THE UNIVERSITY *of* EDINBURGH

Recalibrating Machine Learning for Social Biases: Demonstrating a New Methodology through a Case Study Classifying Gender Biases in Archival Documentation

Lucy Joan Havens

Doctor of Philosophy
Institute for Language, Cognition and Computation
School of Informatics
University of Edinburgh
2024

Abstract

This thesis proposes a recalibration of Machine Learning for social biases to minimize harms from existing approaches and practices in the field. Prioritizing quality over quantity, accuracy over efficiency, representativeness over convenience, and situated thinking over universal thinking, the thesis demonstrates an alternative approach to creating Machine Learning models. Drawing on GLAM, the Humanities, the Social Sciences, and Design, the thesis focuses on understanding and communicating biases in a specific use case. 11,888 metadata descriptions from the University of Edinburgh Heritage Collections' Archives catalog were manually annotated for gender biases and text classification models were then trained on the resulting dataset of 55,260 annotations. Evaluations of the models' performance demonstrates that annotating gender biases can be automated; however, the subjectivity of bias as a concept complicates the generalizability of any one approach.

The contributions are: (1) an interdisciplinary and participatory Bias-Aware Methodology, (2) a Taxonomy of Gendered and Gender Biased Language, (3) data annotated for gender biased language, (4) gender biased text classification models, and (5) a human-centered approach to model evaluation. The contributions have implications for Machine Learning, demonstrating how bias is inherent to all data and models; more specifically for Natural Language Processing, providing an annotation taxonomy, annotated datasets and classification models for analyzing gender biased language at scale; for the Gallery, Library, Archives, and Museum sector, offering guidance to institutions seeking to reconcile with histories of marginalizing communities through their documentation practices; and for historians, who utilize cultural heritage documentation to study and interpret the past. Through a real-world application of the Bias-Aware Methodology in a case study, the thesis illustrates the need to shift away from removing social biases and towards acknowledging them, creating data and models that surface the uncertainty and multiplicity characteristic of human societies.

Lay Summary

Existing approaches to creating Machine Learning (ML) systems build harmful social biases, such as gender and racial biases, into the systems, through their data and models. Drawing on approaches from Galleries, Libraries, Archives, and Museums (GLAM); as well as the Humanities, the Social Sciences, and Design; I propose a new approach to creating ML systems that makes social biases in the systems visible. Research was undertaken for a specific use case: detecting gender biased language in the Archives catalog of the University of Edinburgh's Heritage Collections. Creating a dataset of archival metadata descriptions manually labeled for gender biases, ML models were then trained on the dataset to automatically detect gender biased language. The models' performance was evaluated quantitatively, with typical, numeric ML metrics, and qualitatively, with feedback from Heritage Collections employees. These evaluations show that models can automate the detection of certain types of gender biases to help make them visible in language data, but they cannot automate the detection of all types of gender biases. Moreover, different approaches are needed for different types of gender biases.

The thesis has implications for ML, GLAM, and History. For ML, this thesis shows how all ML systems will inevitably be biased. Consequently, ML systems' creators must consider the context they build and deploy these systems in to avoid harming communities already experiencing discrimination. For GLAM, this thesis provides an ML approach that can help curators, librarians, archivists, and catalogers understand the gender biases in their institutions' catalogs. For historians, who use GLAM catalogs and the collections they describe to study the past, this thesis shows how subjective the catalogs' descriptions are, which cause biases that must be considered when writing about the past. Overall, this thesis calls for a shift away from removing social biases and towards acknowledging them, creating technologies that surface the uncertainty and multiplicity of knowledge.

Acknowledgements

Thank you to Beatrice Alex, Benjamin Bach, and Melissa Terras for their guidance and encouragement, and to Rachel Hosker and the University of Edinburgh's Heritage Collections team for their enthusiastic collaboration. Thanks also go to Adam Lopez, for his willingness to share his feedback with me at the end of each year of my research, to Uta Hinrichs and Mark Kobine, who were always available for a chat and pushed my thinking in creative directions, and to Atoosa Kasirzadeh and Mary Flanagan, for their willingness to examine this thesis and provide feedback for strengthening it. Thank you to my family and friends, for supporting and inspiring me.

Additionally, I extend my gratitude to the Engineering and Physical Sciences Research Council and the School of Informatics Graduate School for funding my Ph.D. research, and to the Centre for Data, Culture, and Society; the Edinburgh Futures Institute; and the Scottish Informatics and Computer Science Alliance for awarding me with grants that enabled me to hire annotators, travel to conferences, and promote my research.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification. The main content of the chapters listed below have been published in peer-reviewed venues; I declare that I substantially contributed to these publications as lead researcher and author.

- **Chapter 3:** This chapter is based on a shorter publication, *Confronting Gender Biases in Heritage Catalogs: A Natural Language Processing Approach to Revisiting Descriptive Metadata*, which will appear in the *Routledge Handbook on Heritage and Gender* (expected 2024). I conducted the literature review for the publication and I wrote the publication as lead author, with my supervisors providing feedback to guide my revisions.

- **Chapter 4:** I originally wrote my thesis' methodology as a paper titled *Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research*. The paper was published in the *Proceedings of the Second Workshop on Gender Bias for NLP* (Havens et al., 2020), as part of the virtual 28th International Conference on Computational Linguistics. I conducted the literature review, developed the methodology, and executed the case study reported in the paper, and I wrote the paper as lead author, with my supervisors providing feedback as I wrote to guide my revisions.

- **Chapter 5:** I originally wrote §5.1 as a paper titled *Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text*. The paper was published in the *Proceedings of the Fourth Workshop on Gender Bias for NLP* (Havens, Terras, et al., 2022), as part of the North American Chapter of the Association for Computational Linguistics Conference. I wrote the paper as lead author, with my supervisors providing feedback as I wrote to guide my revisions. My responsibilities as lead author included executing all Participatory Action Research sessions; recruiting, hiring,

and training annotators; conducting annotations as lead annotator and performing all data transformations and analysis.

- **Chapter 6:** A shorter version of the classification work reported in this chapter was first published and presented at the Digital Humanities Conference (Havens et al., 2023; Appendix K). I wrote the publication as lead author, with my supervisors providing feedback to inform my revisions to the paper. I performed all the programming work to build and evaluate the models reported in the publication.

Lucy Joan Havens

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The increased integration of digital technologies in the lives of many people around the globe, coupled with advances in computing hardware that provide greater data storage and processing capacities, have provided Machine Learning (ML) researchers and practitioners with the ability to train larger models on larger datasets (Bommasani et al., 2021; Kaplan et al., 2020). As harms from deployments of such ML systems have surfaced (Bender et al., 2021; Noble, 2018; O'Neil, 2016; L. Sweeney, 2013), the need for new areas of research on fairness, explainability, ethics, and bias for ML has become evident. These new areas of research identify many potential sources of unfair, unethical, or biased behavior in ML systems, from a model's training data to benchmarks for evaluating models (Hovy and Prabhumoye, 2021; Suresh and Guttag, 2021; Welty et al., 2019). A focus on data has emerged (Aragon et al., 2022; Aroyo et al., 2022; D'Ignazio and Klein, 2020; Jo and Gebru, 2020; Rogers, 2021; Sambasivan et al., 2020; Thylstrup, 2022), as scholars have demonstrated that those excluded and misrepresented in an ML model's training data will be excluded and misrepresented by the ML system (Buolamwini and Gebru, 2018; Crawford and Paglen, 2019; Keyes, 2018; Scheuerman et al., 2019). Recognizing parallels between catalogs in the Gallery, Library, Archives, and Museum (GLAM) sector and ML datasets, scholars have begun looking to GLAM literature for guidance on more critical approaches to ML dataset creation (Agostinho et al., 2019; Denton et al., 2020; Jo and Gebru, 2020; Thylstrup, 2022; Thylstrup et al., 2021).

In the GLAM sector, bias has been conceptualized with greater complexity than in ML research and practice (Havens et al., 2020, 2022). Since the

late 20th century, catalogers, librarians, archivists, and curators increasingly adopted a postmodern philosophy that views GLAM catalogs' metadata and descriptions as constructed, incomplete, and contingent (Duff and Harris, 2002; Tai, 2021). Drawing on experience working with heritage material, researchers and practitioners in GLAM have articulated how the sector's practices of acquiring, preserving, describing, and providing access to heritage are shaped by contextual factors (Caswell, 2022; Caswell and Cifor, 2016, 2019; Cook, 2011; Duff and Harris, 2002; Schwartz and Cook, 2002; Stoler, 2002; Yale, 2015; further discussed in §1.1). Inevitably, these GLAM practices result in simultaneous inclusion and exclusion, determined by the choice of which contextual relationships to make explicit in a cultural heritage record (Bowker and Star, 1999; Duff and Harris, 2002).

That being said, many GLAM institutions existed before the postmodern philosophy was widely adopted, when their practices of acquiring, appraising, preserving, describing, and providing access to heritage were viewed as objective representations of the world. Consequently, GLAM have begun dedicating resources to reviewing the older descriptions of heritage items in their catalogs, aiming to understand their historical biases and add additional context where needed to inform interpretations of the heritage material being described (Antracoli et al., 2019; Berry, 2020; Collections Trust, 2023; Wetli, 2019). In this thesis I refer to this process as "collection reviews;" GLAM researchers and practitioners have also referred to this and similar processes as "inclusive description," (Berry, 2020), "conscious editing" (Berry, 2020), "critical cataloging," (Berry, 2020), "anti-racist description" (Antracoli et al., 2019), and "reparative" and "anti-oppressive" description (Tai, 2021), among other terms. The need to balance describing new heritage material with revisiting existing descriptions opens up a new use case for ML in GLAM: supporting GLAM collection reviews.

Although ML applications to GLAM are not new (Cordell, 2020; Jaillant, 2022), these applications tend to focus on digitization (Hosseini et al., 2022; Nockels et al., 2022; Padilla, 2017; Pal et al., 2016), evaluation and generation of metadata (Berardi et al., 2012; Cordell, 2020; De Bonis et al., 2023; Padilla, 2019; Yilmazel et al., 2004), and analyzing collections (Ames and Havens, 2022; Beelen et al., 2023; Beelen et al., 2022). Recognizing the ML and GLAM communities' shared concern with biased data, and the lack

of research at the intersection of ML, GLAM, and social biases, I focused my Ph.D. research on studying biased language in the descriptive metadata of GLAM catalogs using ML methods. ML methods offer approaches to support reviews of an entire GLAM catalog, which could increase the efficiency of the currently manual collection review process, supporting GLAM institutions' primary aim of making cultural heritage discoverable for the public (Jaffe, 2020; Thomassen, 2002; Welsh, 2016). GLAM methods offer approaches to address the complexities of data, considering the incomplete, uncertain, and contextual nature of data (Adler, 2017; Bowker and Star, 1999; Caswell and Cifor, 2016; Drucker, 2021; Shopland, 2020; Tai, 2021; Thylstrup, 2022).

This thesis thus bridges the ML and GLAM communities, investigating the extent to which ML models can identify bias across collections' descriptions in a GLAM catalog, guided by approaches to bias proposed in GLAM literature. Prioritizing accuracy over efficiency, representativeness over convenience, quality over quantity, and situated thinking over universal thinking (detailed in §2.2), this thesis recalibrates the typical approach to ML to more effectively mitigate harms from social biases in ML systems. Rather than attempting to remove or minimize bias, the thesis proposes that bias be accepted as inevitable and communicated to readers explicitly. Due to the complexities of bias, studying all biased language was deemed too large an undertaking for a Ph.D. of three and a half years. Furthermore, differences in the structure of Galleries, Libraries, Archives, and Museums' catalog metadata made a review of every type of institution's catalog documentation out of scope for a Ph.D. thesis. As a result, I focused specifically on identifying gender biased language, suiting my position as a member of a minoritized[1] gender group, women. I study gender bias in an archival catalog's documentation; Archives' catalogs generally contain lengthier descriptions than those of other types of GLAM institutions (Thomassen, 2002). Working with GLAM documentation meant my data would be text-based, so I narrowed my computational focus within ML to Natural Language Processing (NLP), as this subdiscipline applies rule-based and ML methods to textual data.

While much ML research on bias aims to minimize or remove bias (Andriyansah et al., 2019; Bolukbasi et al., 2016; Bourgeois et al., 2018; Dinan, Fan,

[1] In this thesis I use *minoritization* in the sense D'Ignazio and Klein (2020) use the term: as a descriptor to emphasize a group of people's experience of oppression, rather than using the noun *minority*, which defines people as oppressed.

Williams, et al., 2020; Dixon et al., 2018; Kaneko and Bollegala, 2019; Zhao, Wang, Yatskar, Ordonez, et al., 2018; Zhao, Zhou, et al., 2018), I questioned the underlying assumption of this work: that objective, neutral, or universal technologies could be built. For certain use cases, such as social media (Schmidt and Wiegand, 2017), creating ML systems with the goal of removing biased language in the form of hate speech is feasible, based on relative consensus among nations on definitions of hate speech,[2] and justifiable, based on the social media platforms' terms of use.[3] Biased language that is not explicitly hateful, however, is more difficult to address. Removing this type of language simply erases or hides evidence of society-wide injustices; it does not remove the injustice itself (Gonen and Goldberg, 2019; Hessel, 2023c; Ortolja-Baird and Nyhan, 2022). Additionally, bias is context-dependent: the same word or sequence of words may indicate bias in one document or conversation, but not in another, because the meaning of language comes from contextual relations, such as relationships between speakers, or author and reader, or time and place (Duff and Harris, 2002; Fairclough, 2003; Havens et al., 2020).

Looking to discussions of bias in GLAM literature informed by feminist theories, critical discourse analysis, and authors' experience working with GLAM collections and their descriptions (Caswell and Cifor, 2016; Cook, 2011; Duff and Harris, 2002; Hessel, 2023b; Smith, 2006), I recognized oversimplified conceptualizations of bias in ML literature (a recognition echoed in Crawford, 2017; Blodgett et al., 2020; Stańczak and Augenstein, 2021; and Devinney et al., 2022) that was limiting the ML community's progress on addressing harms from biased ML systems. Rather than ask how to adjust existing ML systems and practices, I sought to create new systems and practices. I began my research on bias in ML by asking different questions. Instead of, *How can I remove biases from ML data and models?*, I asked, *How can I use ML models to make biases in data explicit?* This new starting point assumes that bias is inevitable, and that ML data and models are contingent on the ever-changing context in which they are designed, created, and used. Consequently, this new starting point further assumes that ML systems are intertwined with society, and the power relationships between individuals and institutions in society

[2] See the United Nations' definition of "hate speech:" www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

[3] For example, the Facebook Community Standards: transparency.fb.com/policies/community-standards/hate-speech/

that have caused structural injustices (further discussed in Chapter 2).

