recognition workflow tested by DRAW dataset: Rescuing historical climate observations

Abstract

Millions of valuable weather historical records exist in paper format. Those historical records are gold mines where so much valuable information is buried with them. Data rescuers worldwide have been working hard trying to retrieve those data and transcribe them into digital format to make it easier to preserve, search and analyze. A great portion of the records is written by hand, in print or cursive handwriting. Automatic transcriptions to date have not been reliable or sufficiently accurate on handwritten data, so most of the historical records are currently being keyed in manually. Attempts have been made to integrate artificial intelligence (AI) to automatically transcribe the historical records, but users have to start from scratch, and the results have not been promising. Unfortunately, there is currently no end-to-end workflow to automatically transcribe historical handwritten tabular records into digital datasets for other researchers to build upon. Here I proposed a workflow that uses AI to automate the handwriting transcription process. The workflow is tested using the historical climate records from the Data Rescue: Archives and Weather (DRAW) project, and it can be adapted to other handwritten dataset with proper adjustment as well. This workflow is composed of five steps with EAST algorithm (Efficient and Accurate Scene Test Detector) and Tesseract OCR (Optical Character Recognition) embedded, and it slices the tabular data into rectangles and transcribes them into digital format. These five steps are discrete steps that I broke down to better accommodate future advances (e.g., new training data, better layout detectors). We hope the workflow proposed can also be a guideline that is easily replicable and can be utilized to transcribe other historical datasets.

4.1 Introduction

Historical records are like gold mines. These data are “hidden in the past” waiting for scientists to unearth them out and reveal their value (Kwok, 2017). This valuable information is usually stored in paper form and is difficult to analyze using modern technology. In addition, paper records are subject to degradation and deterioration if not properly preserved, which can lead to the loss of important records. The paper format can also hinder the availability of records for sharing worldwide. One of the unsolved issues for libraries to date has been sharing the textual content of scanned historical records in a format that can be readily used by researchers (e.g., digital spreadsheets, charts) (Fischer et al., 2014). Therefore, researchers have found it important to transcribe historical records into a machine-readable format that can be stored digitally and shared freely. Such a shift would facilitate the organization, archiving and retrieval of historical records in archives, libraries, and repositories around the world (J.-A. Sánchez et al., 2013; Singh et al., 2012; Swindall et al., 2021). It can also benefit society by raising public awareness and a sense of social responsibility (Yasser et al., 2017).

Researchers are also aware of the benefits of transcribing historical records, and many organizations and initiatives are actively transcribing historical records into digital formats (Brönnimann et al., 2019; Wilkinson et al., 2019). Research has shown that by far the most popular method is to manually key records into digital format (World Meteorological Organization, 2016). This includes the recent popularity of citizen science and crowdsourcing projects where volunteer transcribers help transcribe these records, such as Data Rescue: Archives & Weather (DRAW), Old Weather, and Rainfall Rescue (Wilkinson et al., 2019). These projects are much less costly than traditional transcription projects that hire people to transcribe, and they have contributed many valuable datasets.

However, challenges remain with current projects to rescue historical records through manual transcription. I summarized these challenges into three categories. The first challenge is the amount of time required for such projects. One obstacle to current historical data rescue projects is that they are often time consuming (Chen et al., 2018; Fischer et al., 2014; Jenckel et al., 2016; Odunayo et al., 2021). The time may vary from three months to several years, depending on the size of the project. From a researchers’ perspective, this may hinder valuable research

opportunities, as some data may be time sensitive. It is ideal to have machine-readable, readily available data as quickly as possible.

The second challenge is the manpower required for historical data rescue projects. Data rescue projects are very labor-intensive: it usually needs a large number of people, including experts, to complete the transcription of a set of historical data (Chimani et al., 2021; Jenckel et al., 2016; Odunayo et al., 2021). Recruiting transcribers and experts is difficult and expensive. Because data rescue projects are often underfunded, it can be difficult to recruit enough people to complete transcription in a reasonable amount of time. Similarly, for citizen science projects, the project can last a very long time, depending on how many dedicated volunteer transcribers come to help. These issues will cause delays in the delivery of transcribed datasets. Delays in dataset availability will make it difficult for researchers to keep up with current research and limit the time span of studies.

The third challenge is the effort needed for interaction and training. For citizen science projects, a great deal of effort is needed to build a transcription platform like the DRAW or Zooniverse projects from scratch. In addition, a significant effort is needed to provide webinars or tutorials for incoming transcribers so that they can start transcribing selected historical records with less complications. These steps will require substantial resources and manpower, which is not ideal (Terras, 2022). It can be particularly difficult for large projects given the huge efforts required to maintain them, which limit the scalability of manual transcription (Dahl et al., 2021).

The World Meteorological Organization (WMO) alluded that artificial intelligence (AI) augmented transcription may be a way for future research to overcome current challenges (World Meteorological Organization, 2016). There is also a growing number of studies and benchmarks on transcribing paper records, suggesting that AI can be an ideal solution to current data rescue and needs future improvements and developments (Swindall et al., 2021; Yasser et al., 2017). For example, incorporating AI at the scanning phase, possibly with multiple cameras, could help to alleviate some of the image quality issues, such as the shadows, and allow for a hardware solution that minimizes the amount of deleterious artifacts. Researchers believe that AI can alleviate the gap in manual transcription, but this raises new concerns. Many benchmark algorithms have only been tested for processing modern documents and rarely for historical documents (Gatos et al., 2014). Of these, even fewer deal with handwritten records, let alone historical cursive handwriting. Some studies focusing on historical records pointed out that

transcribing handwritten documents, especially cursive documents, is much more difficult than printed documents, and that a large portion of historical documents are handwritten (Chamchong et al., 2019; Firmani et al., 2017; Lehenmeier et al., 2020). The variety of handwriting styles and the poor condition of the original documents make historical handwritten document transcription even more difficult (Alabau & Leiva, 2012). This is a gap that has not been well studied and explored, and researchers expect that AI should be able to help address it.

The goal of this paper is to develop a generic workflow that can automatically transcribe scanned historical handwritten documents into machine readable datasets with good robustness for different types of documents and handwritings. I propose a customized strategy to digitize scanned handwritten tabular records by bringing together different machine learning models. In this work, my contribution is to combine layout analysis and text recognition, two of the most important steps found by data transcription studies but often researched separately, into a robust workflow that employs machine learning techniques to improve performance. I am not concerned with achieving state-of-the-art accuracy rate or examining the performance of different machine learning. Instead, I focused on integrating layout analysis and text recognition with machine learning techniques into a robust workflow to transcribe historical handwritten tabular documents. In this sense, the workflow will serve as a guidance for future research on automated historical document transcription, and the performance can be improved through proper hyperparameter tuning, model training, and algorithm optimization. It will also serve as a benchmark for future research at the intersection of historical data rescue, citizen science, and machine learning.

This workflow is tested on weather records from the McGill Observatory. These observations contain valuable information on the development of science, historical climate studies, historical environmental changes, and their social impacts. The scientific significance of these observations is crucial to understanding long-term climate change over the past centuries, the environmental history of nineteenth century Canada, and the evolution of weather forecasting. Knowledge of these registers is also useful for studying the historical weather records of other countries, as the format of these registers is standard and typical of weather registers from the late nineteenth and early twentieth centuries worldwide.

The paper is organized as follows. I will first summarize past experiences and related works in automating the transcription process. Then, I will present the DRAW dataset on which I tested

this workflow. The next section describes my attempts at building this transcription workflow, which includes detailed documentation of the approaches that were successfully implemented and those that I tried but were not promising. This workflow I built consists of five discrete steps, each of which is well documented and maintained. The discussion section then evaluates the performance of this workflow from three perspectives: input driven, output driven, and model driven. Finally, this paper summarizes the challenges that readers may face when trying to reproduce this workflow and talks about future directions and next steps.

4.2 Related works

The idea of applying automation in data rescue has been extensively discussed in the last decades, especially in climatological studies. However, most projects still rely on manual methods, such as hiring transcriptionists or recruiting citizen science volunteers to key records online or offline in a given template (Brönnimann et al., 2006; World Meteorological Organization, 2016). This is partly because the application of automated data rescue has not been widely tested: only a few projects have attempted to automate the data rescue process (e.g., OCR) in their projects (e.g., Stickler, Alexander et al., 2014; Wilkinson et al., 2019). From these projects, the researchers concluded that the automation techniques (e.g., OCR) currently used in data rescue projects are still under development and need to be improved. Issues identified when testing automated approaches include low accuracy, which is mainly due to poor image quality, difficulty in handling handwritten material, and not being robust enough to handle different formats (e.g., table, text) and content (e.g., digits, characters, alphanumerics) (Blancq, 2010; Stickler et al., 2014; Stickler, Alexander et al., 2014). Regardless of these difficulties, researchers believe that technological advances will help improve the performance of automated data rescue approaches and that state-of-the-art techniques, such as AI, could be the future direction for refining the transcription of historical records (Chimani et al., 2021; World Meteorological Organization, 2016).

AI has been used in many fields, and it can make many things possible that were once impossible. Automatic transcription used to be one of those impossible things. In a 1985 study, researchers found that OCR could not handle a variety of fonts and formats, but they believed that with AI enhancement and proper training, OCR would be able to process a variety of fonts and formats (Harnett, 1985). Later studies have proven this hypothesis. Studies have shown that

AI enhanced OCR has witnessed tremendous improvements and breakthroughs in recent years, and current research on refining OCR is almost always integrated with AI techniques (Memon et al., 2020; Terras, 2022). This proves that AI can be of great help for data rescue projects. A huge advantage of AI augmented data rescue over manual rescue is the amount of time it can save (Fischer et al., 2014). This is very beneficial for research that is time-sensitive or projects that have a large number of records, as manual transcription is almost impossible in terms of time or financially. Therefore, AI may be a great solution to improve automated data rescue projects, and it is crucial to investigate AI-augmented transcription approaches to refine current data rescue projects.

Although there are many benchmark algorithms for OCR and other techniques, there is no all-in-one solution for historical records transcription. In order to provide an end-to-end solution, techniques such as layout analysis and character recognition need to work in combination (J. A. Sánchez et al., 2019). One challenge in providing an automated solution for historical records transcription is that current studies have focused on improving the model and performance of each single algorithm rather than having them work in combination (Dahl et al., 2021). Layout analysis, in particular, is often overlooked when developing an OCR transcription algorithm, such as Tesseract, paddleOCR, and easyOCR (Shen et al., 2021). These algorithms are not equipped with powerful layout analysis capabilities. Especially for tabular data, it was found that transcription tools such as Transkribus, and OCRopy failed to detect the table and locate the text (Lehenmeier et al., 2020). To address this issue, many studies have suggested that we should develop a workflow that includes the necessary combination of techniques for document transcription, which includes but is not limited to preprocessing, layout analysis, text recognition, and post-processing (Chamchong et al., 2019; Neudecker et al., 2019). This idea has been tested in multiple languages and formats and has proven to be very helpful. It provides a direction for future attempts to automate historical records transcription.

Layout analysis is an important part of the transcription workflow. Layout analysis identifies regions of interest and passes them to the recognition algorithm in a recognizable format. On the one hand, there are many state-of-the-art deep learning algorithms for layout analysis, the most commonly used benchmark examples being TableNet (Paliwal et al., 2019), CascadeTabNet (Prasad et al., 2020), and LayoutParser (Shen et al., 2021). Test cases reported that these advanced algorithms sometimes failed to detect the object of interest. Therefore, later studies

have attempted to improve these algorithms by fine-tuning the models, and they have proven to be successful (e.g., Odunayo et al., 2021; Ziomek & Middleton, 2021). On the other hand, there are also off-the-shelf software packages for layout analysis (e.g., Transkribus, ABBYY FineReader). These software also use AI techniques, and test cases show that they do not perform consistently and may introduce errors (e.g., Lehenmeier et al., 2020; Odunayo et al., 2021). Two challenges raised by many studies are analyzing borderless tables and collecting training data to fine-tune the algorithm. There is currently no specific solution to analyze unbounded tables (e.g., tables with faded boundaries), and this is a gap that needs to be filled in future studies (J. A. Sánchez et al., 2019). The challenges of training datasets will be discussed in detail in a later paragraph.