1.1 Research Questions

Conducting Digital Humanities research at the intersection of GLAM and ML, I sought to study the relationship between data and social biases, using NLP methods to identify potentially biased language, and GLAM and Humanities approaches to identify the uncertainties and complexities surrounding measurements of biased language (Risam, 2021; Terras et al., 2013). At the time of writing, no ML systems had yet been developed for identifying social biases in GLAM catalogs. I present the first case study investigating both the capabilities and the limitations of ML methods for locating and communicating biases in GLAM catalogs, applying NLP methods to the descriptive language of an archival catalog, that of the Heritage Collections Archives at the University of Edinburgh. Through the case study, I demonstrate how to recalibrate ML for social biases, creating and implementing alternative approaches to building datasets and models, and evaluating those datasets and models. The promising results of my models' performance evaluations (chapters 6-7) illustrate the value of moving away from attempts to repair existing ML systems, and towards deconstructing and rebuilding ML systems (Hicks, 2018; Morrison, 2021).

In order to answer the question, *How can I use ML models to make biases in data explicit?*, I needed to identify and measure different types of bias, in my case, gender biased language in an archival catalog's descriptive metadata. While many approaches to mitigating biased language in NLP focused on minimizing social biases in abstracted, numeric representations of words (called embeddings) that could be input into an NLP model, research investigating the efficacy of this approach reported limitations and inconclusive results (Goldfarb-Tarrant et al., 2021; Gonen and Goldberg, 2019). Moreover, these approaches failed to provide definitions, or provided overly-simplified definitions, of biased language (Blodgett et al., 2020; Crawford, 2017), especially gender biased language (Devinney et al., 2022; Stańczak and Augenstein, 2021; further discussed in Chapter 4.1.4). I took an alternative approach, studying gender biased language in its original representation: words, how they are arranged in meaningful sequences, and how they are

interpreted in different contexts.

While conceptualizations of gender, bias, and gender biased language are often too simplified in NLP, conceptualizations of these terms in Linguistics, History, and Gender Studies have been theorized in greater complexity. Beard (2017) has written of efforts to silence women that date back to antiquity, noting how women's voices and manner of speaking were described with derogatory terms, such as "bark" and "whine," discouraging women from raising their voice to share their perspective and in turn establishing stereotypes that assert women's less powerful status in society. Problematically, these derogatory terms continue to be used today, perpetuating harmful stereotypes and continuing to encourage the omission of women from public debate and historical records (Beard, 2017; Haines et al., 2016; Talbot, 2003). Hessel (2023a, 2023c) has written of the way in which artists who are women continue to be described in the media as "women artists," and described primarily in relation to men with whom they had relationships (e.g. "lover of," "muse of"), rather than their artistic careers. Such language reveals an underlying assumption of an artist being a man, and of a woman's identity being defined primarily by the men around her. Temporal context adds complexity to identifying gender biased language, though, because words' meanings can shift connotations over time (Bucholtz, 1999; Garg et al., 2018; Schulz, 2000; Shopland, 2020). For example, drag communities have adopted stereotypical "women's language" to assert their femininity in a society with people who attempt to reject that femininity (Bucholtz, 2003). Intersecting identity characteristics also add complexity to gender bias; for example, people of the same gender but different racialized ethnicities will experience privilege and oppression differently (Crenshaw, 1989, 1991).

To investigate these complexities of gender biased language, I defined three research questions that guide the research I report in this thesis:

1. Can existing methods of identifying and categorizing gender biased language in NLP research be applied to archival metadata descriptions? Why or why not?

2. What types of gender bias are present in the language of archival metadata descriptions?

3. Can gender biased language in archival metadata descriptions be

reliably annotated by domain experts to create data on which to train NLP classification models?

These research questions enabled me to develop a definition of gender biased language that better accounts for its complexity, investigate the capabilities and limitations of NLP models for measuring gender biased language, and reflect upon the societal implications of data and ML models' inevitable biases. Prior to investigating my research questions, I developed a new research methodology that approaches ML systems as socio-technical, rather than purely technical, systems. This Bias-Aware Methodology, detailed in Chapter 4, is the first of the five contributions of this thesis.

1.2 Contributions

Contribution 1: Bias-Aware Methodology

Drawing not only on ML, but also the Humanities, Design, and Social Sciences, the Bias-Aware Methodology (the Methodology) I developed for thesis is both interdisciplinary and participatory. From the Humanities, the Methodology draws on critical discourse analysis (Bucholtz, 1999, 2003; Fairclough, 2003; Marston, 2000; Smith, 2006; Talbot, 2003) and feminist theories (Crenshaw, 1989, 1991; D'Ignazio and Klein, 2020; Haraway, 1988; Harding, 1995) as lenses for studying the language of an archival catalog's metadata descriptions. From Design and the Social Sciences, the Methodology draws on Participatory Action Research (PAR) (P. Leavy, 2017a; Martin and Hanington, 2012b; Reason and Bradbury-Huang, 2007; Reid and Frisby, 2008; Swantz, 2008) and the case study method (Martin and Hanington, 2012a), investigating the research questions of this thesis within a specific time and place, among a specific group of people, and for a particular type of social bias. The Methodology consists of three activities: defining the bias of focus, examining power relations, and executing ML methods (in my case, NLP methods, specifically). These three activities are to be executed in parallel, with each activity informing the others.

I created the Methodology due to gaps in ML literature on interdisciplinary and participatory approaches to research. Though scholars had called for interdisciplinary and stakeholder collaboration (Blodgett et al., 2020; Crawford, 2017; Devinney et al., 2022; Stańczak and Augenstein, 2021),

I had yet to see this executed in an ML project end to end, from problem formulation through model evaluation. Narrowing my focus to gender bias research in NLP, I outlined the Methodology and illustrated how I was applying it in my own research, aiming to provide practical guidance for NLP researchers and practitioners on engaging with interdisciplinary methods and with stakeholders. Chapter 4 articulates my research perspective in a bias statement, explains the need for the Methodology, presents the Methodology, and introduces the case study in which I situate the research reported in this thesis, working with the University of Edinburgh's Heritage Collections (HC) team and their Archives' catalog. Through the development and application of the Methodology, I investigate my first research question: ***Can existing methods of identifying and categorizing biased language in NLP research be applied to archival metadata descriptions? Why or why not?*** In chapters 4 and 5 I explain how and why only Hitti et al.'s (2019) approach to identifying and categorizing gender biased language was partially applicable to my research.

The Methodology was originally published in the *Proceedings of the Second Workshop on Gender Bias for NLP* (Havens et al., 2020). In this thesis I extend the Methodology beyond NLP to ML. Chapters 5- 7 report on applications of the Methodology to create four further contributions of this thesis: an annotation taxonomy, annotated dataset, text classification models, and a human-centered model evaluation approach.

Contribution 2: Annotation Taxonomy

Chapter 5, Annotated Data Curation, begins with the creation of a Taxonomy of Gendered and Gender Biased Language (the Taxonomy). The Taxonomy contains three categories of labels with which human annotators, and subsequently NLP models (Chapter 6), annotate metadata descriptions from an archival catalog. Literature from critical discourse analysis, gender studies (including feminist and queer theories), and NLP informed the initial draft of the Taxonomy, which I finalized in a workshop with members of the HC team. During the manual annotation process, I further refined the Taxonomy with hired annotators, based on the content of the dataset of archival metadata descriptions we annotated.

Including gendered language in addition to gender biased language enables

the Taxonomy to inform measurements of gender bias at close and distant views. Gendered language refers to terminology with a grammatical gender association, namely masculine (e.g. "brother," "Sir,"), feminine (e.g. "Queen," "she"), or non-binary (e.g. "Mx.," "they" as a singular pronoun). Grammatical gender does not correlate with gender identity; people of multiple gender identities may refer to themselves with masculine terms, feminine terms, or non-binary terms (Scheuerman, Spiel, et al., 2020; Spiel et al., 2019). That being said, when looking across a text corpus, such as an archives' catalog, counting the frequency of terms that fall into these grammatical gender categories provides an indication of which genders are more likely to have been represented and which genders are more likely to have been excluded, providing a distant, summary view of gender bias in the form of omission. At a close view, annotations of gender biased language, such as language that unjustly reinforces gender stereotypes, provides insight on the variety of ways gender biases manifest in language, while also providing specific instances of bias to be counted across a text corpus for an additional bias measurement.

The Taxonomy was originally published in the *Proceedings of the Fourth Workshop on Gender Bias for NLP* (Havens, Terras, et al., 2022). I use the Taxonomy to guide the creation of annotated data to train gender biased text classification models, applying the labels in the Taxonomy to archival metadata descriptions. Thus the Taxonomy enabled me to investigate my second research question: ***What types of gender bias are present in the language of archival metadata descriptions?*** The Taxonomy offers NLP researchers and practitioners a framework for measuring and analyzing gender biases in a text-based dataset.

Contribution 3: Annotated Datasets

The remainder of Chapter 5, Annotated Data Curation, details the process of curating a dataset to annotate, and annotating the dataset according to the Taxonomy. Participatory action research conducted with the HC team guided my choice of metadata descriptions to extract. This chapter describes the process of using the Open Archives Initiative - Protocol for Metadata Harvesting to extract metadata descriptions from Heritage Collections' online archival catalog, transforming them from eXtensible Markup Language (XML) to Plaintext (TXT) format. The hierarchical format of the metadata

descriptions, organized according to the General International Standard for Archival Description, was flattened for annotation purposes, however every description can be linked back to the *fonds* (the archival term for collection) for which it was written.

Four annotators were hired to annotate, along with myself, 20% of the extracted catalog descriptions according to the Taxonomy. 10% of this subset of descriptions were triply annotated and the remaining 90%, doubly annotated. Several internal institutional grants were successfully applied for to fund the annotation work, which totaled £5,333.76, with the four hired annotators receiving £18.52 an hour to work 72 hours across eight weeks. The annotation process resulted in five annotated datasets, enabling me to respond to my second research question: ***What types of gender bias are present in the language of archival metadata descriptions?*** The most commonly-applied label from the ***Linguistic*** category was *Gendered Pronoun*; from the ***Person Name*** category, *Unknown*; and from the ***Contextual*** category, *Omission*. Measures of overlap between the annotators' labels, called inter-annotator agreement, reflected the subjectivity of the task of identifying gender biased language: labels of gendered language overlapped more than labels of gender biased language. I manually reviewed 97,861 instances of matching or overlapping text spans with different labels assigned by different annotators to determine which labels should be kept in accordance with the annotation instructions. This manual review informed the aggregation of the five annotated datasets into a single aggregated dataset of 55,260 annotations, which then became the training, development, and test data for text classification models.

The explanation of the annotation process was originally published with the Taxonomy in the *Proceedings of the Fourth Workshop on Gender Bias for NLP* (Havens, Terras, et al., 2022). I used the aggregated dataset of annotated archival documentation to train my text classification models with a supervised learning approach. The disaggregated and aggregated annotated datasets offer NLP researchers and practitioners with a text corpus for studying how gender biases manifest in British English language.

Contribution 4: Text Classification Models

Chapter 6 describes the creation of my fourth contribution, NLP models for classifying gendered and gender biased text. This chapter investigates the research question: ***Can gender biased language be reliably annotated by domain experts to train a classification model to automatically annotate gender biased language?*** Several approaches to model training were considered: token, sequence, and document classification; traditional ML (without neural networks) and deep learning (with neural networks); and training on disaggregated data or aggregated data. I settled on a combination of token, sequence, and document classification with traditional ML models on an aggregated dataset.

I evaluate the performance of the classification models quantitatively, with standard NLP metrics, in this chapter. These evaluations indicate that certain types of gender biases can be annotated more reliably than others. The ***Linguistic*** category's gendered language labels, *Gendered Pronoun* and *Gendered Role*, were highly reliable. The ***Person Name*** category of gendered language labels, *Feminine*, *Masculine*, and *Unknown*, were less reliably annotated, reflecting the difficulty of manually annotating names according to the grammatical gender of terminology referring to those names. The ***Linguistic*** category's gender biased language label, *Generalization*, was unreliable, again reflecting the high levels of disagreement among manual annotators with that label. The ***Contextual*** category's *Occupation* label was annotated more reliably than the ***Person Name*** labels, suggesting it could be used to study correlations between certain genders and their jobs, and thus occupational gender biases reinforced with and subverted in the data. The ***Contextual*** category's gender biased language labels, *Omission* and *Stereotype*, were reliably annotated, though not as reliably as *Gendered Pronoun* and *Gendered Role*.

A shorter version of Chapter 6's report of the text classification models was published and presented at the 2023 Digital Humanities Conference (Havens et al., 2023; Appendix K). The models provide GLAM researchers and practitioners with an automated approach to study gender biases in GLAM catalogs' documentation, supporting their existing, largely manual, descriptive practices. Additionally, the models provide NLP researchers and practitioners

with a new approach to addressing gender biased language that makes the biases visible, rather than attempting to minimize the biases.

Contribution 5: Participatory Evaluation

I also evaluated the text classification models with their intended audience: the HC team. Chapter 7 reports on this thesis' final PAR activity with the HC team, for which a workshop was conducted to complement Chapter 6's quantitative evaluation of the models with human-centered evaluations. This qualitative approach provides both ML researchers and practitioners with a new, human-centered approach to evaluating models. The qualitative approach also provides GLAM researchers and practitioners with a framework for collaborating with ML researchers and practitioners, incorporating GLAM domain expertise into evaluations of models' performance for a GLAM use case.

In this thesis, I framed the workshop discussion around two worksheets of data visualizations to facilitate collaborative analysis of the manual annotation and model classification results. The workshop asked archivists, librarians, and curators in HC about the utility of the Taxonomy, based on text visualizations of its application to three example descriptions, and the utility of summary measures of gender bias in the HC Archives' catalog, based on bar charts and tables quantifying the models' annotations. In addition to answering these questions, the workshop discussion also provided insights on the uncertainty of data; the inevitability of bias and complexities with mitigating its harms; the power relationships at play in data curation, description, and access; and the importance of transparency in documentation. The results of the workshop thus contribute to ML (including NLP) and GLAM understandings of the complexities of social biases, and the capabilities and limitations of models for automating approaches to addressing those biases in data.

This thesis proposes a recalibration of ML for social biases, providing a widely-applicable Methodology to guide ML researchers and practitioners towards this recalibration, and a case study demonstrating how to implement the Methodology. Implementing one's own recommendations offers invaluable insights on the practicality of the recommendations, as well as the challenges that may complicate the implementation of those recommendations. Typically, the ML community approaches model creation top-down, aiming for high

scores against quantitative metrics and claiming universal relevance once those high scores are achieved (Raji et al., 2021). Rather than aiming to create the highest scoring models for gender biased language classification, or aiming to create highly generalizable models, this thesis uses a real-world use case to demonstrate the shortcomings of this top-down approach. Chapter 2 details four priorities that characterize the top-down approach to ML research and practice, and proposes four recalibrations for ML research and practice that better accounts for the structural nature of social biases. I explain how each contribution supports this recalibration at the end of chapters 4-7. Chapter 3 summarizes relevant literature on social biases from ML and GLAM that further illustrates the gaps my research addresses and the new directions my research supports.

1.3 Definitions

To ensure clarity of communication, in this section I define key terminology used throughout the thesis.

Machine Learning, abbreviated ML, refers to the discipline and technology that uses algorithms to find patterns in datasets. Though ML and *Artificial Intelligence (AI)* are defined differently in academic literature (Eisenstein, 2018), the terms are widely used interchangeably. AI uses ML in an aim to create systems that have human-like intelligence. This thesis simply uses the term ML to refer to the disciplines and technologies of ML and AI.

Natural Language Processing, abbreviated NLP, overlaps with ML but also includes rule-based methods. NLP is often used interchangeably with the term *computational linguistics*. However, the aim of computational linguistics work is often focused on understanding aspects of language, bringing computational methods to the Linguistics discipline. NLP work, on the other hand, often focuses on large-scale analysis of human-written language, such as summarizing documents, answering questions, extracting information relevant to a given query, or categorizing sentences by their sentiment (Eisenstein, 2018).

Cultural heritage, or simply heritage, as discussed in this thesis relies heavily on the work of Smith (2006), who conceptualizes cultural heritage as inclusive of tangible and intangible records of historical and cultural significance. Moreover, Smith views heritage records as dynamic, rather than static, drawing on anthropological research to define heritage as a process, an "act of passing on and receiving memories and knowledge" (p. 2). Language changes over time, with new terms being introduced and new meanings becoming associated with old terms (Shopland, 2020). Smith's conceptualization of heritage provides a valuable framework for analyzing the impact of this thesis on heritage in the form of archival metadata descriptions, further discussed in chapters 3 and 8.

Galleries, Libraries, Archives, and Museums, or GLAM, are record-keeping institutions that collect, manage, document, and provide access to cultural heritage (Blouin and Rosenberg, 2011; Jaffe, 2020; Schwartz and Cook, 2002; Thomassen, 2002; Welsh, 2016; Welsh and Batley, 2009). GLAM keep records in catalogs that were initially physical and handwritten but since the 1960s, with the creation of Machine Readable Cataloging (MARC) led by Henriette Avram at the Library of Congress (Library of Congress, 2017), have increasingly taken the form of digital databases containing hand-typed descriptions of heritage. The case study of this thesis takes place in an archival institution, that of the University of Edinburgh Heritage Collections. Among GLAM institutions, Archives contain the greatest diversity of materials, ranging from manuscripts to digital recordings to physical artifacts (Thomassen, 2002). In this thesis, the term *Archives* refers to the GLAM institution, rather than informal, personal collections of material that are not held in a GLAM institution (Caswell, 2016). Evidence of archival record-keeping extends back to 8000 BCE in Mesopotamia (present-day Iraq and Kuwait), preceding the invention of writing with the Cuneiform script developed from 3400 to 3100 BCE (ICA, 2021). Archives shape historical narratives and thus have close associations with nationalism, politics, and activism (Blouin and Rosenberg, 2011; Flinn and Alexander, 2015; Schwartz and Cook, 2002; Wood et al., 2014; Yale, 2015).

GLAM documentation refers to the catalog metadata descriptions of GLAM.

Catalog metadata descriptions of a single type of institution are referred to similarly, e.g. archival documentation. Catalogs organize descriptions according to particular schemas based on the type of GLAM institution. For example, archival metadata schemas (e.g. the General International Standard for Archival Description (ICA, 2011)) have deeper hierarchies than library metadata schemas (e.g. Dewey Decimal Classification (OCLC, 2023)). Resources for writing catalog metadata descriptions, on the other hand, overlap across institutions. For example, the Library of Congress publishes Subject Headings that libraries as well as other types of GLAM institutions use, including the University of Edinburgh Heritage Collections' Archives. Hundreds of cataloging resources exist to guide description practices, from controlled vocabularies published by national and international institutions and associations (e.g. the Spectrum collection management standard[4]) to reports of best practices developed through grassroots, public efforts (e.g. The Trans Metadata Collective's *Metadata Best Practices for Trans and Gender Diverse Resources* (2022)). Catalogs' metadata descriptions are written by people who have received cataloging training in universities, or volunteers who at minimum receive training at specific GLAM institutions. As such, the language of GLAM documentation is shaped by university degree programs, GLAM institutions, published cataloging resources, the personal experiences and knowledge of the cataloger who writes the descriptions, and the cultural heritage material itself.

Context in this thesis refers to the characteristics of a situation, including social, economic, temporal, political, cultural, linguistic, geographic, and historical considerations. This thesis writes about the contextual nature of data to describe the way in which these considerations influence the meanings ascribed to data and models. Gitelman (2013) writes of a "mythology" around data as being "decontextualized," which has been reflected in the rhetoric around ML systems built on data (Raji et al., 2021; Verdegem, 2021). In fact, data are simplified reflections of the world; reality is abstracted into data more than it is represented by data (Drucker, 2021; Gitelman and Jackson, 2013). I discuss ML data and models (ML systems) as *socio-technical* to indicate how ML systems are intertwined with the context in which society uses them (Jo

[4] collectionstrust.org.uk/spectrum

and Gebru, 2020; Thylstrup, 2022). The context in which ML systems are created often mismatch the diversity of contexts in which those systems are used. As Aragon et al. (2022) write, "context is recursively defined as information outside measurement" (p. 105), so qualitative research methods, such as interviews and workshops, are required to study ML in context, as socio-technical systems.

Biased language in this thesis refers to language that communicates social biases. More specifically, I define biased language as "written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity, and privileging other people through words or phrases that favor their identity" (Havens et al., 2020). Bias can appear in data of any form (e.g. audio, video, numeric, image-based). In this thesis, most of my discussion of bias revolves around language, given that the data for my case study is text. In ML literature, concepts of bias are closely related to concepts of fairness, ethics, and transparency (Kasirzadeh, 2022; Suresh and Guttag, 2021). While the framework of distributive justice is often applied in an attempt to mitigate bias through a fair allocation of resources, the systematic nature of biases and their resulting harms mean that a framework of structural injustice is also needed to address the systemic issues that cause social biases (Kasirzadeh, 2022). Drawing on feminist theories and critical discourse analysis, this thesis conceptualizes bias as inherent to data, language-based or otherwise, because practices of collecting, categorizing interpreting, and using data reflect societal power relations (Bucholtz, 2003; D'Ignazio and Klein, 2020; Hill Collins, 2000).

Biased language, along with other types of biased data, lead to biases in computational systems, such as ML models, that use that data (D'Ignazio and Klein, 2020; Friedman and Nissenbaum, 1996). Friedman and Nissenbaum (1996) define three types of biases in computational technologies: preexisting, technical, and emergent. *Preexisting bias* originates in personal perspectives and broader social structures that get engineered into computational technologies (e.g. Benjamin's "New Jim Code" (2019)). *Technical bias* refers to mismatches between the real world and representations of the world in computational technologies (e.g. the use of pronouns as a proxy

for gender identity (Spiel et al., 2019)). *Emergent bias* manifests during the use of computational technologies in the real world (e.g. an online query for "black girls" returning racist and sexist search results (Noble, 2018)). For ML systems, Suresh and Guttag (2021) define six types of biases: history, representation, measurement, evaluation, and aggregation. The first three biases occur during data curation and the last three biases occur during model creation. For NLP systems specifically, Hovy and Pabhumoye (2021) identify five sources of bias: data, the annotation process, input representations for models, the models, and research design. Biases are problematic because they unjustly cause harm to certain communities of people while privileging other communities of people. Chapter 2 further articulates my conceptualization of bias.

Gender in this thesis refers to a changeable identity characteristic, one that people determine for themselves rather than having it assigned to them (Keyes, 2018; Scheuerman, Spiel, et al., 2020). Gender is distinct from the grammatical gender of words in language, which may be feminine, masculine, non-binary, or neutral, and does not map one to one with gender identities (Spiel et al., 2019). In the United Kingdom (UK) and United States (US), the term *trans* is often employed as an umbrella for numerous gender identities such as genderqueer, transmasculine, and trans woman, among many others (Costanza-Chock, 2018; Scheuerman, Spiel, et al., 2020). Recognizing that gender terminology such as trans is culturally specific, this thesis employs the term *gender diverse* to refer to gender identities that do not fit within binary conceptualizations of gender, that differ from one's gender assigned at birth, and that cannot be described as trans, based on advice from the Trans Metadata Collective (Burns et al., 2022). Any gender, even a society's most powerful gender, can experience harms from gender biases (Hessel and Beard, 2022).

Chapter 2

Background

> Contrary to popular belief, high technology is often as socially regressive as it is technically revolutionary or progressive.
>
> –Mar Hicks, *Introduction: Britain's Computer "Revolution"* (2018, p. 17)

This chapter describes the contextual factors motivating my Ph.D. research. I begin with an introduction to social biases in Machine Learning (ML) and describe promising approaches for addressing them informed by literature on structural injustice (§2.1). Then, I explain the recalibration of ML demonstrated with my thesis, outlining problematic priorities in existing ML research and my proposed alternatives (§2.2).

2.1 Social Biases in Machine Learning

The challenge social biases pose for ML systems extend beyond the systems' underlying code. As human-made technologies, ML systems are subject to the perspectives of their creators, including their *values* (Birhane et al., 2022; Birhane et al., 2023; R. Dotan and Milli, 2020; Muntean et al., 2017) and *biases* (Bourgeois et al., 2018; S. Leavy, 2018; Markl, 2022b; C. Sweeney and Najafian, 2019). ML systems' creators are not a representative sample of the communities that interact with and are impacted by those systems (Havens

et al., 2020). Rather, the majority of ML systems' creators belong to the dominant social group of white, cisgender, heterosexual men living in Western countries (further discussed in §4.1.3). Consequently, when this dominant social group creates ML systems, they encode and engineer their own biases into datasets and models (Birhane, 2020; Friedman and Nissenbaum, 1996; Hicks, 2018; Scheuerman, Wade, et al., 2020). For example, Perez (2019) collated and summarized the numerous ways in which datasets encode sexism, and Benjamin (2019) has written of the racism, or "new Jim Code," that programmers "engineer" into technology. Friedman and Nissenbaum (1996) refer to these types of ML biases as "preexisting bias," because the biases existed in institutional and personal practices and attitudes before the technology's creation. Hicks (2018, 2021) explains how those in power build social biases into computational technology through an examination of computing history in the UK. The author describes how the exclusion of women from computing led to a labor shortage, which in turn led to the UK losing its technological lead as inventor of the computer. Social biases in ML cannot be addressed as "bugs" (Hicks, 2021) or "glitches" (Benjamin, 2019; Broussard, 2023) because these biases are structural (D'Ignazio and Klein, 2020; Hill Collins, 2000).

Structural issues are difficult to address due to their complexity. There are numerous actors involved with varying degrees of power, making it difficult to trace the impact of one person's action on another. As Young (2011) writes, "It is not difficult to identify persons who contribute to structural processes. On the whole, however, it is not possible to identify how the actions of one particular individual, or even one particular collective agent, such as a firm, has directly produced harm to other specific individuals" (p. 96). Adding ML systems into societal structures further complicates the identification of those responsible for the harms from social biases. Advances in ML technology have increased the computing power and data storage needed to create ML models (Bender et al., 2021; R. Dotan and Milli, 2020). These resource requirements limit access to the underlying data and architecture of ML systems (Bender et al., 2021), ensuring the corporations that can afford the resources maintain the greatest power over them (R. Dotan and Milli, 2020). As data-driven technologies, ML systems rely on the categorization of data, which is an inherently reductive, simplifying process (Bowker and Star, 1999; Drucker, 2021; Gitelman and Jackson, 2013). As a result, normative and

hegemonic categorizations that reflect the perspectives of the dominant social groups working at large corporations are built into those corporations' ML systems (Buolamwini and Gebru, 2018; Crawford and Paglen, 2019; R. Dotan and Milli, 2020; Lucy and Bamman, 2021; Noble, 2018; L. Sweeney, 2013).

Though ML systems have become more integrated into daily life, from guiding online search (Mehdi, 2023; Ng, 2023) to powering voice assistants (Abercrombie et al., 2021) to filtering job applicants' resumes (Rieke and Bogen, 2018), these systems' models function behind the scenes. Most people interact with ML systems through a Graphical User Interface (GUI) or voice interactions, both of which render the inner workings of the ML systems invisible. Consider a Google search: all a user sees is a search bar, and then a list of search results and advertisements. While GUIs and voice interactions may adhere to usability design principles, they also contribute to the invisibility of ML infrastructure. When these infrastructures are invisible, they are difficult to critique and easy to overlook (Adler, 2017; Bowker and Star, 1999; Drabinski, 2013). How can a harmful search result be traced back to the decision of a programmer, or team of programmers? Even if this were feasible, what would tracing that line achieve?

Adopting Young's social connection model of responsibility for ML (Kasirzadeh, 2022; Young, 2011), I aim to support collaborative approaches to mitigating harms from social biases, rather than looking for specific individuals or organizations to blame for creating biased datasets or ML models. Young (2011) characterizes the social connection model of responsibility as "forward-looking," emphasizing action to create a more just future. In this model, society looks to the past not to assign blame, but to understand the origins and manifestations of structural injustices. This understanding can then inform collaborative action to transform societal structures (ibid). Scholars in GLAM, the Humanities, and Digital Humanities describe the value of the past similarly, noting the transformative power in shaping historical narratives (Duff and Harris, 2002; Olson, 2001; Smith, 2006) and the risk of repeating historical harms if the past is not critically reflected upon (Hessel and Beard, 2022; Hicks, 2018, 2021; McGillivray et al., 2020; Risam, 2021). Additionally, the past provides evidence of diverse societal structures that expand our imagination of future possibilities and avoid overly deterministic views of progress. Graeber and Wengrow (2021) write of historical "bureaucracies that work on

a community scale; cities governed by neighbourhood councils; systems of government where women hold a preponderance of formal positions; [and] forms of land management based on care-taking rather than ownership and extraction" (p. 523). Though the existence of these societies has often been overlooked, their existence nonetheless demonstrates the possibility of a more equitable distribution of societal power relations.

To address social biases, the invisible structures reinforcing unjust power relations must be made visible. In the context of ML, this means understanding the data on which a model is trained, how the model categorizes data, and how the model is applied in society. This understanding empowers the public to identify the perspectives included and excluded in an ML system, thus enabling the anticipation of that system's limits and potential harms. As Adler (2017) points out, we can only make improvements to a system if we know how that system is structured. In a critique of knowledge organization in libraries, Adler states, "It is a credit to the institution of librarianship that these tools [classification systems] are open to the public and available for criticism" (p. 9). Drabinski (2013) makes a similar argument in *Queering the Catalog*, explaining how classification systems in libraries can, when visible, be teaching resources that present "knowledge production as a contested project" (p. 108). The key is that the organizing structures of data and models, whether for a GLAM catalog or otherwise, must be visible to the public.

My research takes inspiration from approaches to social biases that have roots in activism, seeking to make changes to technology *and* society. Despite the discourse of "revolution" that often accompanies accounts of computational innovation, computational technologies such as ML have a history of solidifying, rather than challenging, existing social hierarchies (R. Dotan and Milli, 2020; Hicks, 2021). Understanding how social biases manifest in existing ML systems will inform the creation of new ML systems that empower, rather than further oppress, minoritized communities (D'Ignazio and Klein, 2020; Friedman and Nissenbaum, 1996). As the oppressed people, minoritized communities provide the knowledge and experience necessary to identify manifestations of social biases (Young, 2011). Approaches such as value-sensitive design (Friedman and Nissenbaum, 1996), design justice (Costanza-Chock, 2018), data feminism (D'Ignazio and Klein, 2020), and participatory action research (Martin and Hanington, 2012b; Reason

and Bradbury-Huang, 2007; Reid and Frisby, 2008; Swantz, 2008) focus on collaboration with communities while also being action-oriented, aiming to make improvements within a particular community. My Bias-Aware Methodology in Chapter 4 outlines the interdisciplinary, collaborative approach I executed for this thesis to make social biases visible.