OCR is another essential part of the workflow. OCR converts text (e.g., printed, typewritten, handwritten) from a document (e.g., scanned images, photographs) into a machine-encoded format so it can be stored and analyzed digitally. It has been used in many fields such as recognizing license plates, invoices and legal documents (Singh et al., 2012). With the development of OCR, researchers have started to explore its capabilities in preserving historical records (Chimani et al., 2021; Swindall et al., 2021; Yasser et al., 2017). In the past decades, AI techniques have contributed significantly to the advancement of OCR (Memon et al., 2020). Recent state-of-the-art OCR algorithms have made extensive use of deep learning algorithms such as convolutional neural networks (CNNs) (e.g., Firmani et al., 2017), recurrent neural networks (RNNs) (e.g., Parthiban et al., 2020), and hybris of CNNs and RNNs (e.g., Chamchong et al., 2019). However, studies have found that even advanced algorithms are challenging to recognize historical handwritten documents (Alabau & Leiva, 2012; Firmani et al., 2017; Jander, 2016; Swindall et al., 2021). This problem remains a gap, but research suggests that it can be solved by training and fine-tuning the model with proper training datasets (Terras, 2022). Again, this leads to the challenge of collecting training datasets, which will be covered in the next paragraph. Studies also point to future directions for OCR improvement and refinements, that is, making it more robust and easier to use (Jenckel et al., 2016; Memon et al., 2020).

Of all the challenges in building a transcription workflow, the most pressing difficulty is the collection of training datasets. Studies have proven that building a customized training dataset and using it to refine the model is beneficial because it improves accuracy, shortens runtime, and makes the algorithm more robust (Fornés et al., 2017; Holley, 2009; Prasad et al., 2020; Shen et

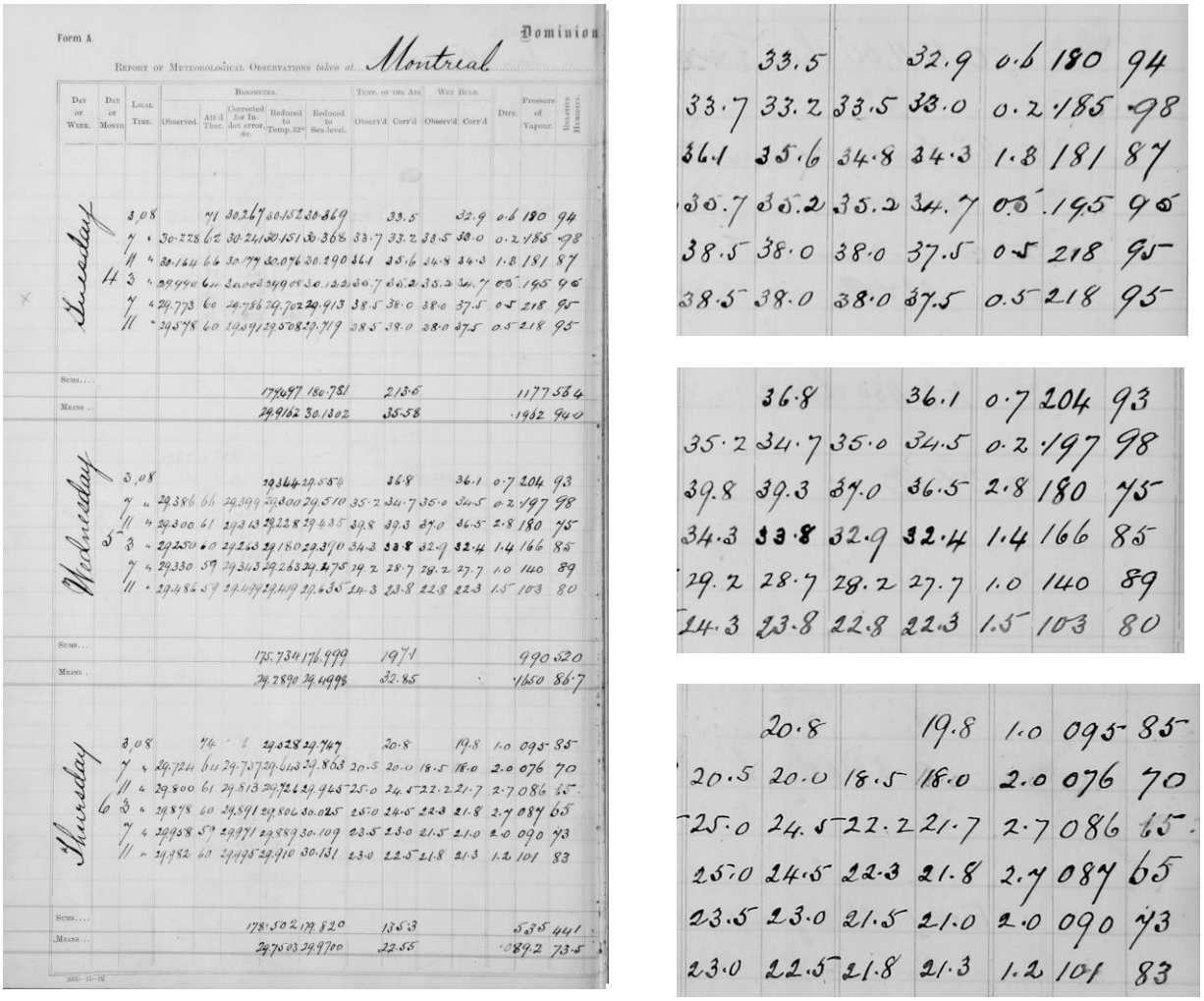
al., 2021; Terras, 2022). However, this is a challenging task in both layout analysis and OCR. One challenge is that there is no dataset suitable for cases, so in most cases it is difficult to find suitable existing training datasets (Dahl et al., 2021; Fischer et al., 2014; Nikolaidou et al., 2022). However, there is no easy way to create training datasets from scratch without prerequisite knowledge (Shen et al., 2021). In addition, it is often difficult to collect enough training data, especially for small collections (Lehenmeier et al., 2020). Even if researchers are able to collect enough data, building a training dataset is very expensive, labor intensive, and time consuming (Swindall et al., 2021; Yasser et al., 2017). Therefore, the difficulties in the training step represent a remaining gap in the development of transcription workflows. Possible solutions such as transfer learning have been proposed and tested as effective (e.g., Ströbel et al., 2022) and it is worthwhile for future studies to explore and find solutions to fill this gap.

4.3 DRAW dataset - paper registers at McGill Observatory

The McGill observatory weather logbooks (hereafter referred to as "registers") are housed at the McGill University Archives and are classified as accession number 1491. These registers are recorded and maintained in paper format in tens of thousands of pages contained in hundreds of registers. The DRAW datasets are those registers that contain sub-daily observations from 1874 to 1964. The registers were categorized into five main register types with a total fourteen subtypes to distinguish the various ways of recording observations. The data are recorded sub-daily in tabular format, primarily containing numbers, although there are letters and words as well (e.g., cardinal directions for wind; descriptions of weather conditions). These records are a part of Canadian cultural legacy and represent an important part of the scientific heritage of McGill University, City of Montreal, and Canada.

In this study, I focused on register type 150, which contains weather records from November 1st, 1884 to December 31st, 1899 (Figure 4.1a). The registers are A2-sized ledger books with three days of observations on each page. The tables are densely packed with digits and are written in Palmer style cursive script. These registers are estimated to contain 1,295,658 observations in 3707 pages of scanned images. It is slightly more than one-third of the entire DRAW data collection. These registers were scanned and stored as tiff and jpeg images, where each image shows a page of the registers. Observations were recorded sub-daily, including barometer, humidity, wind, clouds, rain, snow, and many other readings. In my case study, I selected air

temperature, wet bulb, vapor pressure, relative humidity, and other calculated variables because it included integers and float types. I decided not to focus on the alphabetical characters because, as will be described below, there are no image training datasets for Palmerian alphabetical fonts.



(a) One sample page of the DRAW ledger sheet. (b) Three crops are created per page.

Figure 4.1 - Examples of register type 150 weather records from the DRAW dataset.

4.4 A workflow for future references

The task is to develop a generic and robust workflow to help extract information from historical records using machine learning techniques. There are in total five steps for this workflow: (1) image preprocessing; (2) text line segmentation; (3) bounding boxes detection; (4) optical character recognition; (5) data rearrangement. Figure 4.2 shows a diagram of this workflow. The

result of each step will be the input to its next step. The layout analysis, i.e., the detection of text blocks or regions prior to OCR, is divided into the second and third steps here, as there is no off-the-shelf layout analysis capable of handling the DRAW dataset. Details will be explained in later sections.

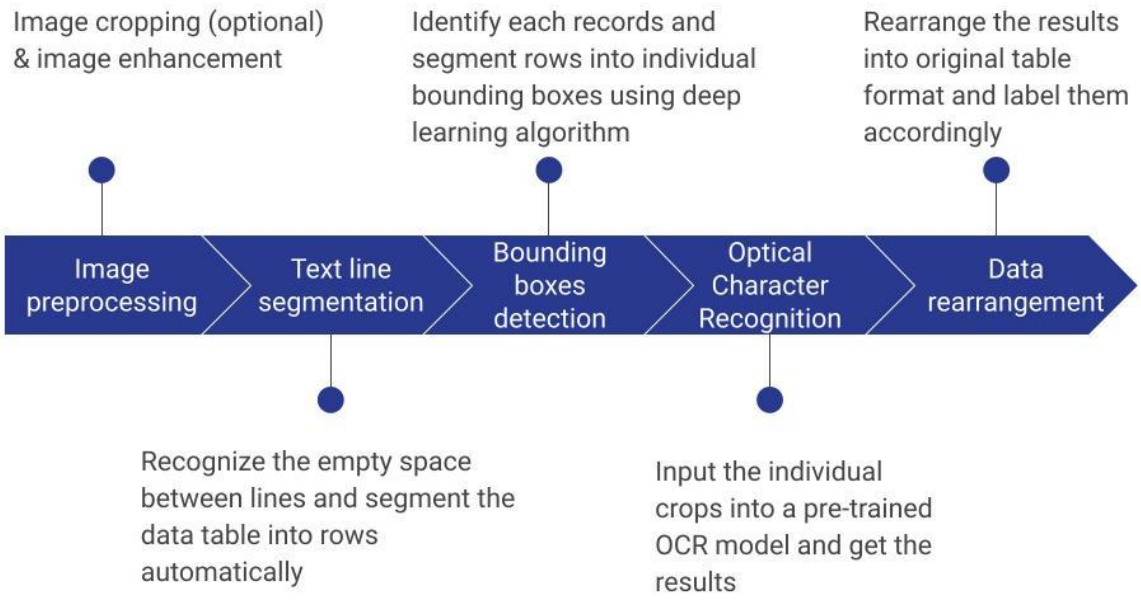


Figure 4.2 - A diagram of this workflow.

Rather than focusing on state-of-the-art performance and accuracy, this workflow emphasizes providing real solutions to the pressing needs of historical records transcription. This section provides details about the proposed workflow. The workflow went through many experiments with various techniques and models, and I have also documented the challenges encountered and the corresponding solutions in this section. In each step, I first describe its purpose. Then I describe the experiment. The python scripts are available at https://github.com/y2749zha/dareOCRthesis.

Prior to imagining a workflow, I experimented with Transkribus because its OCR engine is trained on handwritten cursive.[12] While the character recognition was superior to the LSTM models I eventually used (see below), I found that the layout analysis was not sufficient. Many

[12] https://readcoop.eu/transkribus/

observations in the table were not detected in the first place. This is what drew me to conceive of a workflow.