2.2 Recalibrating Machine Learning

In this thesis I propose and demonstrate a recalibration of ML to address ML systems' social biases. Although the ML community has encouraged research on bias, dedicating conferences and workshops to bias and related topics of fairness and ethics (e.g. *AAAI/ACM Conference on AI, Ethics and Society*,[1] *Workshop on Gender Bias for NLP*,[2] and *ACM Conference on Fairness, Accountability and Transparency*[3]), the priorities underlying most ML research and practice limits the efficacy of much ML bias research. ML approaches to social biases have focused on mathematical representations based on bias as an issue of *distributive* justice, or allocation of resources (Kasirzadeh, 2022). Identifying a gap in ML bias research on understanding social biases, especially the social structures through which they are enacted, I focused my Ph.D. research on creating ML models to identify types of social biases so their manifestations in data could be made visible. Through my literature reviews on social biases in GLAM and ML (summarized in Chapter 3, §5.1.3, and §6.2), I recognized four priorities of ML bias research inhibiting its efficacy. In the remainder of this chapter I describe these priorities and propose alternatives, shifting away from conceptualizing ML biases as a technical issue and towards addressing ML biases as *socio*-technical, rooted in societal structures that data and technology can either reinforce or subvert (Kasirzadeh, 2022).

Priority 1: Quantity. ML research and practice typically prioritizes quantity. Authors critiquing common ML practices note this emphasis on quantity in relation to data size (Birhane et al., 2022; Bommasani et al., 2021; Paullada et al., 2021; Welty et al., 2019), model architecture (Bender et al., 2021),

[1] aies-conference.com
[2] genderbiasnlp.talp.cat
[3] facctconference.org

and performance evaluations (Birhane et al., 2022; Welty et al., 2019). For example, in NLP, Kaplan et al. (2020) and Hoffman et al.'s (2022) comparisons of language models' performance focus unquestioningly on the number of model parameters and the size of model training data. Their publications report no consideration of additional factors that could impact an ML system's performance (i.e. the *representativeness* of a dataset for a model's use case, the *quality* of a dataset). Even the BigScience Workshop's BLOOM model (2022), meant to be an alternative to opaque, corporate, English-focused language models, focuses on quantities, noting the model's 176 billion parameters in its publication's title and 59 languages in its abstract. Such widespread emphasis on large quantities communicates an underlying assumption that bigger is better, yet this focus on quantity sacrifices quality (Bender et al., 2021; Welty et al., 2019). Investigations of training datasets (Kreutzer et al., 2022; Luccioni and Viviano, 2021; Perez, 2019; Sahoo et al., 2022), benchmark datasets (Birhane and Prabhu, 2021; Blodgett, Lopez, et al., 2021; Crawford and Paglen, 2019), and models (Buolamwini and Gebru, 2018; Garg et al., 2018; Jentzsch and Turan, 2022; Lucy and Bamman, 2021; Rudinger et al., 2018; Scheuerman et al., 2019; Scheuerman, Wade, et al., 2020) have uncovered poor quality data, ranging from inaccurate language classifications of text to racist, sexist, trans-exclusive, and sexually explicit text and images.

Recalibration 1: Quality. I propose ML research prioritize quality over quantity, more critically reflecting upon ML dataset creation in relation to the intended task of an ML system. Calls for "data-centric AI" (Brown, 2023) and data perspectivism (Basile, 2022) encourage more focus on the quality of data input to ML models. As Jo and Gebru (2020), Thylstrup et al. (2021), and Havens et al. (2020, 2022) suggest, the GLAM sector provides examples of data curation practices that offer alternatives to those common in ML.

Priority 2: Efficiency. The hype around the promise of ML technologies (Birhane et al., 2023; Raji et al., 2021; Verdegem, 2021), such as the profits they can bring companies and industries (T. Dotan, 2023; Hagey and Cherney, 2023; Kruppa, 2023), has created competition in ML research. This competition contributes to a prioritization of efficiency in the model creation process, where being the first to release a technology is more valued than releasing a thoroughly tested, highly accurate technology (Bommasani et al., 2021; Karen,

2023; Noonan, 2023; Seetharaman, 2023). As a result, in ML bias research concepts of bias and identity characteristics, such as gender, are often vaguely defined, overly simplified, or not defined at all (Blodgett et al., 2020; Devinney et al., 2022; Keyes, 2018; McCradden et al., 2020; Stańczak and Augenstein, 2021; see also §4.1.3, §4.1.4). As a result, dataset creation processes too often reduce subjective tasks to a single interpretation, or ground truth (Basile, 2022; Basile et al., 2021; Davani et al., 2022), and models are often presented as more advanced than they truly are (Bender et al., 2021; Raji et al., 2021). Looking to the GLAM sector, Olson (2001) states, "'better and quicker and cheaper' is always at a price, and the price is the violent reshaping of objects to fit the preconceptions of the knowing subject" (p. 663). Consider the use of proxies for demographic information: using a person's pronoun or name to determine their gender identity may be an efficient way to interpret data, but it is not accurate (Keyes, 2018; Scheuerman, Spiel, et al., 2020; Spiel et al., 2019), undermining the accuracy of any model built on such interpretations. Uncritical interpretations of data risk misgendering, stereotyping, omission, and other forms of oppression.

Recalibration 2: Accuracy. I propose a prioritization of accuracy over efficiency in the process of creating ML systems. Dedicating additional time to consider which metrics are appropriate for a model, what those metrics are capable of measuring, and developing alternative metrics when necessary would help create ML systems that are more accurate relative to real-world contexts (Raji et al., 2021; Welty et al., 2019). ML systems are socio-technical systems (Kasirzadeh, 2022), so their accuracy will change from one social context to another, and as societies evolve over time. Incorporating stakeholders of an ML system in its evaluation is thus an important, qualitative approach to measuring the system's accuracy (Goree and Crandall, 2023; Goree et al., 2023).

Priority 3: Convenience. The prioritization of efficiency and quantity, combined with the availability of large-scale data and crowdwork through online platforms, contributes to a prioritization of convenience in ML. Rather than dedicating time to curating datasets that will best serve the intended audience of a model (Jo and Gebru, 2020; Paullada et al., 2021; Rogers, 2021), many ML researchers and practitioners look for sources of easily-obtainable

data (Bommasani et al., 2021). Researchers at MIT used readily available video recordings of celebrities to create the Speech2Face model, which they report as being capable of generating a person's face based on their voice (Oh et al., 2019). They do not provide any discussion of whether the training data served as a representative sample of a large enough population to make this claim, nor do they include a consideration of the echoes of the historical, pseudoscientific field of eugenics in their model. The report on ChatGPT's language model, GPT-4, states: "Given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar" (OpenAI, 2023, p. 2). While competition as a reason for not detailing the model's creation is arguably problematic (Tahaei, Constantinides, Quercia, and Muller, 2023), the rationale at least holds weight. Safety as a reason, however, does not. In fact, discussion of the ethics of tracking people online, and of using text and images that people publish online without consent, are missing from many reports of new ML models (Crawford, 2021), including BERT (Devlin et al., 2019), Chinchilla (Hoffmann et al., 2022), Speech2Face (Oh et al., 2019), and GPT-4 (OpenAI, 2023).

Recalibration 3: Representativeness. I propose ML research prioritize representativeness over convenience in dataset curation. ML models are statistical at their foundation (Birhane et al., 2023; Jurafsky and Martin, 2023), so to perform as expected in real-world contexts, they require training data that serves as a representative sample of the models' population of stakeholders. Mismatches between training data and the global population have led to countless examples of systematic biases in ML, including sexist and racist search results (Noble, 2018) and facial recognition models (Buolamwini and Gebru, 2018). In a Digital Humanities project, Beelen et al. (2023, 2022) demonstrate an approach to estimating the representativeness of a dataset relative to one's research context. The authors compared the political views represented in a dataset of digitized 19th century British newspapers to the political views of all British newspaper publications that existed during that time. Conducting this "environmental scan" (Beelen et al., 2022) enabled researchers studying 19th century British history to adjust their analysis based on the dataset's biases, ensuring a more accurate understanding of the

political landscape of the time. Interdisciplinary approaches such as Beelen et al.'s (2023, 2022) are key to creating more representative ML systems (Blodgett et al., 2020; Bommasani et al., 2021; Crawford, 2017).

Priority 4: Universal Thinking. Underlying the prioritization of quantity, efficiency, and convenience is an assumption in the generalizability of technology (Birhane et al., 2022). Consequently, much ML research and practice prioritizes universal thinking, with authors publishing ML models as widely applicable (Blodgett et al., 2020; Raji et al., 2021). For example, Tu et al. (2023) discuss "generalist" AI as an aim but focus only on the generalizability of model architecture; the authors do not consider how those represented in the model training data or model evaluation process may limit the generalizability of the model. The way in which Google AdWords associate names common among African American communities with advertisements about arrest records more than names common among white communities (L. Sweeney, 2013), the anti-Semitic (among other harmful) comments of Microsoft's Tay chat bot (Hunt, 2016; Lee, 2016), and the misgendering and erasure of trans bodies that Keyes (2018) and Constanza-Chock and Philip (2018) have documented are only some of the many examples of ways in which ML models fail to exhibit the objectivity that is needed to make them universally applicable. ML, as with all technology, is value-laden (Birhane, 2020; Birhane et al., 2022; Birhane et al., 2023; R. Dotan and Milli, 2020).

Recalibration 4: Situated Thinking. I propose ML research prioritize situated thinking over universal thinking. By situated thinking, I refer to localized or "context-sensitive" (Kasirzadeh and Gabriel, 2023) approaches to creating and evaluating ML datasets and models, such as considerations of language (Ciora et al., 2021; Markl, 2022a; Nekoto et al., 2020), time (Beelen et al., 2021; Rodolfa et al., 2020), politics (Birhane, 2020; Coffey, 2021; Guo et al., 2020; Hicks, 2018, 2021), and history (Samorani et al., 2022), among other contextual factors. Feminist theories of knowledge as partial, multiplicitous, and subjective (Crenshaw, 1989, 1991; Haraway, 1988; Harding, 1995) inspire my prioritization of situated thinking over universal thinking. Only by focusing on the experiences of minoritized communities of people can ML systems be designed to empower, rather than oppress, those minoritized peoples

(Costanza-Chock, 2018; D'Ignazio and Klein, 2020; Kalluri, 2020).

The harmful consequences of socially biased ML systems that result from the prioritization of quantity, efficiency, convenience, and universal thinking motivated my recalibration of typical processes of dataset and model creation and evaluation. At the end of the chapters reporting my thesis' contributions (chapters 4-7), I explain how my approach to each contribution prioritizes *quality* over quantity, *accuracy* over efficiency, *representativeness* over convenience, and *situated thinking* over universal thinking.

Chapter 3

Literature Review

> It is a mad and dangerous wish, however, to break with the past entirely.
>
> –Iris Marion Young, *Responsibility for Justice* (2011, p. 172)

This chapter provides a summary of the literature relevant to my research questions, centering on bias research in the GLAM sector, ML, the Humanities and Digital Humanities, and Design in a Western context. To begin, I summarize literature from the GLAM sector, providing context on the origins and approaches to bias in this sector (§3.1). Next, I shift to a focus on ML research, explaining the origins of bias in ML and approaches to addressing them, noting limitations with many of these existing approaches (§3.2). I close with an explanation of the theoretical triangulation I use to guide my methodology, approach, analysis, and reflection in this thesis (§3.3). This literature review reveals two research gaps to which this thesis makes a contribution to knowledge: (1) ML systems for identifying types of social biases in language and (2) ML approaches for analyzing GLAM documentation for social biases at large scale.

Publication Note: This chapter is based on a shorter publication, *Confronting Gender Biases in Heritage Catalogs: A Natural Language Processing Approach to Revisiting Descriptive Metadata*, which will appear in the *Routledge Handbook on Heritage and Gender* (expected 2024). I wrote the publication as lead author

with my supervisors providing feedback to guide my revisions.

3.1 Framing Bias in GLAM

GLAM acquire, describe, and manage cultural heritage with the aim of recording historical events, people, and places to inform historical narratives, understandings of the present, and possibilities for the future (Blouin and Rosenberg, 2011; Jaffe, 2020; Schwartz and Cook, 2002; Thomassen, 2002; Welsh, 2016; Welsh and Batley, 2009). GLAM documentation is written and organized into metadata fields according to cataloging resources such as classification schemes, metadata standards, and controlled vocabularies (Angel and Fuchs, 2018; Thomassen, 2002; Welsh, 2016; Welsh and Batley, 2009). Together the classification structure and descriptions of cultural heritage items enable GLAM visitors to discover the items in GLAM catalogs (Duff and Harris, 2002; Jaffe, 2020). That being said, records of cultural heritage are inevitably shaped by institutional priorities, cataloging resources, catalogers' education and training, and catalogers' personal experiences, in addition to the content of the collections. As a result, biases permeate GLAM collecting practices, including appraisal processes and the actual items collected (Cook, 2011; Odumosu, 2020; Yale, 2015), classification and metadata structures (Adler, 2016, 2017; Bowker and Star, 1999; Drabinski, 2013; Furner, 2007), and documentation (Duff and Harris, 2002; Salway and Baker, 2020; Schwartz and Cook, 2002). These biases take the form of misrepresentations and omissions that, particularly for national GLAM institutions, aim to assert and reinforce particular national identities, cultural narratives, and social hierarchies (Smith, 2006; Stoler, 2002; Yale, 2015).

Misrepresentations and omissions reinforce the minoritization of some people and the dominance of other people. Misrepresentations include inaccurate and stereotypical representations of certain communities of people. Inaccurate representations in GLAM documentation often require revisiting collection material alongside its documentation, which is out of scope for this thesis. Stereotypical representations occur along intersecting axes of identity characteristics (Crenshaw, 1989, 1991) such as gender (Olson, 2001), sexuality (Adler, 2017), racialized ethnicity (Furner, 2007), and culture (Diao and Cao, 2016), among others. Omissions, meaning silences or absences, occur

when data and information about certain people has not been collected or has been lost, either purposefully or accidentally. Omissions occur when people's identities are not recorded, such as women who are described only with the title "Mrs." and their husband's name (Geraci, 2019); when provenance information has not been documented, such as the contributions of enslaved people (Ortolja-Baird and Nyhan, 2022); when certain types of heritage are not collected and thus are absent from cultural heritage records, such as the intangible heritage of First Nations communities in Australia (Smith, 2006); and when records of certain communities are unjustly discounted or excluded, such as the work of archaeologist Gimbutas on the equitable relations between women and men in societies of the Balkans and Eastern Mediterranean circa 7000 BC to 3500 BC (Graeber and Wengrow, 2021). Omissions contribute to the perpetuation of misrepresentations, because alternatives to the stereotypical narratives are missing from cultural and historical records (Beard, 2017; Graeber and Wengrow, 2021). In this thesis I aim to identify gender biased language in the form of omissions and stereotypes (chapters 5-6).