4.4.1 Image preprocessing

One may think of OCR as text processing, but it actually is image processing. Preprocessing of images is a transformation taken to improve the quality of the original image so that it can be more easily analysed in the later steps. In general, image preprocessing includes image resizing, grayscale conversion, image enhancement, and many other transformations. This step is necessary because the quality of the input images can greatly affect the quality of the output results (e.g., Klijn, 2008). Holley (2009) mentions two common types of image handling that are very useful for OCR workflows, namely grayscale processing and blurring. **Grayscaling** which is the conversion of an image containing any colour to shades of grey. Grayscale images are better and easier to handle in identifying text lines or characters (e.g., grey may be preferable to black and white since grey can better capture light touches of a pen) and can help reduce the complexity of later steps. **Blurring**, also known as image smoothing, is the convolution of an image with a low-pass filter kernel, which is highly effective at removing noise in an image. The preprocessing also can eliminate possible errors for later analysis, which will save time and effort significantly.

I test the image preprocessing step with the scanned images in the DRAW image repository. In this case, I crop the image only to the variables that I decide will be transcribed for this test. Due to the large size of each ledger (A2 size), I choose to work with a subset of variables that have different numerical ranges, including integers and decimals. By including both integer and float type numbers in one page, I hope the workflow testing will be more robust as innovations appear. The selected variables are cropped from the registers into rectangular images. There are three crops per page because each register has three days of observations per page, as shown in Figure 4.1b. The crops are labelled according to the date of observation, which will be important later for traceability and rearrangement. One also could crop or resize to create consistent dots per inch or consistent margins.

I then subject these crops to image enhancement (see Figure 4.3 for an example). The accuracy of the recognition result depends heavily on the quality of the original image (which of course depends on the quality of the original ledger sheet). Quality here refers to pixel noise and

character edges. Less pixel noise and sharper character edges is a precondition for OCR accuracy. I preprocess images with the OpenCV library,[13] a computer vision software library, as it is a widely used software library for computer vision that supports multiple interfaces and most operating systems.

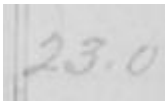


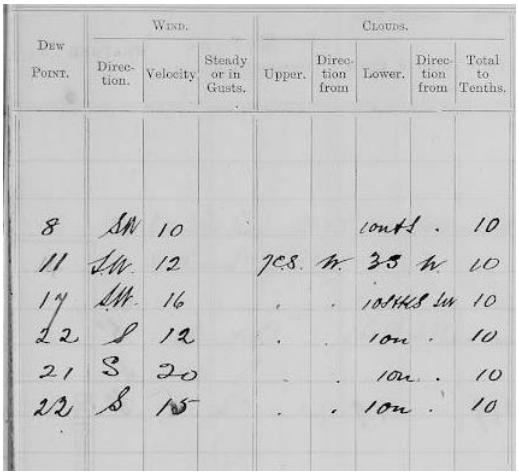
Figure 4.3 - Example of a value that may require enhancement (could be a number that entered in pencil that was erased).

I used two preprocessing techniques to increase machine readability of the images, grayscaling and blurring. We performed grayscale processing on the DRAW registers, and then experimented with blurring on the grayscale images. I read the OpenCV documentation, which has examples of noise and blurring remedies. We manually examined a few crops and found they contained what is called salt-and-pepper noise. So, we chose a median blur, which is very effective against that noise. However, those steps may not be necessary for other documents that are not from DRAW. This preprocessing is contingent on the quality of the input images. The user will need to do a manual inspection of a sample of pages to determine which processing is optimal. Experimentation may also be needed to determine the specifications of the processing. The choices here or in the later steps can be interactive. In other words, experiments in the subsequent steps can modify the prior steps. For example, in this step, various morphological transformations are initially experimented to enhance the detection of text lines and bounding boxes. However, it does not work well with the OCR step, so I put these transformations as an intermediate layer in the text line segmentation step, which will only be a guidance to the segmentation and will not be reflected in the segmented stripes.

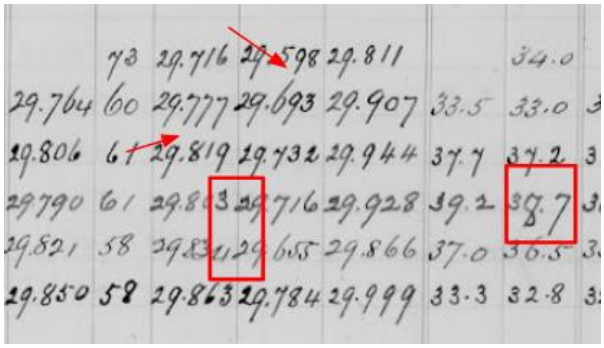
One challenge found after preprocessing is the faded table border shown in Figure 4.4a. In general, a well-defined table boundary as well as content is necessary to recognize the table content. Even after we did preprocessing, we found that the table outlines of the registers were blurred. This is a problem for off-the-shelf computer algorithms to accurately identify the layout of the tables. We tried several state-of-the-art software and algorithms for layout analysis (e.g.,

[13] https://opencv.org/

TableNet, CascadeTabNet, LayoutParser) to detect the table structure of the DRAW registers. However, none of these software and algorithms can detect every entry of the table. They are not suitable to process handwritten tabular documents with faded table outlines like the DRAW dataset. This makes it difficult to automate the transcription process. Therefore, special care needs to be taken in layout analysis for historical records. Layout analysis is the detection of the document structure, in this case the detection of text regions. To solve this problem, I separated layout analysis into two steps - text line segmentation and bounding box detection - to achieve recognition of each table entry. The details are explained in the following section.



(a) Poorly defined table boundaries.



(b) Values that are not well aligned.

Figure 4.4 - Challenges of layout analysis.

In this step, a total of 37 pages are processed into 100 crops.

4.4.2 Text line segmentation

Text line segmentation is an important step in document image processing, where text lines are individually detected and segmented. It also can be used for tabular records where the table lines are blurred or faded. Instead of detecting table lines, text line segmentation detects the gaps between text lines and automatically splits the data table into rows based on the invisible grid. This step is necessary because it can help overcome the challenges in analysing the layout of historical records. We take for granted that OCR can easily identify the boundaries of a number or a word. However, Shen et al. (2021) and Lehenmeier et al. (2020) tested and demonstrated that current OCR software and algorithms are not equipped with sufficient layout analysis. OCR

may use lines as a guide to that identification. One of the major drawbacks of historical tabular records is that their table boundaries are not well defined and often faded, as shown in Figure 4.4a. As a result, it is difficult for OCR to locate the text in the tables. Another challenge is the values that are not well aligned with the rows and columns. Each cell of the table should contain only one value. However, due to varying writing styles of different observers, some cells may have more than one reading, as shown in Figure 4.4b. Some of the values also may be in between rows or columns. This can interfere with the OCR. Therefore, text line segmentation is an important intermediate step to make the text ready to be processed by the character recognition algorithms. It also should be robust. The hope is that, with minor modifications, the segmentation algorithm should be able to be used for other similar records as well.

Two common means of aiding text segmentation are thresholding and reducing. *Thresholding* is the process of assigning pixel values to zero or maximum value based on the provided threshold value. It is only performed on grayscale images. Thresholding distinguishes the pixels we are interested in from other pixels. *Reduce* function reduces the image into a set of one-dimensional vectors.

In testing the DRAW registers, I found that I needed to further refine the image processing to prepare the data for this step. These processes are not reflected in the segmentation result; it only serves as an intermediate layer for text line detection. The results of the image preprocessing step first undergo some morphological transformations, which is optional and depends on the conditions of records. For example, *opening* is applied to remove the noise next to the object to make the segmentation more accurate. Here, a 5 * 5 rectangular kernel is used for the opening transformation. Other transformations include *erosion* and *dilation*. *Thresholding* is performed after the morphological transformations, which highlights the text to some extent and helps later operations to detect the text lines. The result of the thresholding process is passed to the *reduce* function, where the upper and lower edges of the text line are calculated and detected according to a threshold value adapted to the DRAW registers.

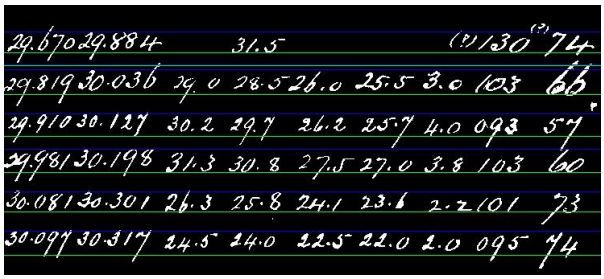
In our case, thresholding makes it easier to identify characters and text lines. Otsu's thresholding is used in our test to help detect gaps between rows. It distinguishes the pixels we are interested in from other pixels and avoids having to choose a value by automatically determining an optimal threshold value. Thresholding is also required for reducing. We used REDUCE_AVG, where the average pixel value of each row is calculated and stored in the output array. The gaps

can be calculated and detected by comparing the average pixel value of each row, where zero represents black and 255 represents white in grayscale images.

Originally, I thought this would be a process of subdividing the crops into strips. So I tested the DRAW registers on *single split*, where it effectively divides the crops into discrete and non-overlapping strips. As shown in Figure 4.5a, this approach divides the text line in the middle of the gap between upper and lower contours by averaging the upper and lower contour lines. This is usually used for well aligned, neatly written documents, where no noise (here splotches but also parts of characters) are found in the gaps. However, for pages that are not neatly written or have ink spots in the middle of two rows, this *single split* approach does not remove unwanted noise in the gaps that will interfere with the recognition step. I then came up with a *double split* method. *Double split*, as shown in Figure 4.5b, segments the text lines according to its contour. It helps truncate the outline bits of characters and makes the writing of the resulting rows more uniform. Therefore, the double split is usually used for documents that are not aligned neatly or have characters that are frequently out of the line.

Both of these splitting approaches are calculated based on the upper and lower contours returned by the reduce function. In the testing of DRAW pages, I switched between these two segmentation methods depending on the condition of the page to make the segmentation result neater. I tested both methods through the subsequent steps and found that the single split would suffice for this particular register type. By making the results neater, the bounding box detection and character recognition steps will be able to process the images with ease.

(a) Example of single split.



(b) Example of double split. The rows from blue lines to green lines are the output result.

Figure 4.5 - Examples of line segmentation approaches.

Overall, recent research on layout analysis tends to downplay the segmentation of text lines and focus on using neural networks to identify all elements in a page (e.g., Paliwal et al., 2019; Prasad et al., 2020). Methods that focus on segmentation are also using neural networks (e.g., Dahl et al., 2021; Fischer et al., 2014; Lehenmeier et al., 2020). As suggested by Shen et al. (2021), reproducing and adapting an existing deep learning algorithm can be time-consuming and frustrating, and there is no easy way to finetune and retrain the model. Neural networks in this case are too convoluted for reproduction, as I found it sufficient to use the basic algorithm in this step. I discovered that both single split and double split can efficiently segment the text line for the DRAW registers and thus save the time and resources needed to train and set up deep learning models.

I tested columnar segmentation on DRAW crops, but I found the result not as promising as row segmentation. The biggest problem is that in some edge cases, a column of floating numbers can be split into two columns at the decimal point due to the way it is written. This can truncate numbers or words if they overlap with nearby cells. In contrast, row segmentation is less prone to similar problems and is very accurate.

The segmented rows are numbered according to the observation date and number of rows automatically by the algorithm. This nomenclature ensures traceability, so that any image can be found in the original record at a later step. It will bring context to images that are difficult to recognize in the later step, which is important for data quality control and future error checking. This will also enhance the sustainability of any data rescue workflow.

In this step, the 100 crops are segmented into 581 rows by single split.

4.4.3 Bounding boxes detection

Bounding box detection is a common, albeit hidden, step in OCR. A bounding box is a rectangle that contains the object on the image to be detected. It makes it easier for computer algorithms to find what they are looking for and saves computational resources. In effect, bounding box detection is about finding where characters are located in a field of pixels and isolating them, and then the OCR will try to identify which characters are the mixture of grey, black and white. This means each data entry needs to be passed separately to the character recognition so that it can accurately recognize the text. Therefore, bounding boxes detection is important and necessary.

This step takes the segmented rows from the previous step as input, detects the bounding boxes of each observation entry, and crops them into individual ordered images accordingly. Separate bounding boxes help the OCR algorithm to detect the target region and transcribe the alphanumeric contents within the region. I added a label to each box to enhance traceability of each observation and this region separation can also eliminate the risk of the recognition algorithm transcribing two closely recorded observations into one.

Bounding box detection can generate lots of boxes for a single object as well as overlap objects (e.g., two numbers). In OCR packages, this is opaque to the user; however, it is necessary to understand this in this workflow. The ultimate goal of bounding box detection is to efficiently encapsulate one and only one object, and you want to be confident that an object exists in the bounding box. Therefore, special care must be taken beyond bounding box detection to ensure this.

Bounding box detection has been a challenging task because historical handwritten records were often contributed by many people, which would result in different writing styles. Research has shown that some writers minimize the space between two observations, while others may leave too much space (Clinchant et al., 2018). This inconsistent writing style makes it extremely challenging to separate each observation entry without human intervention. This challenge has also hindered the transcription of many historical documents.

In the tests with the DRAW ledgers, a bounding box is constructed for each observation to separate the observations in the text line segmentations (Figure 4.6). We wrote python code that called a pretrained model called Efficient and Accurate Scene Text Detector (EAST) (Zhou et al., 2017) to detect and draw the bounding box of each observation entry. This detector utilizes a convolutional neural network to perform word or text line level detection. The model is pretrained on ICDAR 2015 dataset and ICDAR 2013 dataset. ICDAR 2015 consists of 1000 images for training, which is used in Challenge 4 of ICDAR 2015 Robust Reading Competition (Karatzas et al., 2015), and ICDAR 2013 is comprised of 229 images for training, which is used in Challenge 2 of ICDAR 2013 Robust Reading Competition (Karatzas et al., 2013). The two training datasets are annotated by word level rectangular bounding boxes.

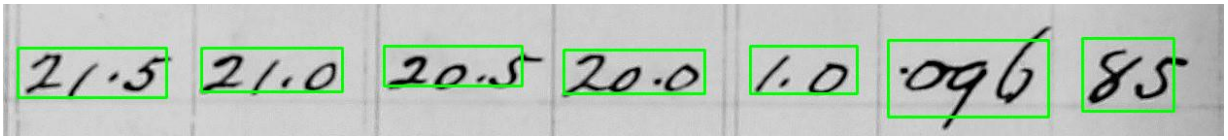
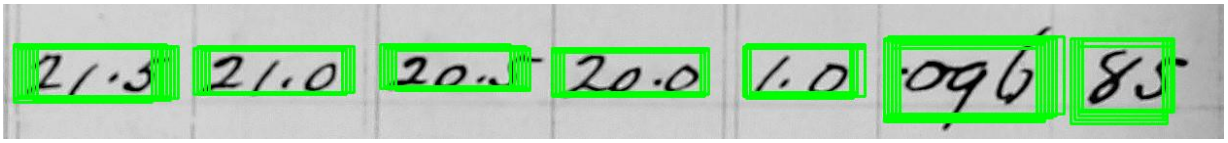
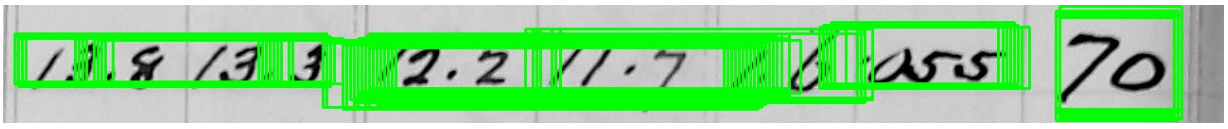


Figure 4.6 - An example of bounding boxes identified on a segmented row.

Figure 4.7 shows that output bounding boxes may not be discrete, as described above. For instance, a bounding box may contain two observations if they were written closely together. I did a visual inspection of a few of my samples to determine a confidence threshold to eliminate overlap. A confidence score here is a value between zero and one, which indicates the likelihood that an object is present in the bounding box (e.g., a confidence of 0.8 would indicate that there is an 80 percent chance that an object exists in that bounding box). This is an iterative process to attempt to minimize the overlap. A default value is also available if visual inspection is not preferred. The EAST documentation suggests that you apply non-maximum suppression, and among the overlapping bounding boxes, only those with maximum confidence are kept valid, while the others are suppressed. These remaining bounding boxes are the final result and are cropped into individual images.



(a) An example row with overlapping bounding boxes for each number.



(b) An example row with overlapping bounding boxes, sometimes one containing two observations.

Figure 4.7 - Bounding boxes detection without non-maximum suppression and thresholded confidence scores. It is important to eliminate overlapping bounding boxes and those that contain more than one observation.

I originally tried generating bounding boxes from entire intact ledger sheets. Experiments show that smaller input images produce better results than larger ones. Due to the large dimensions of

DRAW registers (A2 size), crops are important and necessary to accurately detect the bounding boxes of each entry, although I found that even crops were too large. When a whole page is entered for bounding boxes detection, it is likely that some data entries are not detected. This can be very problematic in transcriptions, as the missing values can cause a lot of chain effects. However, input like the segmented rows do not have similar problems. All data entries can be detected as bounding boxes, so no values will be missed for transcription. This is why text line segmentation is a necessary step before bounding box detection. We grant that this requires considerable human intervention. The user could take a sample and visually compare with the text lines. For example, the user could start with the default settings of the bounding box, as we did, and then shift the threshold.

Similar to text line segmentation, we label the cropped images. The python script for this step includes code to automatically assign each image name, the observation date, row number, and column number. The labelling retains traceability to the original handwritten registers and also ensures transparency in the workflow, where errors made at each step can be easily tracked and corrected. It will also encourage thorough documentation among each step. Innovation and improvements happen every day so I attempted to design a workflow in the hopes that algorithms can be swapped in and the workflow can be refined. That way, modifications can be made more easily to each step without having to interpret each step from scratch. Traceability also makes it easier to manage these images in a database for future usage, such as for citizen science projects and training dataset construction.

In our code, EAST detects the bounds. Our code then crops 3954 bounding boxes from the previous 581 text lines. The Reduce function above allowed us to identify and produce crops for empty cells (boxes with no values) as well as full cells.

4.4.4 Optical character recognition

To reiterate from above, OCR is the conversion of text images into machine-encoded text. It is important to use the cropped images as input because, unlike restaurant receipts, most OCR engines do not have the ability to accurately process a full page of weather or dense text without any layout analysis (Lehenmeier et al., 2020; Shen et al., 2021). An OCR step separated from preprocessing and layout analysis also will aid in tracing the source of errors.

The cropped images from the third step are fed into OCR for transcription, and the output are the transcription results. For the DRAW registers, I chose Tesseract version 5 character recognition. As with earlier versions, Tesseract is open source and works with multiple natural languages; version 5 employs a long short-term memory (LSTM) neural network that uses back propagation to refine the recognition of dynamically changing text (Sak et al., 2014).

The training data and the trained model are as important as the algorithm itself. The quality of the recognition results depends heavily on the suitability of the trained models, and similarly, the quality of the trained model depends almost entirely on the quality of the training dataset. There is no publicly available model trained on handwritten Palmerian cursive used in the DRAW dataset. Consequently, I used a total of three models in this workflow: the baseline Tesseract model (eng.traineddata) and two custom retrained models. The two latter models are retrained on the baseline model with different training datasets. One custom model is retrained on the MNIST dataset (LeCun & Cortes, 1998), which is considered to be one of the most cited dataset for handwritten digits in a normalized form (mnist.traineddata). The other model is retrained on printed float digits in various fonts (printed_digits.traineddata).

As previously, I initially experimented with OCR on uncropped full pages. As anticipated, the recognition output was difficult to interpret (e.g., 2589 52.5896 63256 instead of 25.89 52.5 89 66 32 56); whole rows were not transcribed; individual values were missed; empty cells (i.e., null values) that were essential for correct interpretation were ignored. Recognizing empty cells is a crucial gap in OCR of tabular data. The benchmarking algorithms focus on identifying existing objects and neglects the empty space. In the DRAW dataset, empty cells represent that no observations were taken on that day, and it is important to retain this information for proper interpretation.

This confirms past research that current OCR engines perform poorly without introducing a layout analysis step (e.g., Lehenmeier et al., 2020; Shen et al., 2021). Whole text lines also were tried as input to the OCR. The results improved but remained difficult to distinguish different number readings when two entries were close together. The cropped individual data entries proved the best improvement and were selected as the OCR input.

Regardless of the training dataset, Tesseract usually needs to be configured for the elements to be detected. In this case, it was set to recognize the input as integers or floating point numbers as

well as other factors.[14] (We should note that there are alphabetical characters in the ledgers that we did not test here.)

I performed character recognition for each trained model (see discussion for performance). The transcription result of each data entry is paired with its confidence score, and the pair is labeled with its original position in the registers. This confidence score is returned by the Tesseract itself and can be used as an alternative to evaluate transcription results when ground truth value (i.e., the observation has been transcribed) is not available. I write code to label the results with the confidence score.

4.4.5 Data rearrangement (maintaining the layout)

Rearranging the data is usually achieved by labeling the transcription results. Rearrangement is very important because without it, the output is just arbitrary texts that are not arranged into datasets. This means that the output results are not associated with any row and column information, so the labeling serves the need of associating the results with the original sheets. It also serves traceability to prior steps, for instance, in cases where parameters should be adjusted to improve bounding boxes. This step usually involves putting the results in a list (e.g., a matrix with confidence values) and arranging the results following the original layout of the page.

In the DRAW example, the input images are automatically sorted according to the observation date, row number and column number in the recognition step (e.g., 1884-02-13_1_3.jpg). These labels can be used to track the location of each observation on the original ledger sheets; row numbers are associated with the time of observation (e.g. 7:03 am) and column numbers are associated with type of observation (e.g., vapor pressure). The transcription results and confidence score for each observation are matched and categorized by the corresponding transcription models. At last, the transcription result is collated into data frames along with confidence scores and its original positions and exported as CSV files. There is no extra parameterization for the DRAW records.

In this workflow, the results are rearranged into their original structure based on the layout retained in each step. They are stored in a database in their original structure, so that date ranges and different observations can be easily queried, and files of transcription results can be easily

[14] https://github.com/tesseract-ocr/tesseract/blob/main/doc/tesseract.1.asc

retrieved. This also means that each result can be easily traced back to each step, which is extremely useful in post-processing when we need to figure out which step caused the error. The result can be referred back to its original position referencing the name I gave to each individual observation. The transcription results of different models are also stored separately as different versions of the transcription results. It can serve as a reference for future attempts to improve this workflow.

4.5 Discussion: performance evaluation

This section presents preliminary results of this transcription workflow using the DRAW registers as a case study. It is worth noting that the focus of this workflow is to serve as a guidance for future research, rather than to achieve state-of-the-art accuracy rate. In this regard, the performance of this workflow can be further improved by proper optimization and training. This workflow was tested on 100 randomly selected days in the DRAW dataset register type 150, which contains a total of 3954 single observations. These randomly selected ledger sheets are all in tabular format and the observations are in cursive handwriting.

With the current development of historical record transcription workflows and possible future improvements, it is essential and valuable to establish long-term preservation and storage of different versions of transcription results, metadata, and links between results and original records for traceability. Therefore, an appropriate system must be available to preserve the OCR results, their corresponding individual cropped images, metadata, and their original structure. Different versions of the transcription results from the same record need to be preserved and cross-referenced overtime for research purposes. This documentation of different versions of results can help to progressively improve the performance of transcription workflows and OCR techniques at different stages of development and improvement. Overall, a good preservation and storage benefits both the data holders (e.g., archival researchers) and the data users (e.g., historical data researchers and interested amateurs) by uncovering valuable information buried in historical records.

The performance of this workflow will be evaluated in three perspectives: input-driven performance, output-driven performance, and model-driven performance. The performance to be evaluated here includes aspects such as accuracy, efficiency, proportion of manual work, and complexity. The input-driven performance focuses on evaluating the results from step 1: image

preprocessing; the output-driven performance focuses on the accuracy of output results; and the model-driven performance focuses on steps after step 1.