Given the history of misrepresentations and omissions in GLAM, it is no surprise that minoritized people have created community-run GLAM and GLAM resources (Flinn et al., 2009). Community-focused GLAM institutions that reject what Smith (2006) terms the "authorized heritage discourse" dominating cultural heritage narratives (described further in §3.3.3) include the Glasgow Women's Library[1] (which also has Archives and Museum collections) in Scotland, the Black Cultural Archives[2] in England, the Digital Transgender Archives[3] and Gerber/Hart Archives (for the LGBTQ+ community)[4] in the US, and the Native Museum of Mashteuiatsh (a First Nations community)[5] in Canada. GLAM cataloging resources developed for greater accuracy and inclusion of minoritized communities include the *Metadata Best Practices for Trans and Gender Diverse Resources* (Burns et al., 2022), the linked data vocabulary *Homosaurus*,[6] the *European Women's Thesaurus* (Drenthe and van der Sommen, 1998), and the Traditional Knowledge and Biocultural

[1] womenslibrary.org.uk
[2] blackculturalarchives.org
[3] www.digitaltransgenderarchive.net
[4] gerberhart.org
[5] collection.cultureilnu.ca
[6] homosaurus.org

labels.[7] Additional GLAM resources developed with the values of minoritized communities in mind include the Mukurtu Content Management System[8] and Archive of Our Own,[9] both developed with the aim of empowering communities to manage and provide access to their cultural heritage themselves. These institutions and resources recognize the need for localized, or *situated*, classification schemes, cataloging standards, and controlled vocabularies.

The gap in this existing work that I address relates to biases in the descriptive language of GLAM documentation; most work has focused on critiquing and changing the terminology of metadata fields and the structure of classification schemes. I posit that the lack of work on analyzing GLAM documentation results from three factors. First, the skills and resources needed to analyze large text corpora differ from the skills and resources with which GLAM experts are typically equipped (McGillivray et al., 2020; Terras et al., 2018). While the digitization of cultural heritage content has enabled greater collaboration between GLAM and computational experts (Padilla, 2017, 2019; Terras, 2015), the majority of these collaborations work with "collections as data" (Padilla, 2017), meaning digitized collection material (e.g. Ames and Havens, 2022; Beelen et al., 2021; Coll Ardanuy et al., 2020; Filgueira et al., 2021; Filgueira et al., 2019; Hinrichs et al., 2015; Hosseini et al., 2022; Lamb et al., 2022), rather than collections' documentation (e.g. Baker and Salway, 2020; Salway and Baker, 2020).

Second, GLAM continuously acquire new heritage items, meaning catalogers have an overwhelming amount of new items to describe so visitors can discover them (Blouin and Rosenberg, 2011). The Library of Congress, for instance, adds 10,000 new items to its collection each working day,[10] and in total, the British Library's catalog documents over 170 million items.[11] Consequently, revising GLAM documentation historically has not been prioritized. Third, as Welsh (2016) writes, "Perhaps we are so used to the fundamental concept that the catalog record is a surrogate for the material itself and to the argument that what users want is full-text access that it becomes easy for us to overlook the status of the catalog itself as data" (p. 327). Nonetheless,

[7] localcontexts.org
[8] mukurtu.org
[9] archiveofourown.org
[10] www.loc.gov/about/fascinating-facts
[11] www.bl.uk/about-us

especially following the *#metoo* and *Black Lives Matter* movements in the early 2000s, GLAM are increasingly dedicating resources to reviewing existing documentation (Antracoli et al., 2019; Berry, 2020; Collections Trust, 2023; Wetli, 2019).

3.1.1 Origins of Bias in GLAM Documentation

Why do we need research on GLAM documentation, in addition to GLAM collections, classification schemes, metadata standards, and controlled vocabularies? The little existing computational research on biased language in GLAM documentation indicates that the biased language of catalogs' metadata standards and classification structures also exists in the catalogs' descriptive language (Baker and Salway, 2020; Geraci, 2019; Salway and Baker, 2020). For example, Baker and Salway (2020) and Salway and Baker (2020) analyzed the influence of one cataloger, Mary Dorothy George, from the British Museum's catalog in the UK to the Lewis Walpole Library's catalog in the US, finding traces of George's historical perspective in harmful descriptions of people in these contemporary catalogs. Moreover, digitized collection material provides a biased sample of an entire GLAM institution's collections (Beelen et al., 2022; Hauswedell et al., 2020). Studying GLAM collections' documentation offers an approach to understanding GLAM collections more comprehensively.

GLAM documentation provides records of particular perspectives throughout history, not only providing access to cultural heritage, but also producing knowledge and shaping the interpretation of cultural heritage (Blouin and Rosenberg, 2011; Duff and Harris, 2002; Schwartz and Cook, 2002; Welsh, 2016). Studying GLAM documentation provides insight on how particular cultural narratives have been recorded and perpetuated, and how those narratives contribute to the uneven distribution of societal power relationships today. Reflecting on how past events and attitudes have shaped the present enables the identification of which perspectives have been misrepresented and omitted from cultural and historical records, a necessity in working towards a more equitable and just society (Flanagan and Jakobsson, 2023).

Historically, the GLAM community trained catalogers to classify and describe cultural heritage objectively. This approach originates in the Enlightenment's

positivist philosophy, from which the Archival and Library Sciences emerged, along with methods designed to record history neutrally (Duff and Harris, 2002). That being said, due to how minoritized communities of people have been classified and described, heritage of particular relevance to these communities has been rendered difficult to discover (Adler, 2017; Noble, 2018; Olson, 2001). Thanks to the way in which GLAM make cataloging resources publicly available, however, people have long been able to critique the practices of GLAM institutions for the biases they perpetuate and amplify.

3.1.2 New Directions

From the late 20th and early 21st centuries, catalogers, librarians, archivists, and curators began pushing back against the supposed neutrality and objectivity of heritage collections. Rejecting the Enlightenment's positivist philosophy, archivists and librarians began putting forth a postmodern philosophy (Duff and Harris, 2002). Postmodernism asserts that no record of history can be neutral (Tai, 2021). Duff and Harris (2002) write, "Every representation, every model of description, is biased because it reflects a particular world-view and is constructed to meet specific purposes" (p. 275). This postmodern view aligns with feminist theories' view of knowledge as *situated*, meaning knowledge changes depending on the context in which it is produced and received (Crenshaw, 1989, 1991; Haraway, 1988; Harding, 1995; Hill Collins, 2000). Postmodern, feminist views of cultural heritage shift the role of GLAM institutions and encourage new approaches to GLAM processes.

Scholars have put forth new approaches for the collection and description of heritage that aim to empower minoritized communities. Caswell's (2022) "feminist standpoint appraisal" prioritizes the perspectives of minoritized people, recognizing that due to their minoritized status, they offer unique points of view that people in dominant positions of society cannot offer. Feminist standpoint appraisal is thus about more than inclusivity; it is about enriching the knowledge records of GLAM institutions to improve research and scholarship (ibid.). For appraisal and other collection practices to evolve through this enrichment, GLAM institutions must reframe their relationship with the people they represent and serve. Iacovino (2010), Caswell and Cifor (2016), Tai (2021), and Caswell (2022), among other scholars, call for

participatory, human-centered, and community-oriented relationships where minoritized people are collaborators and agents, rather than subjects and visitors. Voluntary crowdsourcing and gaming approaches provide examples of how GLAM can develop such relationships at large scale (Flanagan and Carini, 2012; Flanagan et al., 2014; Manzo et al., 2015; Ridge, 2013, 2016).

For individuals working in GLAM institutions, scholars have proposed a reframing of the work of catalogers, librarians, archivists, and curators. Rather than viewing oneself as an expert, or as being responsible for cultural competence, these individuals can frame themselves as stewards (Tai, 2021) and caregivers (Caswell and Cifor, 2016). Cook (2011) and Tai (2021) propose that individuals working in GLAM adopt a framework of cultural humility. Cultural humility acknowledges that attaining cultural competence is an ongoing process and that biases are inevitable. This framework aims to "normalize not knowing" (Tai, 2021, p. 3) and adjust power distributions in GLAM, positioning minoritized communities as collaborators. Havens (2021) uses a Speculative Design approach to envision what this could look like for a GLAM institution's online interface to its catalog (Chapter 8).

When collaborating with communities to add to or revise collections and their descriptions, GLAM should document the changes to records in their catalogs (Drabinski, 2013; Duff and Harris, 2002). Such documentation makes the biases characterizing a collection or description explicit (Duff and Harris, 2002) and reminds GLAM's visitors of the constructed, subjective nature of collection, classification, and description (Drabinski, 2013). Changing collections and their descriptions to reflect the perspectives of particular communities empowers those communities, enabling them to control their identity and how they are perceived in political contexts (Smith, 2006). As Duff and Harris (2002, p. 272) write,

> *The power to describe is the power to make and remake records and to determine how they will be used and remade in the future. Each story we tell about our records, each description we compile, changes the meaning of the records and recreates them.*

The origins of bias and new directions for addressing it in GLAM have parallels in ML. In the next section, I detail the challenges with, origins of, and approaches to bias in ML.

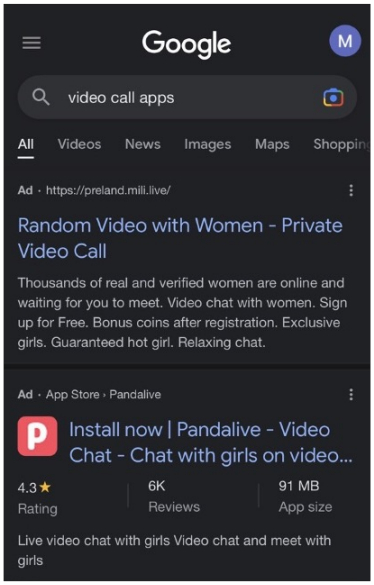


Figure 3.1: **Biased search results.** A friend's query for "video call apps" with Google's search engine in October 2022 yielded results for online video applications to chat with girls and women. The top result promises, "Guaranteed hot girl."

3.2 Bias in ML Systems

Awareness of biases in ML has grown as harmful consequences of ML models' applications become evident. Sweeney (2013) wrote of the criminalization of Black people in the ML model behind Google AdWords. O'Neil (2016) wrote of qualified, respected teachers who lost their jobs due to uncritical applications of ML models to teachers' performance evaluations. Constanza-Chock and Philip (2018) wrote of the "anomolous" categorization of trans bodies in TSA scanners at the airport. Noble (2018) wrote of the sexualization of Black girls perpetuated through the ML model powering Google's search engine; four years later, my friend's Google search confirmed that this issue of the sexualization of girls and women had not been adequately addressed (Figure 3.1). These examples exhibit the two types of harms that may result from biases in ML models: representational and allocative harms (Crawford, 2017). *Representational harms* result in negative consequences for a person due to their identity, such as the lack of recognition of trans bodies and the sexualization of women. Representational harms often lead to *allocative harms*, which result in the denial of a resource or opportunity, such as teachers losing their jobs due

to an unfair evaluation process enacted in an ML model.

Awareness of the harms resulting from biases encoded in ML systems have led to growing interest in bias and related areas of research in ML. Venues such as the *ACM Conference on Fairness, Accountability and Transparency*, the *Workshop on Gender Bias in NLP*, and the *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, among others, encourage research on bias, fairness, ethics, human values, and explainability. Publications address these topics in research for ML generally (Birhane et al., 2022; Holstein et al., 2019; Suresh and Guttag, 2021), as well as for the more specific areas of Artificial Intelligence (Crawford, 2021), Computer Vision (Bennett and Keyes, 2020; Birhane and Prabhu, 2021), Natural Language Generation and Conversational Agents (Abercrombie et al., 2021; Kasirzadeh and Gabriel, 2023), Search and Information Extraction (Noble, 2018; Shah and Bender, 2022; Zheng et al., 2017), Automatic Speech Recognition (Markl, 2022a), and NLP (Bender et al., 2021; Hovy and Prabhumoye, 2021; Tenney et al., 2020). Research in these areas includes investigation of the biases in existing ML models (Buolamwini and Gebru, 2018; Jentzsch and Turan, 2022) and their datasets (Crawford and Paglen, 2019; Luccioni and Viviano, 2021), the harms and risk of harms from using those models (C. Bird et al., 2023; Weidinger et al., 2022), and approaches to mitigating those harms (Bordia and Bowman, 2019; Zhao et al., 2017, 2018). Scholars have also suggested changes to steps within the ML pipeline, such as dataset creation (Basile et al., 2021; Davani et al., 2022; Jo and Gebru, 2020; Paullada et al., 2021; Rogers, 2021) and documentation (Bender and Friedman, 2018; Gebru et al., 2018; Mitchell et al., 2019). Progress on mitigating biases and its resulting harms in ML remains limited, though.

Overly simplistic conceptualizations of bias and inadequate evaluation approaches have limited the efficacy of ML approaches to bias. For example, in Computer Vision research, Buolamwini and Gebru (2018) and Keyes (2018) report on how facial recognition models have been developed without adequately diverse data and identity categorizations, resulting in performance differences across genders and racialized ethnicities. In NLP, Blodgett et al.'s (2020) survey on bias, and Stańczak and Augenstein (2021) and Devinney et al.'s (2022) surveys on gender bias note authors' frequent reduction of gender to a binary and the lack of clarity in authors' definitions of bias.

Regarding measurement approaches, Welty, Praveen, and Aroyo (2019); Paullada et al. (2021); Raji et al. (2021); and Jacobs and Wallach (2021) describe shortcomings in dataset and model evaluations, calling for more critical approaches to metrics and performance reporting to avoid overstating ML systems' capabilities and generalizability. As Irani (2016) writes, culture and language evolve as time passes, and computers do not have the "cultural fluencies" needed to interpret much of the data upon which ML systems rely.

Despite many technological advances, manual annotation and other forms of human labor are very much needed for data-driven technologies to function (Crawford, 2017; Irani, 2015, 2016; Taylor, 2018). For example, Samorani et al.'s (2022) approach to creating an ML model to automate medical appointment scheduling required human intervention to avoid perpetuating racial discrimination. When defining the task for the ML model, the authors considered the history of racism in the US that has led to a correlation between a patient's racialized ethnicity and likelihood to miss a medical appointment. Acknowledging oppressive societal structures as the cause of this correlation, the authors were able to define the model's task in a way that avoided perpetuating racism in the appointment scheduling system, focusing on minimizing the maximum wait time among patients rather than using patients' demographic information. In this thesis, I focus on identifying biases through context-informed approaches, aiming to understand the variety of ways social biases may manifest in language to develop a more complex conceptualization of bias in ML systems and GLAM documentation. Although this added complexity presents challenges, it contributes to a more *accurate* understanding of ML systems as socio-technical, enabling improvements to evaluation approaches for these systems. I draw on interdisciplinary literature and employ quantitative and qualitative methods to identify, measure, and reflect upon gender biased language and its contextual nature.