4.5.1 Input-driven performance

This section will evaluate the performance of step 1: image preprocessing. This step includes image cropping and image enhancement, so these two aspects will be evaluated separately. We will focus on examining the proportion of manual work and its complexity. The first sub-step to be evaluated is image cropping. In this step, it is done fully manual, because this test only processes a small number of images and manual cropping is fast enough. This task is straightforward, less challenging, and should be manageable by someone with basic computer skills. However, if there are a large number of ledgers to process, one may consider automatic image cropping. It would be easier if one is dealing with materials that have similar formats.

As for image enhancement, it is done fully automatically by running a python script. The parameters are pre-set to default values that will accommodate most images. Users are encouraged to fine-tune the parameters according to the characteristics of their images to achieve the best results. Image enhancement, which requires knowledge about python language and the OpenCV library, will be more complex than the image cropping procedure. However, users do not need to fully understand the script in order to run this procedure.

4.5.2 Output-driven performance

This section will evaluate the output performance, and the performance evaluated here is accuracy. The accuracy of the output results is measured by comparing the confidence scores and accuracy of the baseline model and the retrained models. Character error rate (CER) and word error rate (WER), which are the most common evaluation metrics used to examine OCR performance (J. A. Sánchez et al., 2019), are used here to evaluate accuracy. CER indicates the percentage of characters that were incorrectly predicted, including substitutions, deletions, and insertions. WER is defined in the same way but at a word level. A lower value indicates better accuracy.

Table 4.1 shows the performance of the baseline model and the two retrained models. It shows that the confidence score of both the *MNIST* model and *printed_digits* model is significantly better than those of the baseline model. To evaluate the accuracy, I selected a subset of

transcription results to represent the average accuracy of 3912 observations. A sample of 350 was randomly selected and was annotated with ground truth values under the supervision of domain experts. This sample size reflected a 95 percent confidence level and a five percent margin of error. Both retrained models have better CER and WER values than the baseline model; the *MNIST* model has a slightly better CER than the *printed_digits* model, and their WER values are comparable. This indicates that with proper training data and models, the accuracy of this workflow can be improved substantially. Since Tesseract has the advantage of being open source, it can easily be trained and modified, compared to other OCR software (e.g., ABBYY FineReader).

Table 4.1 - The performance of the three OCR models.

	Models		
	MNIST	printed_digits	baseline
Confidence score (%)	74.6	59.2	18.1
CER (%)	47.3	51.7	63.0
WER (%)	84.3	83.6	88.1

This finding is consistent with the view described in the literature review that automating the transcription of historical handwritten records is a challenging task and is far from error-free (Alabau & Leiva, 2012; Firmani et al., 2017; Holley, 2009; Jander, 2016; Jenckel et al., 2016; J. A. Sánchez et al., 2014; Swindall et al., 2021). However, I found that accuracy was improved by modifying the training data and thus the trained model, which confirms Odunayo et al. (2021) and Rakshit et al. (2010) that this is an important next step. This means, refining and customizing the training data can largely improve the accuracy. Another finding is that the average confidence score can reflect the accuracy of the results to some extent. It can be used to compare the accuracy between different models; a higher average confidence score is likely to imply an average more accurate result. However, it is not reliable to be used as a reference for each individual transcription. This explains why recent studies have never used the confidence scores as the sole evaluation of the results.

4.5.3 Model-driven performance

This section will evaluate the model-driven performance, which includes every step after image preprocessing. The performance will be evaluated in four aspects: accuracy, efficiency (runtime), proportion of manual work, and complexity.

As for accuracy, this section will focus on evaluating layout analysis and rearrangement, i.e., step 2: text line segmentation, step 3: bounding boxes detection, and step 5: data rearrangement. From the results, all text lines were segmented (step 2) and all cells were cropped (step 3) according to the bounding boxes after parameter tuning. Notably, this includes every empty cell, which fills the gap of identifying empty spaces in the layout analysis. By identifying all cells of poorly-defined tables in the DRAW dataset, this result also alleviates the difficulty of analyzing the layout of unbounded or faded tables as raised in past studies (Prasad et al., 2020; Ziomek & Middleton, 2021). In data rearrangement (step 5), every entry was labeled correctly, whereby each result can be traced back to its position in the original ledger. This also includes empty cells. In conclusion, the results of these three steps were very accurate. It not only fills the gap of identifying and documenting empty cells, but also has the potential to solve the problem of unbounded tables that has been raised in past studies.

The time efficiency is assessed by comparing the runtime of the steps to the speed of average human transcribers. Considering that OCR (step 4) takes up most of the time when running the workflow and that it is the only step that can be quantified and compared to the runtime with human transcribers, we will focus on evaluating the efficiency of this step. To evaluate the time efficiency in the OCR step, the run time of different OCR models were compared to the speed of average human transcribers. The running time of the three models ranged from 13m44s to 21m55s under the same hardware and software settings, as shown in Table 4.2. For human transcribers, the DRAW project estimates their average transcription speed to be 300 observations per hour. This means that for 3954 observations, an average human transcriber will need to work nonstop for 13.04 hours to complete the transcription. However, the same number of observations can be automatically transcribed in less than 25 minutes if you only look at the runtime in the OCR stage. This result suggests that this workflow is much faster than manual transcription in the transcription part alone, which can significantly improve the speed of transcription. This confirms Brönnimann et al. (2006), Wilkinson et al. (2019) and World

Meteorological Organization (2016) that automation can greatly accelerate the transcription part of historical data rescue, and thus, with appropriate development and refinement, has the potential to speed up the entire historical data rescue process. Unfortunately, this does not mean it will be accurate; just significantly faster.

Table 4.2- Runtime of three OCR models.

	Models		
	MNIST	printed_digits	baseline
Runtime *compared with 13.04hr if done by average human transcriber*	21m55s	13m44s	16m46s

Then we will assess the proportion of manual work needed in these steps. The manual work required in the text line segmentation, bounding boxes detection and OCR steps is limited to parameter tuning, which is similar to benchmark algorithms and software to date. As for the last step, data rearrangement, no manual work is needed; everything is embedded in the script and no parameters need to be adjusted. The amount of manual work is similar to other benchmarks, except that this workflow has the benefit of combining procedures such as layout analysis, OCR and data arrangement as one product.

At last, we will evaluate the complexity of these steps. Each step requires different knowledge. For the text line segmentation, bounding boxes detection and OCR steps, knowledge such as building training datasets, training models and a moderate understanding of the script is required. That is, when dealing with dissimilar materials, it may be necessary to build training datasets of other fonts, and it may be necessary to train new models to handle the material better. These have long been difficult and unsolved challenges raised by past research, albeit very beneficial to the model performance (e.g., Dahl et al., 2021; Fornés et al., 2017; Shen et al., 2021). A moderate understanding of the code is also necessary in situations such as adjusting parameters or debugging programs. For the data rearrangement step, knowledge and skills of database management system (DBMS) are required to manage, arrange and label the transcription result into an easily accessible and retrievable database. In conclusion, it is difficult to use this workflow if the user does not have a moderate understanding of the background technology. This can be frustrating to many researchers who do not possess the technical skills but could benefit

greatly from this workflow. The next section will discuss future steps that can mitigate potential challenges like this one.

4.6 Conclusions and next steps

In the last few decades, we have seen tremendous progress in the field of research of OCR transcription in the humanities, which has made it possible to learn more from the historical records than ever before. Manually making tens of millions of historical records machine-readable is not always an easy task, thus automation is gradually being considered as an alternative. This preliminary workflow is a cornerstone for future attempts to automate the transcription of historical records, and there is still much to consider as to what the path forward will be.

The key innovation achieved by inventing this workflow is not only to provide a robust system that takes scanned historical records as input and outputs a labeled and arranged digital dataset, but also it provides direction to solve existing conundrums such as automatically analysing the layout of poorly defined tables, identifying empty cells, and providing an end-to-end automated transcription solution. While this workflow sounds very promising, there are also many difficulties and challenges. Here in Table 4.3, we present the summary of challenges that readers may face in doing OCR of their historical records. I also offer possible solutions.

Table 4.3 - Summary of possible challenges and their recommendations.

Theme	Description of Challenge	Approach/Recommended actions
Step 1: Image Processing		
Cropping	Cropping is done through manual intervention. This may affect the robustness and speed of this workflow when dealing with different document structures.	A robust layout segmentation tool may help alleviate this issue. Automation may speed up this step.
Step 2: Text Line Segmentation		
Horizontal segmentation	This method is not applicable to documents that have many author comments between	Special care is needed in the layout analysis. For example, a

	lines (as this problem does not exist in the sample dataset). It will remove these comments, which can be a problem if they are valuable information.	special text block recognition where these extra comments can be extracted before the line segmentation.
Step 3: Bounding Boxes Detection		
Text documents	For text documents (e.g., diaries, newspapers), bounding box detection may not be optimal. Text recognition may be more accurate if the recognition is done on a sentence-by-sentence basis.	Natural language processing may be a useful alternate step to enhance semantic understanding.
Step 4: Optical Character Recognition		
Accuracy	The accuracy of this workflow is lower than the accuracy of manual transcription. Therefore, it may not be appropriate at this time to replace manual transcription. Greater effort needs to be invested to improve accuracy in the future.	Training data that contains historical cursive font may be necessary to increase the accuracy, although this is expensive and time-consuming
Evaluation - post processing	Confidence scores may not be a reliable source of evaluation. As shown in the results, confidence scores are not reliable in reflecting the accuracy of individual transcription; they only reflect average accuracy. Another approach, sample test on ground truth values, is laborious for large datasets or in the long term.	Sufficient funding to ground truthing can partially address this challenge. Ground truth will always be difficult for historical observations so new post-processing methods and sampling techniques might need to be created. For example, one might create a captcha for validation.
Step 5: Rearrangement/Layout Analysis		
Changes to Layout	Workflow can be hyper parameterized to fit a certain page layout and suddenly the page layout changes.	Some pages may always have to be handled manually, or you may have to visually inspect the corpus beforehand.

Embedded traceability	Currently, traceability is achieved through nomenclature embedded in each step. Traceability may be compromised if substantial changes are made to each step, resulting in the need to change naming conventions.	Detailed logs may be generated and kept for each step so that everything can be traced back.
Workflow		
Human intervention	For an automated transcription workflow, a fair amount of human intervention is still required to get the workflow operating. Human interventions include, but are not limited to, visual inspection, parameter tuning, and model selection.	One possible solution to this challenge is to use newer technologies where we can minimize human intervention and automate every step.

The following sections will discuss the outlook of AI-enhance historical record transcription with a sustainable future. The discussion focuses on three aspects: technology advancement, community effort, and transparency.

4.6.1 Technology advancement

The accuracy of the transcription results suggests that the transcription step is far from optimal, but it provides guidance on how to refine and improve in the future. We can see that the accuracy increases when an appropriate model is used for recognition. Appropriate here means that the writing styles of the training dataset used to train the model has sufficient similarity to the input writing styles. In this case, models trained with printed fonts or modern handwriting fonts are not sufficient to produce good accuracy for historical cursive handwritings. I believe that a model trained with historical cursive fonts will greatly improve the transcription accuracy. As a result, building a training dataset of historical cursive handwriting with ground truth values can be the next step for improvement. Transfer learning can also be an option to minimize the effort of building a training dataset.

Since AI is a quickly developing field, new transcription techniques and algorithms will gradually emerge. Tesseract will not be the only open-source and robust option for character recognition. A more powerful and flexible transcription engine could greatly improve accuracy in the future. Since this workflow is very robust and flexible, new transcription engines can be

easily adapted to the workflow, and it can easily be modified to connect to the preceding and following steps.

My workflow may very well be a stopgap until a better system is developed. Transkribus, as an example, has revolutionized OCR for the humanities. It will continue to improve, in training data sets, methods, and UI/UX. But we have to balance cost, proprietary, closed and inflexible systems with better support and better UI/UX.