3.2.1 Origins of Biases in ML

Drawing on conceptualizations of the origins of bias in computer and ML technologies (Crawford, 2017; Friedman and Nissenbaum, 1996; Hovy and Prabhumoye, 2021; Suresh and Guttag, 2021), I identify three overarching sources of bias in ML: data, models, and human decision-making. By *data*, I

refer to limitations of data for representing reality and the difficulty of creating datasets of balanced samples. By *models*, I refer to the way in which models approach a task, interpret data, and perform against particular metrics. By *human decision-making*, I refer to processes of data collection, curation, and interpretation; and processes of defining tasks for, training, and evaluating models.

Though data are always summaries, partial representations of a population or phenomenon (Bowker, 2008; D'Ignazio and Klein, 2020; Gitelman and Jackson, 2013), there is a dangerous consensus that *large quantities* of data can elicit the truth and, when combined with an algorithm, create models that accurately predict the future (Bender et al., 2021; Paullada et al., 2021; Thatcher et al., 2016). Data, whether structured, such as in a tabular format, or unstructured, such as in paragraphs of a book or news article, are situated, incomplete representations of reality (Bowker, 2008; Drucker, 2021; Gitelman and Jackson, 2013). The etymology of "data" brings one back to Latin, with "data" as the plural of "datum," meaning "that which is given" (OED, n.d.). Today, however, I argue that what are referred to as data are more often gathered than given. Problematically, when data are gathered for ML datasets, the gatherers implicitly assume the data are representative of their ML models' audience. There is no standard approach to evaluate the *representativeness* of ML datasets.

Due to insufficient interdisciplinary collaboration, the ML community has repeated mistakes with data collection that are well-documented and reflected upon in GLAM (Blodgett et al., 2020; Crawford, 2017; McGillivray et al., 2020). ML datasets have been found to contain gender biases (Crawford and Paglen, 2019; Hube, 2017), racial biases (Buolamwini and Gebru, 2018), sexually explicit content (Birhane and Prabhu, 2021; Luccioni and Viviano, 2021), and derogatory language (Crawford and Paglen, 2019; Luccioni and Viviano, 2021), among other problematic contents. Meanwhile, scholars from GLAM have long reflected on the power exerted and oppression caused through data collection practices (Cook, 2011; Flinn et al., 2009; V. Harris, 2002; Odumosu, 2020; Schwartz and Cook, 2002; Stoler, 2002). The *convenience* of gathering data at scale online has outweighed considerations of *quality* as well as *representativeness*, including the ethics of using data without the data creators' consent.

The ML community has also repeated mistakes with model categorizations of data that are well-documented and reflected upon in GLAM. ML models label and organize data by particular attributes to search for meaningful patterns (Eisenstein, 2018; Jurafsky and Martin, 2023). The attributes models choose to label and organize data are similar to metadata standards and classification schemes that label and organize GLAM collections: both are necessary yet overly simplistic (Bowker and Star, 1999; Gitelman and Jackson, 2013; Jo and Gebru, 2020; Thylstrup, 2022; Thylstrup et al., 2021). Categorizing data and information allows us to make sense of the world and our relationships within it, but this categorization also does away with the messiness of the world, confining us to well-delineated categories (Bowker and Star, 1999). These categories often reflect and reinforce social biases (Adler, 2017; Furner, 2007; Olson, 2001). Computational approaches such as counterfactual modelling (Lu et al., 2020) aim to add data to models that represent more diverse groups of people than the gathered data contain. With these approaches, though, it is up to the researchers to determine who is omitted or misrepresented in their data. Similarly, evaluations of models depend on human-chosen metrics. Benchmarks and metrics for evaluating model performance are often inadequately grounded in real-world applications, so the performance measures fail to capture shortcomings with models' categorizations (Blodgett, Lopez, et al., 2021; Denton et al., 2020; Orgad et al., 2022; Raji et al., 2021; Welty et al., 2019).

This leads to the last, and arguably most important, source of bias in ML: human decision-making. Choosing which data to gather and how to train, develop, and evaluate models are decisions humans make. Humans are inevitably biased by their own experiences of the world (Friedman and Nissenbaum, 1996; Haraway, 1988; Harding, 1995; Young, 2011), so while ML models may serve a diverse and global audience (e.g. Google's search engine, Open AI's ChatGPT), the creators of the models are limited by their unavoidable *situatedness* in a particular context (Aroyo and Welty, 2015; Haraway, 1988; Harding, 1995). Additionally, creators of models are situated within the discipline of ML, which leads them to make different assumptions than people situated in other disciplines such as GLAM and the Humanities. Posner (2016) cautions, "...most of the data and data models we [Digital Humanists] have inherited deal with structures of power, like gender and race,

with a crudeness that would never pass muster in a peer-reviewed Humanities publication," echoing the critiques of approaches to bias in ML summarized earlier (Blodgett et al., 2020; Devinney et al., 2022; Raji et al., 2021; Stańczak and Augenstein, 2021). For example, in NLP gender bias research, whether or not a dataset includes labels indicating that a person's gender cannot be determined from the text in the dataset, model creators have trained their models to guess a gender (Dinan, Fan, Wu, et al., 2020; Webster et al., 2018). These interpretive decisions build social biases into ML models (Benjamin, 2019; Broussard, 2023; Hicks, 2021).

3.2.2 New Directions

Within ML, researchers and practitioners have begun advocating for new, more critical approaches to dataset and model creation and evaluation. These directions incorporate interdisciplinary engagement, localized framing of research, and stakeholder collaboration. Scholars in ML and Data Science have looked to feminist theories, the Social Sciences, and GLAM. In *Data Feminism*, D'Ignazio and Klein (2020) encourage critical reflection on the social context of data, helping dataset and model creators to recognize the situated nature of their work and reject a universal truth, whether this truth be a dataset's representation or a model's categorization of people. Publications focused on documenting and understanding large datasets have begun to turn toward GLAM and Archives for guidance (Jo and Gebru, 2020; Thylstrup et al., 2021). In this thesis, I also look to approaches from the disciplines of Design (Gaver et al., 2003) and Data Visualization (Hinrichs et al., 2019; Wexler et al., 2019; Whitelaw, 2015), to work with the uncertainty that exists in datasets but is often overlooked, as well as the Digital Humanities (Beelen et al., 2023; Beelen et al., 2022; Beelen et al., 2021; Kizhner et al., 2021), to reflect upon the generalizability of my research on gender bias in NLP and Archives relative to research on bias in ML and GLAM.

Promising directions for minimizing harms from bias in ML include *situated* approaches to model creation. Ciora et al. (2021) investigate existing Machine Translation models' biases in a Turkish context. The authors write, "We advocate for the inclusion of language-specific differences and the design of mitigation models that are linguistically and socially grounded" (p. 55). Keyes

et al. (2021) investigate ML models' ability to perform specific tasks within the context of case studies on autism and sexuality. Markl et al. (2022a) investigate commercial Automatic Speech Recognition models for linguistic bias in the context of the British Isles. Looking to GLAM and the Digital Humanities, I consider the ways in which cataloging and research projects are framed in a specific context, due to the contextual nature of knowledge (Dunsire and Willer, 2014) and the differences between historical datasets and the contemporary data on which most ML models are trained (Beelen et al., 2021; Filgueira et al., 2019; Hosseini et al., 2022). The work of my thesis prioritizes *situated* thinking end to end, presenting my work as a case study from problem definition (Chapter 4) through model evaluation (Chapter 7).

Incorporating human-centered research methods to collaborate with ML models' stakeholders offers another promising direction for minimizing harms from bias in ML (Aragon et al., 2022; Blodgett, Madaio, et al., 2021; Caselli et al., 2021; Tahaei, Constantinides, Quercia, Kennedy, et al., 2023). Rodolfa et al. (2020) collaborated with the Recidivism Reduction and Drug Diversion unit of the Los Angeles Police Department in the US to create an ML model, along with a framework for choosing a suitable fairness metric for one's social context. The authors describe their research as a "work-in-progress," recalling Tai's (2021) concept of "cultural humility" for GLAM, and explain that in order to ensure the model evolves with changing legal and social contexts, "an effective implementation will require ongoing evaluation of both the performance and fairness of the model's predictions over time" (Rodolfa et al., 2020, p. 155). Nekoto et al. (2020) report on collaboration with 400 participants across 20 countries for the Masakhane project, which aims to "strengthen and spur NLP research in African languages, for Africans, by Africans."[12] Aragon et al. (2022) summarize a range of human-centered methods and their value to data-driven work. In this thesis, I incorporate human-centered research methods through Participatory Action Research (Aragon et al., 2022; Martin and Hanington, 2012b; Reid and Frisby, 2008; Swantz, 2008), informing my problem definition, dataset creation, and model creation processes (Chapter 4).

In regards to data annotation in NLP more specifically, scholars have begun advocating for *data perspectivism* (Basile, 2022). Data perspectivism

[12] masakhane.io

recognizes that a single "ground truth" or "gold standard" may not be a suitable goal to achieve in a data annotation process undertaken to create an NLP model using supervised learning, due to the subjectivity of language (Basile et al., 2021; Davani et al., 2022. To allow for NLP research that incorporates multiple perspectives, data perspectivists advocate for the publication of disaggregated datasets, and for creating models based on multiple annotators' annotations even if they conflict (Basile, 2022). Data perspectivism informs my approach to creating an aggregated dataset that incorporates multiple annotators' perspectives, and my decision to publish each individual annotator's dataset alongside the aggregated dataset (Chapter 5). Additionally, data perspectivism's aim of encoding multiple annotators' viewpoints in a dataset aligns with feminist theories' conceptualization of knowledge as multiplicitous and subjective, discussed below in §3.3.

3.3 Theoretical Triangulation

For this thesis, I analyze and reflect upon my research of gender biases in archival documentation with NLP models through three theoretical lenses: (1) feminism, (2) critical discourse analysis, and (3) heritage as a process. These theories have distinct roots yet complement one another in their approach to knowledge and interpretation, highlighting parallels between GLAM and ML practices (Kushner and Morrow, 2003; P. Leavy, 2017a).

3.3.1 Feminism

I draw on a combination of feminist theories in my approach to research and analysis of my research outputs, focusing on theories from Western, English-speaking authors due to the context of my research in the UK. Feminism has roots in activism, with an aim of enacting societal change (Pilcher and Whelehan, 2004), so participatory approaches often draw on feminist theorizing (Moore, 2018; Reid and Frisby, 2008). Though many feminist theories have been put forth, at the foundation of all is the subjectivity and multiplicity of knowledge, and the rejection of a universal, neutral truth or perspective (Haraway, 1988; Harding, 1995; Pilcher and Whelehan, 2004). Data and information can only be understood and put to use

through interpretation, and that interpretation will always be situated, partial, and subjective (Haraway, 1988). Moreover, Moore (2018) writes that the commonalities across feminist and other critical studies are "a far-reaching critique of the practices of Western science and philosophy which produce inequalities and marginalities" (p. 11). Feminism itself has produced marginalities that have motivated the introduction of new types of feminist theories.

In response to the failure of earlier feminist theories to address the experiences of Black women, scholars put forth Black Feminism (Combahee River Collective, 1979) which considers gender alongside racialized ethnicity and other identity characteristics. Crenshaw (1989, 1991) coined the term "intersectionality," which refers to the way in which the intersection, or combination, of an individual's identity characteristics determine that individual's experience of privilege and oppression. Hill Collins (2000) put forth the "matrix of domination" to explain and study the way in which societies organize power and oppression. The matrix of domination consists of four domains:

- "Structural," where oppression is organized (e.g. national courts),
- "Disciplinary," where oppression is managed (e.g. bureaucracies),
- "Hegemonic," where oppression is justified (e.g. ideologies), and
- "Interpersonal," where oppression is enacted and experienced (e.g. social interactions between people).

D'Ignazio and Klein (2020) discuss the relevance of the matrix of domination, intersectionality and other feminist theories for Data Science and data-driven work such as ML. The authors introduce "data feminism" as an approach to working with data with that aims to change society's imbalanced distribution of power. Data feminism has seven principles that encourage a focus on justice, empowerment, reflective practice, and situated thinking (ibid., p. 17-18):

1. Examine power,
2. Challenge power,
3. Elevate emotion and embodiment,
4. Rethink binaries and hierarchies,

5. Embrace pluralism,
6. Consider context, and
7. Make labor visible.

These principles inform my research methodology (Chapter 4) and approach to creating my thesis contributions (chapters 5-7).

3.3.2 Critical Discourse Analysis

Critical Discourse Analysis (CDA) is a branch of linguistics that considers language in its context of use (Bucholtz, 2003; Fairclough, 2003; Gee and Handford, 2014; Marston, 2000; Smith, 2006; Talbot, 2003). *Discourse* refers not only to the words of written or spoken language, but also to the social relationships and practices in which language is produced and received (Fairclough, 2003). *Critical* theories (e.g. critical race theory, queer theory, feminist theories) grew out of interdisciplinary research and social justice movements; they focus on local contexts and an ethics of care by considering unequal distributions of power, agency and negotiation, and social and cultural reproduction (Kushner and Morrow, 2003; P. Leavy, 2017b). CDA analyzes meanings and actions in and surrounding language, whether spoken or written (Gee and Handford, 2014). According to CDA, to understand the meaning of language, two components must be considered: the internal relations of language, which are based on individual words and how they come together, and the external relations of language, which are based on contextual factors such as culture and politics (Fairclough, 2003).

CDA's conceptualization of language aligns with the situated nature of knowledge that feminist theories put forth. CDA offers a useful lens for critically reflecting on NLP models in relation to the people who are represented in, use, and are impacted by the models. The challenge with NLP, and all ML, models, however, is that while model developers may have an intended use case for their work, they cannot predict every possible future use case.

In GLAM, applications of NLP are further complicated by diachronic changes in language. Language evolves over time as new terminology is introduced, meanings of existing terminology change, and certain terminology falls out of use (Garg et al., 2018; Schulz, 2000; Shopland, 2020). Consequently, when applying NLP models to a context that differs from the context of their training

data, there will be a mismatch of perspectives. Though Digital Humanities researchers have begun experimenting with fine-tuning and re-training models from scratch on new data to customize a model for particular historical contexts (Beelen et al., 2021; De Toni et al., 2022; Manjavacas and Fonteyn, 2022), further work in this area is needed to understand the extent to which a model's contemporary foundation impacts its performance on historical data.

3.3.3 Heritage as a Process

Smith (2006) extends the definition of heritage from tangible objects of innate value to a process, the process of adapting past understandings in response to a present-day political, cultural, or social contexts. In conceptualizing heritage, the author draws on CDA, describing heritage as a discourse, or social practice. Smith (2006, p. 2) writes of heritage as,

> *the act of passing on and receiving memories and knowledge. It also occurs in the way that we then use, reshape, and recreate those memories and knowledge, to help us make sense of and understand not only who we 'are,' but also who we want to be.*

This description recalls the earlier quote from Duff and Harris (2002) on the process of describing an archival item as an act of "recreation," and as an act that has the power to influence how the archival item will be used in the future. In this way, heritage is a *process*: an experience and an act of storytelling, where values and identities are debated, regulated, and validated (Duff and Harris, 2002; Smith, 2006). Smith distinguishes between an "authorized" heritage discourse and a "dissenting" heritage discourse. A dissenting heritage discourse can challenge values embedded in an authorized heritage discourse, as Smith has observed in her ethnographic research with First Nations communities in Australia. Through a dissenting heritage discourse, communities have reclaimed control over their identities, particularly how they are perceived by national governments and GLAM institutions (Smith, 2006). Heritage is thus relational and dynamic, rather than static and tangible. Smith (2006) explains that communities can engage in the process of heritage to push back against their misrepresentation, or lack of representation (i.e. omission), in the authorized heritage discourse, regaining power to influence historical narratives and present-day politics.