4.6.2 Community effort

OCR in the humanities and in historical data rescue is very much a community effort. Automating the rescue of historical records requires community efforts. This is a multidisciplinary project involving researchers interested in historical records, archival researchers, and experts in computer science (e.g., computer vision). As stated in the literature review, automation requires a combination of technologies, and likewise, it requires a combination of efforts from different communities. These communities include archival research community, historical record research community and computer science community. As discussed in Zhang and Sieber (2022), many historical data rescue projects do not have the resources to follow through with building automation from scratch or to move forward after the project is implemented. The resource here can be money, but most importantly it can be experts with computer science knowledge who can maintain the transcription workflow overtime. A lot of projects are hesitant to implement automation because of the lack of experts in the computer science community. We hope that multiple communities will collaborate and support each other to improve and refine the automation of data rescue in the future.

While it sounds promising to have a cursive training dataset, which could greatly improve the accuracy of the workflow, it is difficult to build a training dataset. As mentioned in literature review, building a suitable training dataset has been an issue in past research (Dahl et al., 2021; Fischer et al., 2014; J. A. Sánchez et al., 2019), and this has not changed so far. Building a training dataset from scratch is still expensive, labor-intensive, and time-consuming. In addition, there will always be character sets left out, for example, Persian, Arabic, or handwritten indigenous languages. If less care is given to these communities, valuable information stored in these languages may not have the opportunity to be analyzed. These problems may be mitigated if people from multiple fields work together.

Innovations and improvement also require the collective support from these multiple communities. In the computer science community, fields like computer vision and image processing have created a lot of state-of-the-art algorithms, with a focus on deep learning technologies (e.g., Chamchong et al., 2019; Paliwal et al., 2019; Prasad et al., 2020). However, as Shen et al. (2021) point out, researchers who benefit from these innovations often lack the expertise to put them to use. Let alone building a one customized workflow from scratch. The lack of collaborative efforts hinders the strength of innovation and holds back the improvement of similar matters. We hope to improve and refine the automation of data rescue by having experts from these communities contribute their ideas and collaborate.

4.6.3 Transparency

Acceptance of new technologies is always an issue to consider before introducing innovative improvements to a field. Zhang and Sieber (2022) found that the acceptance is partially hindered by the lack of transparency of new technologies. By transparency, I mean the openness of the source code. In the case of automation, it also includes open access to detailed documentation of any automation attempts and past experiments. Neudecker et al. (2019) also indicated that transparency is necessary for a sustainable future of automated record transcription. However, transparency can be good and bad for automation.

Transparency can be an important element of sustainable automation. These source code and documentations are valuable references to future research and can help avoid unnecessary detours. Tafti (2016) pointed out that commercial automation software (e.g., ABBYY FineReader, Transkribus) do not share the underlying technology. Without transparency, automation is more like a black box that is difficult for other researchers to interpret and replicate, and therefore poorly disseminated. This can hinder the advancement of automation because documentations and experiments are not shared, which is an essential factor for sustainable improvement.

However, transparency can also hinder the development of automation and may not be necessary. Many existing commercial OCR software has a comprehensive user interface that allows for easy customization and makes installation straightforward. This way, users do not feel the need to know the “black box” behind the scenes, as long as the software functions as it is advertised. Similar suggestions have also been brought up by Jenckel et al. (2016) and Memon et

al. (2020). In this case, transparency can have the opposite effect, such as being overwhelmed by the complexity of the code and hesitant to automate. It may also cause confusion and conflict when there are too many people working on the same code (e.g., consistency may be compromised). This may hamper the transition to automation and thus slow down the pace of development.

As a result, transparency in automation is something that we should handle carefully as we proceed. The amount of transparency may determine how much progress we can make in automation and how willing researchers are to make the transition to it. The impact of transparency is contingent on many factors and there is no absolute answer to what should be done in the future. In order to achieve a sustainable future for automation, how to balance this impact should be carefully considered.

References

Alabau, V., & Leiva, L. (2012). Transcribing handwritten text images with a word soup game. *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, 2273–2278. https://doi.org/10.1145/2212776.2223788

Tafti, A. P., Baghaie, A., Assefi, M., Arabnia, H. R., Yu, Z., & Peissig, P. (2016). OCR as a Service: An Experimental Evaluation of Google Docs OCR, Tesseract, ABBYY FineReader, and Transym. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, F. Porikli, S. Skaff, A. Entezari, J. Min, D. Iwai, A. Sadagic, C. Scheidegger, & T. Isenberg (Eds.), *Advances in Visual Computing* (pp. 735–746). Springer International Publishing. https://doi.org/10.1007/978-3-319-50835-1_66

Blancq, F. L. (2010). Rescuing old meteorological data. *Weather*, *65*(10), 277–280. https://doi.org/10.1002/wea.510

Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R., Brugnara, Y., Brunet, M., Brunetti, M., Chimani, B., Cornes, R., Domínguez-Castro, F., Filipiak, J., Founda, D., Herrera, R. G., Gergis, J., Grab, S., Hannak, L., Huhtamaa, H., … Wyszyński, P. (2019). Unlocking Pre-1850 Instrumental Meteorological Records: A Global Inventory. *Bulletin of the American Meteorological Society*, *100*(12), ES389–ES413. https://doi.org/10.1175/BAMS-D-19-0040.1

Brönnimann, S., Annis, J., Dann, W., Ewen, T., Grant, A. N., Griesser, T., Krähenmann, S.,

Mohr, C., Scherer, M., & Vogler, C. (2006). A guide for digitising manuscript climate data. *Climate of the Past*, *2*(2), 137–144. https://doi.org/10.5194/cp-2-137-2006

Chamchong, R., Gao, W., & McDonnell, M. D. (2019). Thai Handwritten Recognition on Text Block-Based from Thai Archive Manuscripts. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1346–1351. https://doi.org/10.1109/ICDAR.2019.00217

Chen, J., Riba, P., Fornés, A., Mas, J., Lladós, J., & Pujadas-Mora, J. M. (2018). Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 528–533. https://doi.org/10.1109/ICFHR-2018.2018.00098

Chimani, B., Auer, I., Prohom, M., Nadbath, M., Paul, A., & Rasol, D. (2021). Data rescue in selected countries in connection with the EUMETNET DARE activity. *Geoscience Data Journal*, *9*(1), 187–200. https://doi.org/10.1002/gdj3.128

Clinchant, S., Dejean, H., Meunier, J.-L., Lang, E. M., & Kleber, F. (2018). Comparing Machine Learning Approaches for Table Recognition in Historical Register Books. *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 133–138. https://doi.org/10.1109/DAS.2018.44

Dahl, C. M., Johansen, T. S. D., Sørensen, E. N., Westermann, C. E., & Wittrock, S. F. (2021). Applications of Machine Learning in Document Digitisation. *ArXiv:2102.03239 [Cs, Econ, Stat]*. http://arxiv.org/abs/2102.03239

Firmani, D., Merialdo, P., Nieddu, E., & Scardapane, S. (2017). In Codice Ratio: OCR of Handwritten Latin Documents using Deep Convolutional Networks. *Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage (AI*CH 2017)*, 9–16.

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., & Ingold, R. (2014). *The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries* (pp. 91–106). https://doi.org/10.13140/2.1.2180.3526

Fornés, A., Megyesi, B., & Romeu, J. M. (2017). Transcription of Encoded Manuscripts with Image Processing Techniques. *DH*.

Gatos, B., Louloudis, G., Causer, T., Grint, K., Romero, V., Sánchez, J. A., Toselli, A. H., & Vidal, E. (2014). Ground-Truth Production in the Transcriptorium Project. *2014 11th IAPR International Workshop on Document Analysis Systems*, 237–241.

https://doi.org/10.1109/DAS.2014.23

Harnett, J. (1985, November 14). Developments in OCR for automatic data entry. *Proceedings of Translating and the Computer 7*. TC 1985, London, UK. https://aclanthology.org/1985.tc-1.12

Holley, R. (2009). How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs. *D-Lib Magazine*, *15*(3/4). https://doi.org/10.1045/march2009-holley

Jander, M. (2016). Handwritten Text Recognition—Transkribus: A User Report. *ETRAP Research Group, Institute of Computer Science, University of G¨ Ottingen, Germany*, 3.

Jenckel, M., Bukhari, S. S., & Dengel, A. (2016). anyOCR: A sequence learning based OCR system for unlabeled historical documents. *2016 23rd International Conference on Pattern Recognition (ICPR)*, 4035–4040. https://doi.org/10.1109/ICPR.2016.7900265

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., Shafait, F., Uchida, S., & Valveny, E. (2015). ICDAR 2015 competition on Robust Reading. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 1156–1160. https://doi.org/10.1109/ICDAR.2015.7333942

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., Bigorda, L. G. i, Mestre, S. R., Mas, J., Mota, D. F., Almazàn, J. A., & de las Heras, L. P. (2013). ICDAR 2013 Robust Reading Competition. *2013 12th International Conference on Document Analysis and Recognition*, 1484–1493. https://doi.org/10.1109/ICDAR.2013.221

Klijn, E. (2008). The Current State-of-art in Newspaper Digitization: A Market Perspective. *D-Lib Magazine*, *14*(1/2). https://doi.org/10.1045/january2008-klijn

Kwok, R. (2017). Historical data: Hidden in the past. *Nature*, *549*(7672), 419–421. https://doi.org/10.1038/nj7672-419

LeCun, Y., & Cortes, C. (1998). The MNIST database of handwritten digits. *Http://Yann.Lecun.Com/Exdb/Mnist/*.

Lehenmeier, C., Burghardt, M., & Mischka, B. (2020). Layout Detection and Table Recognition – Recent Challenges in Digitizing Historical Documents and Handwritten Tabular Data. In M. Hall, T. Merčun, T. Risse, & F. Duchateau (Eds.), *Digital Libraries for Open Knowledge* (pp. 229–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-54956-5_17

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, *8*, 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., & Herrmann, E. (2019). OCR-D: An end-to-end open source OCR framework for historical printed documents. *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, 53–58. https://doi.org/10.1145/3322905.3322917

Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A Survey of Historical Document Image Datasets. *ArXiv:2203.08504 [Cs]*. http://arxiv.org/abs/2203.08504

Odunayo, O., Sookoo, N. N., Bathla, G., Cavallin, A., Persaud, B. D., Szigeti, K., Van Cappellen, P., & Lin, J. (2021). Rescuing historical climate observations to support hydrological research: A case study of solar radiation data. *Proceedings of the 21st ACM Symposium on Document Engineering*, 1–4. https://doi.org/10.1145/3469096.3474929

Paliwal, S. S., D, V., Rahul, R., Sharma, M., & Vig, L. (2019). TableNet: Deep Learning Model for End-to-end Table Detection and Tabular Data Extraction from Scanned Document Images. *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 128–133. https://doi.org/10.1109/ICDAR.2019.00029

Parthiban, R., Ezhilarasi, R., & Saravanan, D. (2020). Optical Character Recognition for English Handwritten Text Using Recurrent Neural Network. *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–5. https://doi.org/10.1109/ICSCAN49426.2020.9262379

Prasad, D., Gadpal, A., Kapadni, K., Visave, M., & Sultanpure, K. (2020). CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2439–2447. https://doi.org/10.1109/CVPRW50498.2020.00294

Rakshit, S., Kundu, A., Maity, M., Mandal, S., Sarkar, S., & Basu, S. (2010). Recognition of handwritten Roman Numerals using Tesseract open source OCR engine. *ArXiv:1003.5898 [Cs]*. http://arxiv.org/abs/1003.5898

Sak, H., Senior, A., & Beaufays, F. (2014). *Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition* (arXiv:1402.1128). arXiv. https://doi.org/10.48550/arXiv.1402.1128

Sánchez, J. A., Bosch, V., Romero, V., Depuydt, K., & de Does, J. (2014). Handwritten text

recognition for historical documents in the transcriptorium project. *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 111–117. https://doi.org/10.1145/2595188.2595193

Sánchez, J. A., Romero, V., Toselli, A. H., Villegas, M., & Vidal, E. (2019). A set of benchmarks for Handwritten Text Recognition on historical documents. *Pattern Recognition*, *94*, 122–134. https://doi.org/10.1016/j.patcog.2019.05.025