Smith's theory of heritage as a process aligns with feminist theorizing and its activist origins aiming to enact social change. Heritage functions largely in the hegemonic domain of the matrix of domination (Hill Collins, 2000), with GLAM institutions residing in the disciplinary domain and being influenced by the structural domain; the experiences of oppression that result from the misrepresentations and omissions of communities in heritage exist in the interpersonal domain (see p. 44). Data feminism's (D'Ignazio and Klein, 2020) principles (see p. 44) echo the aim of Smith's theory to challenge the simplistic, hegemonic concept of heritage and the power it wields over cultural and historical records (principles 1 and 2), expanding heritage to have a plurality of meanings (principles 4, 5, and 6).

3.3.4 Applying the Theories

Extending concepts from theories of feminism, CDA, and heritage as a process to communities of minoritized genders, I investigate how NLP models can challenge the patriarchal and cisgender assumptions of heritage in the form of archival documentation. The dynamic nature of heritage and language extends to the dynamic nature of text corpora, including the GLAM documentation that serves as my thesis' dataset, opening a path to reflecting upon the challenges of applying contemporary NLP models to historical GLAM documentation.

There are technical and conceptual challenges to applying contemporary NLP models to GLAM documentation. From a technical perspective, the models may struggle to perform well. Beelen et al. (2021) found that a pre-trained BERT model (Devlin et al., 2019) performed better on historical text when the model was fine-tuned on data relevant to the historical text's time period. From a conceptual perspective, because the meaning of language changes with context, the validity of approaches to certain research questions may be undermined. Consider the research of this thesis, which aims to understand the types of biased language in GLAM documentation: using state-of-the-art models such as BERT (Devlin et al., 2019) or GPT-4 (OpenAI, 2023) risks the injection of biases from those models' training data into measures of biased language for my corpus of archival documentation. Although model fine-tuning offers an approach to customizing pre-trained models to particular domains, at the time of writing, no approach to disentangling biases in a pre-trained

model's word representations from those of the in-domain data exists.

On the other hand, perhaps applying NLP models that have a foundation in contemporary worldviews provides a way to contribute to heritage. If one adopts Smith's (2006) definition of heritage as a process, rather than something static and unchanging, could applying an NLP model to GLAM documentation contribute to this process? If applied in support of authorized heritage discourse, perhaps not; but if applied in support of dissenting heritage discourse, I argue that NLP models could contribute to this process by challenging the authorized discourse. Heritage can reinforce memories and knowledge, but it can also recreate and reshape memories and knowledge (Duff and Harris, 2002; Smith, 2006). As Odumosu (2020) notes, heritage in GLAM catalogs, in addition to GLAM collections, has this power: "metadata could be rethought as a cataloging space with the potential to alter historical imbalances of power" (p. 1). Controlling heritage means controlling identities at the level of the individual, the community, and the nation (Adler, 2017; Duncan, 2005; Olson, 2001; Schwartz and Cook, 2002; Smith, 2006; Yale, 2015). In this thesis, I conduct a case study to research the extent to which NLP methods applied to archival documentation can support a redistribution of power, away from men and towards women and trans and gender diverse communities, by calling attention to stereotypical representations and omissions of those gender groups.

This literature review has summarized the origins of social biases in GLAM documentation and ML systems, as well as existing approaches to addressing bias in GLAM and ML. Despite the evidence of social biases being a shared concern across GLAM and ML, there has been a lack of work investigating applications of ML to GLAM for addressing biased language in heritage collections' documentation. This thesis adds a contribution to knowledge at this intersection of GLAM and ML, specifically creating a methodology, annotation taxonomy, annotated data, models, and a model evaluation approach to (1) identify types of gender biased language and (2) enable large-scale analysis of gender biases in archival documentation.

Chapter 4

Methodology

> [N]umbers cannot determine what has moral value, nor what is socially desirable.
>
> –*Montréal Declaration for a Responsible Development of Artificial Intelligence* (2018, p. 7)

This chapter contributes my Ph.D. thesis' methodology. The research question investigated with this chapter is: ***Can existing methods of identifying and categorizing biased language in NLP research be applied to archival metadata descriptions? Why or why not?*** A literature review of existing approaches to identifying and categorizing gender biased language found that the taxonomy of Hitti et al. (2019) could serve as a foundation, but that additional categorizations were needed for identifying gender biases in archival documentation (Chapter 5). That literature review motivated me to develop a new research methodology, the Bias-Aware Methodology. Publications called for greater interdisciplinary engagement and stakeholder collaboration in ML (Blodgett et al., 2020; Crawford, 2017), yet at the time, I found no examples of projects integrating these activities into the entire ML system creation process. Alternative approaches to working with data and technology that considered societal power relations had been published (Costanza-Chock, 2018; D'Ignazio and Klein, 2020), but there was no *methodology* combining human-centered or otherwise interdisciplinary research methods with ML methods. With my Bias-Aware Methodology, I define three activities (visualized

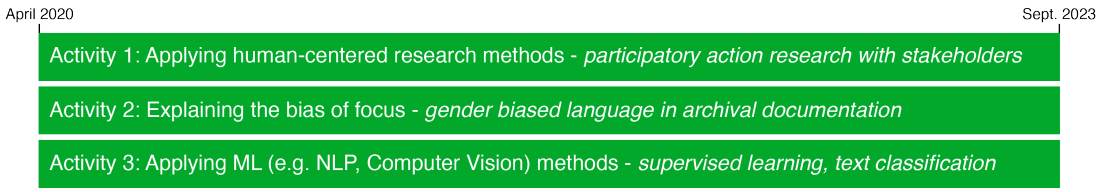


Figure 4.1: **The Bias-Aware Methodology's Activities.** The three parallel activities of my Bias-Aware Methodology with details of my execution of them in italics.

in Figure 4.1 and detailed in §4.1.5) that researchers and practitioners can follow when creating ML models to facilitate interdisciplinary engagement and stakeholder collaboration. Due to the focus of my thesis on working with text data, §4.1 presents the Bias-Aware Methodology as particularly relevant to NLP research. §4.2 explains the relevance of the Methodology to ML research more broadly, and explains how my execution of the Methodology demonstrates a recalibration of ML research for social biases.

Publication Note: I originally wrote my thesis' methodology as a paper titled *Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research*. The paper was published in the *Proceedings of the Second Workshop on Gender Bias for NLP* (Havens et al., 2020), as part of the virtual 28th International Conference on Computational Linguistics. I wrote the paper as lead author, with my supervisors providing feedback as I wrote to guide my revisions. I conducted the research (i.e. the literature review and data extraction, transformation, and analysis) reported in that paper and in this chapter. For the paper as it is presented here (§4.1), I made small changes to the original publication to keep my terminology and formatting consistent across chapters, and to update referenced statistics and literature.

4.1 The Bias-Aware Methodology

4.1.1 Introduction

Analysis of computer systems has raised awareness of their biases, prompting researchers to make recommendations to mitigate harms that biased computer systems cause. Analysis has shown computer systems exhibiting biases through racism[1] (Noble, 2018), sexism[2] (Perez, 2019), and classism[3] (Eubanks, 2017). This list of harms is not exhaustive; biased computer systems may also harm people based on ability, citizenship, and any other identity characteristic. To mitigate harms from biased computer systems, researchers have recommended actions, methods, and practices. However, none of the recommendations comprehensively address the complexity of the problems bias causes.

Considering the numerous *types* of bias that may enter an NLP system, *places* that bias may enter, and *harms* that bias may cause, we propose a Bias-Aware Methodology to comprehensively address the consequences of bias for NLP research. Our Methodology integrates critical reflection on social influences on and implications of NLP research with technical NLP methods. To scope our research direction and inform our Methodology, we draw on an interdisciplinary selection of literature that includes work from the Humanities, Arts, and Social Sciences. We intend the Methodology to (a) support the reproducibility of NLP research, enabling researchers to better understand which perspectives were considered in the research; and (b) diversify perspectives in NLP systems, guiding researchers in explicitly communicating the social context of their research so others can situate future research in contexts that have yet to be investigated.

We begin with our bias statement (§4.1.2) and motivations for proposing our Bias-Aware Methodology (§4.1.3). Next, we summarize the interdisciplinary literature informing the Methodology (§4.1.4), and explain (§4.1.5) and demonstrate it with a case study of our research with archival documentation (§4.1.6). We end with a summary and vision for future NLP research (§4.1.7).

[1] "A belief that one's own racial or ethnic group is superior" (OED, 2013c).

[2] "[P]rejudice, stereotyping, or discrimination, typically against women, on the basis of sex" (OED, 2013d).

[3] "The belief that people can be distinguished or characterized, esp. as inferior, on the basis of their social class" (OED, 2013a).

4.1.2 Bias Statement

We situate this paper in the UK in the 21st century, writing as authors who primarily work as academic researchers. We identify as three women and one man; and as American, German, and Scots. Together we have experience in NLP, Design, Human-Computer Interaction, Data Visualization, Digital Humanities, and Digital Cultural Heritage. In this paper, we propose a Bias-Aware Methodology for NLP researchers. We define **biased language** as:

> written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity; and privileging other people through words or phrases that favor their identity.

Biased language causes representational harms (L. Sweeney, 2013; Vainapel et al., 2015), or the restriction of a person's identity through the use of hyperbolic or simplistic language (Blodgett et al., 2020; Talbot, 2003). NLP systems built on biased language become biased computer systems, which "*systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others" (Friedman and Nissenbaum, 1996, p. 332, emphasis in the original). Representational harms may cause inequitable system performance for different groups of people, leading to allocative harms (Noble, 2018; H. Zhang et al., 2020), or the denial of a resource or opportunity (Blodgett et al., 2020). The people who experience harms from biased NLP systems varies with the context in which people use the system and with the language source on which the system relies. Moreover, people may not be aware they are being harmed given the black-box nature of many systems (Koene et al., 2017). That being said, whether or not people realize they are being prejudiced against, the people experiencing the most harm will be those excluded from the most powerful social group.

4.1.3 Why does NLP need a Bias-Aware Methodology?

Statistics report a homogeneity of perspectives among students in computer-related disciplines that do not reflect the diversity of people affected by computer systems, risking a homogeneity of perspectives in the technology workforce and

the computer systems that workforce develops (Hicks, 2018). For academic year 2018/19, statistics on students in the UK[4] report that the dominant group of people studying computer-related subjects overwhelmingly are white males without a disability.[5,6] Moreover, differences in total numbers of surveyed students across identity characteristics (e.g. sex, ethnicity, disability) skew the statistics in favor of those reported as white, male, and without a disability. Lack of diverse perspectives among students in computer-related disciplines may limit the diversity of perspectives in the workforce, where the development of NLP and other computer systems occurs. As of 2019, the Wise Campaign reported that women comprise 24% of the core-STEM workforce in the UK.[7,8] Lack of diverse perspectives in the development of NLP and other computer systems risks technological decisions that exclude groups of people ("technical bias"), as well as applications of computer systems that oppress groups of people ("emergent bias") (Friedman and Nissenbaum, 1996).

That being said, even if student demographics in NLP and computer-related disciplines become more balanced, the data underlying NLP systems will still cause bias. Theories of discourse state that language (written or spoken) reflects and reinforces "society, culture and power" (Bucholtz, 2003, p. 45). In turn, NLP systems built on human language reflect and reinforce power relations in society, inheriting biases in language (Caliskan et al., 2017) such as stereotypical expectations of genders (Haines et al., 2016) and ethnicities (Garg et al., 2018). Drawing on feminist theory, we argue that all language is biased, because language records human interpretations that are situated in a specific time, place, and worldview (Haraway, 1988). Consequently, all NLP systems are subject to biases originating in the social contexts in which the systems are built ("preexisting bias") (Friedman and Nissenbaum, 1996). Psychology research suggests that biased language causes representational harms: Vainapel et al. (2015) studied how masculine-generic language (e.g. "he") versus gender-neutral language (e.g. "he or she") affected

[4]Situating our research in the UK, we reference statistics from the UK's Higher Education Statistical Agency (HESA).

[5]hesa.ac.uk/news/16-01-2020/sb255-higher-education-student-statistics/subjects

[6]HESA changed its reporting in the 2019/20 academic year so comparable statistics are not available for more recent years.

[7]wisecampaign.org.uk/statistics/2019-workforce-statistics-one-million-women-in-stem-in-the-uk

[8]As of June 2022, the proportion of women in the core-STEM workforce was reported to have increased to 26.9%; see: wisecampaign.org.uk/updated-workforce-statistics-june-2022.

participants' responses to questionnaires. The authors report that women gave themselves lower scores on intrinsic goal orientation and task value in questionnaires using masculine-generic language in contrast to questionnaires using gender-neutral language.[9] The study provides an example of how biased language may harm select groups of people, because the participants reported as women experienced a restriction of their identity, influencing their behavior to conform to stereotypes.

Acknowledging the harms of biased language and biased NLP systems, researchers have proposed approaches mitigating bias, though no approach has fully removed bias from an NLP dataset or algorithm. To mitigate bias in datasets, Webster et al. (2018) produced a dataset of gendered ambiguous pronouns (GAP) to provide an unbiased text source on which to train NLP algorithms. However, the GAP dataset reverses gender roles, assuming that gender is a binary rather than a spectrum (Scheuerman, Spiel, et al., 2020). Any NLP system that uses the GAP dataset thus adopts its preexisting gender bias. Efforts to mitigate bias in algorithms are similarly limited, focusing on technical performance rather than performance in social contexts. Zhao et al. (2018) describe an approach to debias word embeddings, writing, "Finally we show that given sufficiently strong alternative cues, systems can ignore their bias" (p. 16). However, the paper does not explain the intended social context in which to apply the authors' approach, risking emergent bias.[10] Additionally, Gonen and Goldberg (2019) demonstrate how this debiasing approach hides, rather than removes, bias. In our Bias-Aware Methodology, we describe documentation and user research practices that facilitate transparent communication of biases that may be present in NLP systems, facilitating reflection on how to include more diverse perspectives and empower underrepresented people.

[9]The authors report that men showed no difference in their intrinsic goal orientation and task value scores with masculine-generic versus gender-neutral language in the questionnaires; impacts on people who do not identify as either a man or a woman are unknown as the study groups participants into these two gender categories (Vainapel et al., 2015).

[10]While earlier paragraphs in the paper indicate a focus on gender bias and stereotypes related to professional occupations, the authors do not define *bias* or *gender bias*, nor do they identify the types of *systems* to which they refer.

4.1.4 Related Work

To inform our proposed Bias-Aware Methodology, we draw on an interdisciplinary corpus of literature from Computer Science, Data Science, the Humanities, Design, and the Social Sciences.