Sánchez, J.-A., Mühlberger, G., Gatos, B., Schofield, P., Depuydt, K., Davis, R., Vidal, E., & de Does, J. (2013). *tranScriptorium: A european project on handwritten text recognition*. 227–228. https://doi.org/10.1145/2494266.2494294

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *ArXiv:2103.15348 [Cs]*. http://arxiv.org/abs/2103.15348

Singh, A., Bacchuwar, K., & Bhasin, A. (2012). A Survey of OCR Applications. *International Journal of Machine Learning and Computing*, 314–318. https://doi.org/10.7763/IJMLC.2012.V2.137

Stickler, A., Brönnimann, S., Valente, M. A., Bethke, J., Sterin, A., Jourdain, S., Roucaute, E., Vasquez, M. V., Reyes, D. A., Allan, R., & Dee, D. (2014). ERA-CLIM: Historical Surface and Upper-Air Data for Future Reanalyses. *Bulletin of the American Meteorological Society*, *95*(9), 1419–1430. https://doi.org/10.1175/BAMS-D-13-00147.1

Stickler, Alexander, Brönnimann, Stefan, Jourdain, Sylvie, Roucaute, Eméline, Sterin, Alexander M, Nikolaev, Dmitrii, Valente, Maria Antónia, Wartenburger, Richard, Hersbach, Hans, Ramella Pralungo, Lorenzo, & Dee, Dick P. (2014). *ERA-CLIM Historical Upper-Air Data 1900-1972, supplement to: Stickler, Alexander; Brönnimann, Stefan; Jourdain, Sylvie; Roucaute, Eméline; Sterin, Alexander M; Nikolaev, Dmitrii; Valente, Maria Antónia; Wartenburger, Richard; Hersbach, Hans; Ramella Pralungo, Lorenzo; Dee, Dick P (2014): Description of the ERA-CLIM historical upper-air data. Earth System Science Data, 6(1), 29-48* [Application/zip]. 813 datasets. https://doi.org/10.1594/PANGAEA.821222

Ströbel, P. B., Clematide, S., Volk, M., & Hodel, T. (2022). Transformer-based HTR for Historical Documents. *ArXiv:2203.11008 [Cs]*. http://arxiv.org/abs/2203.11008

Swindall, M. I., Croisdale, G., Hunter, C. C., Keener, B., Williams, A. C., Brusuelas, J. H., Krevans, N., Sellew, M., Fortson, L., & Wallin, J. F. (2021). Exploring Learning

Approaches for Ancient Greek Character Recognition with Citizen Science Data. *2021 IEEE 17th International Conference on EScience (EScience)*, 128–137. https://doi.org/10.1109/eScience51609.2021.00023

Terras, M. (2022). Inviting AI into the archives: The reception of handwritten recognition technology into historical manuscript transcription. *Archives, Access and AI: Working with Born-Digital and Digitised Archival Collections*, 179–204. https://doi.org/10.1515/9783839455845-008

Wilkinson, C., Brönnimann, S., Jourdain, S., Roucaute, E., Crouthamel, R., Brohan, P., Valente, A., Brugnara, Y., Brunet, M., & Team, I. (2019). *Best Practice Guidelines for Climate Data Rescue*.

World Meteorological Organization. (2016). *Guidelines on Best Practices for Climate Data Rescue*. https://library.wmo.int/doc_num.php?explnum_id=3318

Yasser, A. M., Clawson, K., & Bowerman, C. (2017). *Saving Cultural Heritage with Digital Make-Believe: Machine Learning and Digital Techniques to the Rescue*. https://doi.org/10.14236/ewic/HCI2017.97

Zhang, Y., & Sieber, R. (2022). *A survey on attitude and perception of AI-augmented data rescue among leaders of data rescue community* [Manuscript in preparation]. McGill University.

Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). *EAST: An Efficient and Accurate Scene Text Detector*. https://arxiv.org/abs/1704.03155v2

Ziomek, J., & Middleton, S. E. (2021). GloSAT Historical Measurement Table Dataset: Enhanced Table Structure Recognition Annotation for Downstream Historical Data Rescue. *The 6th International Workshop on Historical Document Imaging and Processing*, 49–54. https://doi.org/10.1145/3476887.3476890

Chapter 5. Discussion, conclusions, and future directions

5.1 Thesis discussion and summary

The aim of this research is to provide theoretical and empirical guidance and reference for future attempts of artificial intelligence (AI) in historical weather data rescue and historical record transcription studies. To address the aim, I asked two research questions: How do researchers and practitioners perceive the challenges and opportunities of using AI-augmented data rescue? If AI-augmented data rescue is useful then what might an automated system look like? The AI-augmented data rescue here refers to using AI technology to automate the transcription process and improve on the manual data rescue process. It is hoped that the result of this study will contribute to filling the gap in the application of AI to this multidisciplinary issue. These questions are answered in Chapter 2, 3, and 4 through a literature review, survey study, and the establishment of an automated transcription workflow.

Chapter 2 conducted a literature review of the evolution of past and present data rescue research and what role AI has played in related studies. The literature concludes that although AI techniques sound promising in many ways (e.g., time, money), current historical weather data rescue projects prefer to use the manual keying methods for transcription (Brönnimann et al., 2006; World Meteorological Organization, 2016). Many studies have found that manual keying methods, including hiring paid transcriptionists and recruiting citizen science volunteer transcriptionists, are labor-intensive and time-consuming (Brohan, 2017; Craig & Hawkins, 2020; Gura, 2013), so more and more studies are considering the possibility that AI may contribute to the data rescue and transcription process (e.g., Chen et al., 2018; Chimani et al., 2021; World Meteorological Organization, 2016). However, studies have shown that there is a disconnect between the computer science (e.g., AI) experts and historical records researchers. The former are domain-agnostic, and they focus on finding general solutions rather than customizing an automated solution for a specific domain, such as rescuing historical weather data. The latter, historical records researchers, usually do not have the expert knowledge needed to utilize AI, even though they may benefit most from it (Shen et al., 2021). There is also no precedent for studying the perceptions of data rescue researchers on the use of AI in either community, and the opportunities and challenges data rescuers face.

Given that there are challenges in using AI, it is important to hear from the data rescue and citizen science community what they think about implementing AI in their fields. Chapter 3 conducted a survey study among 50 principal investigators or leading scientists of data rescue or citizen science transcription projects. Here, respondents were primarily interested in transcribing historical weather records, but I also sought responses from citizen science projects who are interested in transcription projects in general. Unfortunately, there was no significant diversity of perspectives across communities, backgrounds, years of experience, technical skill levels, etc. The survey results revealed that the respondents are aware of AI but have barely used it in their research to transcribe the records. Nevertheless, most are willing to try AI in their research, but there are certain obstacles, such as the accuracy of the results, availability of funding, and the time needed to invest. Respondents believe that AI will be an opportunity in the field of historical weather data rescue and historical records transcription more generally, but equally, they express concerns about the potential elimination of public participation and loss of valuable information that cannot be automatically transcribed at this stage. There is clearly a gap between what AI-augmented automation approaches promise and what can be achieved today. Respondents hope that a hybrid approach will emerge in which human and AI work together will fill this gap.

If opportunities are envisaged for AI in historical weather data rescue, what would it look like? How would we implement it? Chapter 4 established an AI-augmented end-to-end historical record transcription workflow and tested it with a specific historical weather data rescue project. The Data Rescue: Archival and Weather (DRAW) project is an online citizen science project that rescues weather records from 1874 to 1964. The DRAW weather records are dense tables filled with cursive handwritten weather observations (e.g., barometer readings, vapor pressures, wind direction). These dense ledgers have a complex layout and were a standard and typical format throughout the world in the late nineteenth and early twentieth centuries. Therefore, testing on these records may open the opportunity to rescue many other historical weather records of the time. I created a workflow that was divided into five discrete steps to allow for better adaptation to future advances: (1) image preprocessing; (2) text line segmentation; (3) bounding boxes detection; (4) optical character recognition (OCR); (5) data rearrangement. There was considerable trial and error in determining the appropriate number and type of steps. Testing was performed on 100 randomly selected daily weather observations, which included air temperature, wet bulb, vapor pressure, relative humidity, and other meteorological variables (e.g., some pages

contained central tendencies like averages). These 100 pages of observations were preprocessed, segmented into 581 rows, cropped into a total of 3,954 individual observations, processed and transcribed by an OCR augmented by a long short-term memory (LSTM) neural network, labeled with confidence scores, dates, and original layout locations, and arranged into an easily searchable and accessible format. I argue that this workflow is robust because it is adaptable to possible future advances; for example, each step can be swapped and adapted to other technologies. Additionally, the workflow combines multiple technologies from different fields that work together to achieve historical records transcription. Specifically, the detection of numbers in bounding boxes outperforms existing popular and commonly used algorithms. Although the resulting transcription accuracy is not ideal, it can be improved with appropriate parameter tuning and adequate training.

This research provides a solution for the multidisciplinary problem about using AI to automate the transcription of historical records. The results in Chapter 3 indicate that researchers acknowledge the opportunities in AI-augmented data rescue and are eager to experiment with this new technology, although they are limited in terms of resources and skill levels. They also are concerned about the impact on citizen science participation. The workflow creates a path to implement and test AI-augmented data rescue. Unfortunately, challenges remain, such as the lack of funding and support staff to maintain the system over time. There are findings across chapters. Specifically, Chapter 3 informs Chapter 4, and they confirm each other. Chapter 3 has anticipated that automating the data rescue process would be difficult and would require a joint and continuous effort, and Chapter 4 confirms this. There was a lot of back and forth testing and a lot of trials and errors in creating the workflow. The obstacles identified in Chapter 4 when creating the workflow were also concerns for the respondents in Chapter 3. For example, the respondents suggested that the transcription accuracy may not be optimal, and one possible way to improve it is to find a training dataset, which can be difficult. The results of this thesis answer both research questions by providing a detailed understanding of researchers' perspectives on the use of AI in their transcription research and by presenting an AI-augmented transcription workflow with detailed documentation for future refinements.

This research has several limitations. First, due to the resource and time constraints, the survey study in Chapter 3 consulted with 50 principal investigators and leading scientists in the related fields. The perspectives on the use of AI in historical records transcription may be more diverse

if a larger number of people were surveyed. For example, if more people in the related fields were recruited, there may be more correlations between the respondents' field of interests, work experience, etc. Second, although the workflow is generalizable, it has not been tested in historical records with letters, in other languages, or in other domains (e.g., humanity transcription); it has only been tested with numeric Palmerian-handwritten weather records. Resource, time and financial constraints made testing difficult to do thoroughly. This workflow would be able to be tested in a wider range of cases if the condition and situation allows. In other words, more robustness and refinement can be achieved if time, resource and funding constraints allow.

5.2 What are the future directions?

The present study has several aspects that may be improved in future studies. First, there is still room for improvement in the transcriptional accuracy of the workflow. The suboptimal accuracy may be due to the lack of customized training datasets. It is difficult to find suitable existing training datasets and to collect and build training datasets from scratch (e.g., Dahl et al., 2021; Fischer et al., 2014; Nikolaidou et al., 2022; Shen et al., 2021). The recognition results may be improved by advances in technology, in which case custom training datasets may be easier to build or not needed. It would be useful if some proprietary company or government invested in building a handwriting training model, since there are so many diaries, records, etc. that are in danger of being lost. There is also an argument for having collective funding to develop and maintain such transcription systems. It was already difficult to create and test a workflow from scratch by myself. I cannot imagine how difficult it will be if researchers decide to implement my workflow. Services such as building an intuitive user interface, as mentioned below, and troubleshooting may be new issues to consider. The economics of creating and maintaining such workflow are new, and hopefully future studies will find a way to balance all stakeholders.

Second, this AI-augmented data rescue workflow still requires a small amount of human intervention and expert knowledge to run and oversee this workflow. Several solutions have been proposed in this study to address this issue. Possible future directions include implementing an intuitive user interface (UI) and user experience (UX) and making the workflow hybrid so that the algorithm works under human supervision. The latter solution could also address some of the

concerns of researchers, such as retrieving valuable comments that cannot be transcribed by computers with current technologies.

5.3 Concluding remarks

These concerns and challenges, while not pressing, are still issues that need to be addressed in future studies. This means that a great deal of effort still needs to be invested in the research on the use of AI in data rescue and historical record transcription projects. In other words, this is an ongoing process that needs continuous efforts. This thesis provides a comprehensive investigation of the two research questions as it presented the views of researchers and practitioners on the challenges and opportunities of using AI-augmented data rescue, and also proposed and tested an automated transcription workflow. What AI-augmented data rescue promises to achieve is very helpful for historical records transcription, as long as the concerns and challenges identified are addressed. Hopefully, the future studies will build on this research and by addressing these issues in the next step, AI-augmented data rescue and historical records transcription can become a proven solution to the tens of millions of historical records that have not been digitized.

References

Brohan, P. (2017, August 17). *RealClimate: Data rescue projects*. https://www.realclimate.org/index.php/archives/2017/08/data-rescue-projects/

Brönnimann, S., Annis, J., Dann, W., Ewen, T., Grant, A. N., Griesser, T., Krähenmann, S., Mohr, C., Scherer, M., & Vogler, C. (2006). A guide for digitising manuscript climate data. *Climate of the Past*, *2*(2), 137–144. https://doi.org/10.5194/cp-2-137-2006

Chen, J., Riba, P., Fornés, A., Mas, J., Lladós, J., & Pujadas-Mora, J. M. (2018). Word-Hunter: A Gamesourcing Experience to Validate the Transcription of Historical Manuscripts. *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 528–533. https://doi.org/10.1109/ICFHR-2018.2018.00098

Chimani, B., Auer, I., Prohom, M., Nadbath, M., Paul, A., & Rasol, D. (2021). Data rescue in selected countries in connection with the EUMETNET DARE activity. *Geoscience Data Journal*, *9*(1), 187–200. https://doi.org/10.1002/gdj3.128

Craig, P. M., & Hawkins, E. (2020). Digitizing observations from the Met Office Daily Weather

Reports for 1900–1910 using citizen scientist volunteers. *Geoscience Data Journal*, *7*(2), 116–134. https://doi.org/10.1002/gdj3.93

Dahl, C. M., Johansen, T. S. D., Sørensen, E. N., Westermann, C. E., & Wittrock, S. F. (2021). Applications of Machine Learning in Document Digitisation. *ArXiv:2102.03239 [Cs, Econ, Stat]*. http://arxiv.org/abs/2102.03239

Fischer, A., Bunke, H., Naji, N., Savoy, J., Baechler, M., & Ingold, R. (2014). *The HisDoc Project. Automatic Analysis, Recognition, and Retrieval of Handwritten Historical Documents for Digital Libraries* (pp. 91–106). https://doi.org/10.13140/2.1.2180.3526

Gura, T. (2013). Citizen science: Amateur experts. *Nature*, *496*(7444), Article 7444. https://doi.org/10.1038/nj7444-259a

Nikolaidou, K., Seuret, M., Mokayed, H., & Liwicki, M. (2022). A Survey of Historical Document Image Datasets. *ArXiv:2203.08504 [Cs]*. http://arxiv.org/abs/2203.08504

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. *ArXiv:2103.15348 [Cs]*. http://arxiv.org/abs/2103.15348

World Meteorological Organization. (2016). *Guidelines on Best Practices for Climate Data Rescue*. https://library.wmo.int/doc_num.php?explnum_id=3318

Appendices

Appendix A. Survey questions about AI/machine learning & data rescue from Chapter 3

1. You consent to be identified by name in reports (required).

   A. Yes
   B. No

2. You selected yes in question 1, please add your name here.

3. You consent to have your organization's name used (required).

   A. Yes
   B. No

Background

4. What do you predominantly "identify" as?

   A. Citizen science researcher
   B. Data rescue researcher
   C. Other, please specify [ ]

5. How many years have you been working on citizen science or data rescue?

6. Do you have experience in the field of meteorology or climatology? Choose the one closest to your experience.

   A. I have a degree in or work in the field of meteorology or climatology
   B. I have taken one or more meteorology or climatology related courses
   C. I regularly read/watch meteorology or climatology materials
   D. I have been involved in meteorology or climatology related projects/initiatives
   E. I have a strong interest in meteorology or climatology, but have done none of the things above
   F. I consider myself an amateur

7. Machine learning refers to algorithms that learn and adapt without needing to follow explicit instructions. What experience do you have related to machine learning? Check all that apply.

   A. I use it extensively in my research/work

B. I've coded some algorithms
C. I have run machine learning algorithms written by others
D. I took at least one machine learning course
E. I have watched some videos or online materials on machine learning (30 minutes or more)
F. I know what machine learning is, although I've never used it
G. I have no experience in using or learning about machine learning

Please tell us about your transcription projects

8. What kinds of data are you transcribing? Check all that apply.

A. logbooks
B. diaries
C. newspapers
D. museum labels
E. Other, please specify [ ]

9. Which computer-related technologies/software do you use to capture/collect the information (e.g., spreadsheet, online platform)? Please select.

A. Microsoft Excel
B. Database system
C. Local/Web server
D. OCR algorithm software
E. Online platform/resource (please specify in 'other')
F. Other, please specify [ ]

10. You may be involved in several data rescue or citizen science projects. Generally, what is your role in the project/activity?

A. Project manager
B. Involved with system development, maintenance
C. Researcher
D. Amateur transcriber
E. Other, please describe [ ]

11. Generally, how are the transcription projects you work on funded? Check the one item that best applies.

A. I or my colleague need to write grants or contracts to fund the work
B. It is part of my job (i.e., my employer pays)
C. My own pocket
D. A mix of above

E. I did not setup the system so I do not know
F. Unknown
G. None of the above, please describe [ ]

12. Citizen science is scientific research conducted, in whole or in part, by amateurs (or non-professionals). Sometimes it is referred to as public participation in scientific research. Have you been involved in any citizen science projects? If you have, please explain.

13. Data rescue is a concept, predominantly in the historical weather community, to preserve the data from being lost or deteriorated by transferring the data into computer compatible formats and updating the records continuously to support various media versions. Have you been involved in any data rescue projects? If you have, please explain.

Perception of Goals of Citizen Science/ Data Rescue

14. What do you think is the most important goal of citizen science related to weather? (Please answer even if your work is not directly related to citizen science.)

A. Provide researchers with useful data
B. Make citizens a part of advancing science
C. Citizens can better protect the environment via their contributions
D. Improve citizen education and literacy
E. Make citizens more aware of the history of weather
F. Provide an opportunity for recreation
G. Other, please specify [ ]

15. What do you think is the most important goal of data rescue related to weather? (Please answer even if your work is not directly related to data rescue.)

A. Provide researchers with useful data
B. Preserve fragile records
C. Improve weather/climate data interoperability
D. Help build a global model of the past climate
E. Validate climate models with records of the past
F. Aid in better understanding human's role in climate change
G. Other, please specify [ ]

Perception and Knowledge of Automation of Transcription

16. If the transcription portion of your citizen science/data rescue activity could be automated (e.g., via optical character recognition {OCR}), would you consider automation?

A. Yes
B. No

C. I don’t know
D. Other, please specify [ ]

17. If automation would provide you with a pre-trained algorithm with a customized user interface where you can adjust the parameters according to your project, what might be reasons not to automate your citizen science/data rescue process? If there are multiple reasons, please rank them in the next question.

A. Lack of funding
B. Too much work and effort needed to shift the process
C. The performance of automation is not guaranteed
D. Do not have experts to maintain the automation
E. There are no obvious benefits to do so
F. Have tried automation already, but the result is not promising
G. There are benefits to involving citizens that would be reduced if transcription was automated
H. Automation algorithms are just too “black box”, opaque to trust
I. Other, please specify [ ]

18. Please rank them by sequence.

19. If it took an acceptable short amount of time to set up automated data transcription, would you use it to make the data machine readable? How much time would you consider to be an acceptable amount to invest in setting up the process?

20. Given what you may know about attempts to automate transcription, how confident are you that automation will accurately transcribe the data? Please choose from “Not at all confident”, “Not very confident”, “Somewhat confident”, “Very confident”, and “Don’t know”.
And please briefly explain.

21. What, in your opinion, might we gain or lose if we started to use automation instead of humans (e.g., citizen science or paid transcribers)?

22. We are part of a research network attempting to develop an AI augmented OCR system. How confident would you be if you were asked to do the following, which may be a part of a system like this? For each one, please clearly indicate whether you are “Very confident”, “Somewhat confident”, “Not very confident”, “Not at all confident”, or “Don’t know.”

	Not at all confident	Not very confident	Somewhat confident	Very confident	I don’t know
Build a database on your local machine					
Use SQL to query a database					

Configure a virtual environment on your local computer					
Code in a programming language like Python, C, R, or Java					
Run a machine learning algorithm such as linear regression, K-nearest neighbor, Bayes, and random forest					
Use software to produce a coefficient matrix or a scatterplot of paired data					
Run an Optical Character Recognition script/software to digitize documents					
Run a deep learning algorithm, such as Convolutional Neural Network (CNN), Reinforcement Learning, and Long-short term Memory (LSTM)					
Use machine learning library like TensorFlow or PyTorch to program and run a script					
Use HTML, CSS, JavaScript, or PHP to write a web page with interactive functions					
Use GitHub to document your workflow					
Solve technical issues using platforms like StackOverflow					

Conclusion

23.Do you have any other comments or feedback that you would like to share?

24. Thank you for taking your time to complete this survey. If you would like a copy of the results then please supply your email address:

Appendix B. Research Ethics Board approval for Chapter 3

McGill

Research Ethics Board Office
James Administration Bldg.
845 Sherbrooke Street West. Rm 325
Montreal, QC H3A 0G4

Website: www.mcgill.ca/research/research/compliance/human/

Research Ethics Board 1
Certificate of Ethical Acceptability of Research Involving Humans

REB File #: 21-05-005

Project Title: A Comparison of Citizen Science and Machine Learning Techniques on Handwritten Text Recognition: A Case Study of Historical Climate Data Rescue

Principal Investigator: Yumeng Zhang

Status: Master's Student

Dept: Geography

Supervisor: Prof. Renée Sieber

Funding: FQRNT (PI-Frederic Fabry)

Approval Period: June 28, 2021 to June 27, 2022

The REB-1 reviewed and approved this project by delegated review in accordance with the requirements of the McGill University Policy on the Ethical Conduct of Research Involving Human Participants and the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans.

Deanna Collin
Senior Ethics Review Administrator

---

* Approval is granted only for the research and purposes described.
* Modifications to the approved research must be reviewed and approved by the REB before they can be implemented.
* A Request for Renewal form must be submitted before the above expiry date. Research cannot be conducted without a current ethics approval. Submit 2-3 weeks ahead of the expiry date.
* When a project has been completed or terminated, a Study Closure form must be submitted.
* Unanticipated issues that may increase the risk level to participants or that may have other ethical implications must be promptly reported to the REB. Serious adverse events experienced by a participant in conjunction with the research must be reported to the REB without delay.
* The REB must be promptly notified of any new information that may affect the welfare or consent of participants.
* The REB must be notified of any suspension or cancellation imposed by a funding agency or regulatory body that is related to this study.
* The REB must be notified of any findings that may have ethical implications or may affect the decision of the REB.

Appendix C. *p*-value of confidence in accuracy from Chapter 3

	Whether respondents are confidence that automation will accurately transcribe the data:		
	Yes	**No**	***p*-value**
Identity			
Self-identified as both data rescuer and citizen science researcher	12	10	0.7597
Self-identified as one of data rescuer and citizen science researcher	13	8	
Willingness of using automation			
Yes	23	15	0.2972
No	1	3	
Years been working			
⩽10 years (median)	12	10	0.7597
>10 years (median)	13	8	
Confidence in doing automation related technical tasks			
Confident	19	11	0.3318
Not confident	6	7	
Whether respondents have experience in machine learning before			
Yes	11	3	0.0574
No	13	15	
If respondents are working in climatology related fields			
Yes	20	14	0.7061
No	4	4	