NLP and ML scholars have recommended actions to diversify perspectives in technological research, recognizing the value of diversity to bias mitigation. Blodgett et al. (2020) and Crawford (2017) recommend interdisciplinary collaboration so researchers can learn from humanistic, artistic, and sociological disciplines regarding human behavior, helping researchers to more effectively anticipate harms that computer systems may cause, in addition to benefits they may bring, addressing risks of emergent bias. They also recommend engaging with the people affected by NLP and other computer systems, testing on more diverse populations to address the risk of technical bias, and rethinking power relations between those who create and those who are affected by computer systems to address the risk of preexisting bias. Though these recommendations address the three types of bias that may enter an NLP system, they do not articulate how to identify relevant people to include in the development and testing of NLP systems. Our Bias-Aware Methodology builds on recommendations from Blodgett et al. (2020) and Crawford (2017) by outlining how to identify and include stakeholders in NLP research (§4.1.5.1).

D'Ignazio and Klein (2020) propose data feminism as an approach to addressing bias in data science. They define data feminism as, "a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought" (p. 8).[11] Data feminism has seven principles: examine power, challenge power, elevate emotion and embodiment, rethink binaries and hierarchies, embrace pluralism, consider context, and make labor visible. These principles facilitate critical reflection on the impacts of data's collection and use in social contexts. Our Bias-Aware Methodology tailors these principles to NLP research, outlining activities that encourage researchers to consider influences on and implications

[11] Intersectionality refers to the way in which different combinations of identity characteristics from one individual to another result in different experiences of privilege and oppression (Crenshaw, 1989, 1991). In feminist thought, multiple viewpoints are needed to understand reality; viewpoints that claim to be objective are, in fact, subjective, because knowledge is the result of human interpretation (Haraway, 1988).

of their work beyond the NLP community (§4.1.5.1).

Within the NLP research community, Bender and Friedman (2018) recommend improved documentation practices to mitigate emergent, technical, and preexisting biases. They recommend all NLP research includes a "data statement," which they describe as, "a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software" (p. 587). Aimed at developers and users of NLP systems, data statements reduce the risk of emergent bias. The authors also note: "As systems are being built, data statements enable developers and researchers to make informed choices about training sets and to flag potential underrepresented populations who may be overlooked or treated unfairly" (p. 599), helping authors of data statements reduce the risk of technical and preexisting biases. A data statement serves as guiding documentation for the case study approach we propose in our Bias-Aware Methodology (§4.1.5.2), documenting the specific context in which NLP researchers work. Our Bias-Aware Methodology guides research activities before, during, and after the writing of a data statement: for researchers reading data statements to find a dataset for an NLP system, our Methodology guides their evaluation of a dataset's suitability for research; for researchers writing data statements, our Methodology guides their documentation of the data collection process.

In addition to technological disciplines, our Methodology draws on Critical Discourse Analysis (CDA) (Fairclough, 2003; van Leeuwen, 2009), Participatory Action Research (PAR) (Aragon et al., 2022; Reid and Frisby, 2008; Swantz, 2008), intersectionality (Crenshaw, 1989, 1991), feminist theories (D'Ignazio and Klein, 2020; Haraway, 1988; Harding, 1995; Moore, 2018), and Design (Martin and Hanington, 2012a). CDA studies language in context. PAR provides a way for NLP researchers to diversify perspectives in their research, engaging with the social context that influences and is affected by NLP systems. Intersectionality reminds researchers of the multitude of experiences of privilege and oppression that bias causes, because no single identity characteristic determines whether a person is "dominant" (favored) or "minoritized" (harmed) (D'Ignazio and Klein, 2020). The case study approach common to Design methods enables a researcher to make progress

on addressing bias through explicitly situating research in a specific time and place, and conducting user research with people to understand their power relations in that time and place (Martin and Hanington, 2012a). Feminist theories value perspectives at the margins, encouraging researchers to engage with people who are excluded from the dominant group in a social context. Feminist theorist Harding (1995) writes, "In order to gain a causal critical view of the interests and values that constitute the dominant conceptual projects...one must start from the lives excluded as origins of their design - from 'marginal' lives" (p. 341). Our Bias-Aware Methodology includes collaboration with people at the margins of NLP research in an effort to empower those minoritized people.

4.1.5 Activities of the Bias-Aware Methodology

Our Bias-Aware Methodology has three main activities: applying human-centered research methods (§4.1.5.1), explaining the bias of focus (§4.1.5.2), and applying NLP methods (§4.1.5.3). Though we discuss the activities individually, we recommend researchers execute them in parallel because each activity informs the others. We aim for the Methodology to include activities that researchers may adapt to their own research context, be their focus on algorithm development, adaptation, or application; or on dataset creation. We hope for this paper to begin a dialogue on tailoring a Bias-Aware Methodology to different types of NLP research.

4.1.5.1 Applying Human-Centered Research Methods

Stakeholder Identification An NLP researcher executing the Bias-Aware Methodology will document the distribution of power in the social context relevant to their research and language source. In the Bias-Aware Methodology, a researcher considers language to be a partial record that provides knowledge situated in a specific time, place, and perspective. To understand which people's perspectives their language source ("the data") includes and excludes, an NLP researcher will identify *stakeholders*, or those who are represented in, use, manage, or provide the data. Specifically, NLP research stakeholders are:

1. The researcher(s),
2. Producers of the data,

3. Institutions providing access to the data,
4. People represented in the data, and
5. People who use the data.

To investigate their stakeholders' power relations, an NLP researcher will observe who dominates the social setting(s) relevant to their research, and who experiences minoritization in the same setting(s). After identifying the stakeholders, the researcher will document their roles as dominant or minoritized, along with any limitations to their identification.

Stakeholder Collaboration To understand how privilege and oppression are experienced among stakeholders, an NLP researcher will conduct PAR (or another human-centered research method; see Aragon et al., 2022 for examples) with representative individuals from all five stakeholder groups. Researchers who conduct PAR attempt to establish collaborative relationships with representatives from their groups of stakeholders. Researchers are not experts bringing NLP systems to stakeholders; rather, researchers and stakeholders collaboratively study a social context to understand how NLP systems could empower people, particularly minoritized people. Instead of seeking an objective perspective, researchers foreground individual stakeholder perspectives, recording them as situated in a specific time and place, and using their multiplicity to gain insight into the complexity of the research's social context. To understand how NLP research can empower people in a specific social context, we propose four *power relations questions* [12] for NLP researchers to answer:

1. Who or what is included in the research?
2. Who or what is excluded from the research?
3. How will the research define knowledge?
4. Who has agency and who can be empowered?

To understand the impacts of dominant people's interests and values, research following the Bias-Aware Methodology will begin from the perspective of minoritized people, those who are typically excluded as a result (even if

[12] We adapted these questions from Moore's work on feminist community archiving (2018).

unintentional) of the interests and values of dominant people. The research will define knowledge as situated in specific times, places, and perspectives. The widespread availability of language as digital data may give the illusion of universal representation. However, CDA reminds the NLP researcher that their data, composed of discourses,[13] are "socially constructed ways of knowing some aspect of reality" (van Leeuwen, 2009, p. 141). Social hierarchies influence the data that becomes widely available, rendering minoritized groups of people invisible due to their exclusion from the data, or misrepresenting them due to their exclusion from the data collection process.

An NLP researcher will weigh insights gathered from different stakeholder groups equally, making the research's knowledge multi-faceted. Explicit documentation of the time, place, and perspective that produced the knowledge will inform future NLP research. Should a future researcher wish to reproduce the research, the documentation will guide the future researcher in seeking the proper social context. Should a future researcher wish to build upon the research, they will be able to compare and contrast the research's social setting with their own, guiding them in determining potential contributions.

Unavailable Stakeholders In situations where the researcher cannot conduct PAR with stakeholders, the researcher will write a data biography.[14] A data biography documents where data were collected and stored, who collected and owns the data, and why, when, and how the data were collected (Krause, 2019). Writing a data biography facilitates critical reflection on the social influences on and social implications of a dataset, informing technical decisions when applying NLP methods. Datasets may circulate oppression of minoritized groups through inclusion and through omission. The key to recognizing who is dominant and minoritized is understanding that an individual may be both; power relations vary with the context of research.

4.1.5.2 Explaining the Bias of Focus

When explaining the type of bias on which NLP research focuses, a researcher will provide a definition and explain how this type of bias relates to other

[13] "A connected series of utterances by which meaning is communicated" (OED, 2013b).
[14] We All Count has a free data biography tool at: wac-survey-rails.herokuapp.com.

Structural Bias		Contextual Bias	
Gender Generalization	**A lawyer** must always carry **his** phone.	Societal Stereotype	The event was **sports-themed** for all **fathers** volunteering.
Explicit Marking of Sex	The role of **a waitress** is overlooked by the restaurant owners.	Behavioral Stereotype	**All girls** are **sensitive**.

Table 4.1: **Examples of Gender Biases in Text from Hitti et al., 2019.**

types of bias. For example, AllSides.com's ratings may guide the classification of political bias in news,[15] Hanson et al.'s (2015) *Accessible Writing Guide* may inform research with stakeholders who include people with disabilities, and Hitti et al. (2019) provide a model for how to clearly define and classify gender bias in collaboration with interdisciplinary experts. Table 4.1 provides examples of gender biased language organized into Hitti et al.'s (2019) gender bias taxonomy. When following the Bias-Aware Methodology, NLP research to create annotated datasets for other types of bias will similarly include collaboration with relevant disciplinary experts (i.e. racial bias with critical race theory experts) to define and categorize types of bias relevant to the research. When writing a data statement's *Curation Rationale* (Bender and Friedman, 2018), an NLP researcher will include a definition of their bias of focus.

In the answers to the power relations questions, an NLP researcher will describe how they consider intragroup differences within their stakeholder groups, in addition to differences between dominating and minoritized stakeholder groups, because the intersection of identity characteristics, rather than one identity characteristic in isolation, determines how people experience oppression (Crenshaw, 1989, 1991). Due to the complexity that intersecting identity characteristics add to evaluations of bias, in the Bias-Aware Methodology, an NLP researcher will use case studies.

Case studies gather information in a clearly-defined context and present the resulting knowledge as connected to a specific time, place, and people. To conduct a case study, an NLP researcher will "determine a problem, make initial hypotheses, conduct research through interviews, observations, and

[15] See the Media Bias Ratings at: www.allsides.com/media-bias/media-bias-ratings.

other forms of information gathering [such as PAR], revise hypotheses and theory, and tell a story" (Martin and Hanington, 2012a, p. 28). Feminist theories' focus on agency and lived experience as situated in a specific context adds value to PAR by helping a researcher anticipate and critically examine the implications of PAR's drive towards action (Reid and Frisby, 2008). When documenting their case study in blogs, presentations, or publications, an NLP researcher will discuss potential applications of the research beyond the case study's context, anticipating potential benefits and harms. Potential harms may outweigh potential benefits, making the best decision not to build an NLP system (Crawford, 2017; Graeff, 2020).

4.1.5.3 Applying NLP Methods

When applying NLP methods in the Bias-Aware Methodology, an NLP researcher should acknowledge biases found with any algorithms they use in their data statement. For example, when applying word embeddings, an NLP researcher could look to Bolukbasi et al. (2016), Caliskan et al. (2017), and Kurita et al. (2019) on gender bias; Swinger et al. (2019) on racial bias; Diaz et al. (2018) on age bias; Papakyriakopoulos (2020) on sexuality and nationality bias; and Gonen and Goldberg (2019) on the inadequacy of debiasing word embeddings. When applying part-of-speech tagging, dependency parsing, or machine translation, an NLP researcher could look to Garimella et al. (2019) and Stanovsky et al. (2019) for understanding how these methods have been shown to exhibit gender bias. If an NLP researcher will train an algorithm on their language source, research documentation will describe the training process and results. If the research includes annotation, documentation will include instructions given to annotators.

For NLP research on algorithms, we recommend considering approaches to making bias transparent, in addition to reducing the biased behavior of algorithms. Research from Kaneko et al. (2019) and Zhao et al. (2018) on mitigating bias in word embeddings provide starting points for algorithmic bias research, as their methods have yet to be evaluated in diverse contexts. However, Gonen and Goldberg (2019) have shown the limits of debiasing word embeddings. We argue that the situated nature of data, and thus the situated nature of knowledge drawn from data, makes the elimination of bias impossible. Investigating how to make bias transparent provides an alternative

direction for NLP researchers interested in mitigating bias in NLP systems. Whether making bias transparent or reducing biased behavior of algorithms, NLP researchers following the Bias-Aware Methodology will collaborate with relevant disciplinary experts and minoritized stakeholders in determining how to evaluate a model for bias.

To support the training of algorithms in diverse contexts, NLP research on datasets will define the context of its language source's collection and annotation. An NLP researcher will provide data statements to inform algorithms' training and evaluation, ensuring reproducibility and avoiding unintended harms from misapplications of algorithms (Bender and Friedman, 2018). Similarly, dataset research will include disciplinary experts and minoritized stakeholders in datasets' creation, annotation, and evaluation.

4.1.6 Case Study

In this section we describe how we are implementing the Bias-Aware Methodology for NLP research in a case study on bias in documentation of the archival collections of the Heritage Collections department at the University of Edinburgh ("the HC Archives").[16] The HC Archives' documentation describes a variety of heritage collections and items, such as letters, journals, photographs, degree certificates, and drawings; on a variety of topics, such as religion, research, teaching, architecture, and town planning, largely as relevant to Edinburgh, Scotland, and the University of Edinburgh. The HC Archives' documentation is in the public domain, available for browsing at archives.collections.ed.ac.uk. The dates of the descriptions in the HC Archives' catalog are often unknown; for the circa 30% of descriptions that are dated, the dates range from the 16th century to the present (the HC Archives is actively collecting and describing heritage material). The extracted dataset documents 1,081 collections of varying sizes and consists of 2,754,044 tokens.

For consistency with the outline of a Bias-Aware Methodology, we group our case study into the same three activities, explaining our examination of power relations, our bias of focus, and then our application of NLP methods.

[16]Metadata documents information about collections of cultural heritage records. Archival catalogs have numerous metadata fields that contain descriptions written by people who Archives hire to document their collection items. These descriptions are the language source we refer to as *archival documentation* (Angel and Fuchs, 2018).

Each subsection includes accomplished, ongoing, and planned future work. To demonstrate how we execute the three activities in parallel, as proposed in §4.1.5, we first provide a chronological overview.

Initially, our research began with information gathering linked to a PAR method. We reviewed literature on bias in NLP and Archives, and on Digital Humanities research.[17] We also met with employees at the HC Archives to better understand the HC Archives' policies, which guide cataloging and description practices, such as the metadata standards used. The employees described how they are proactively challenging the inherited metadata and inherited practices of the HC Archives. After the literature review and meeting we began writing data statements for the HC Archives' documentation and for our research.

Due to the limited research on NLP methods applied to archival metadata, and limited large-scale analysis of metadata descriptions, we undertook a pilot data project,[18] walking through the process of extracting metadata descriptions from a single archival collection, adding historical context to our documentation of the extracted descriptions, and calculating corpus analytics using ElementTree[19] and NLTK[20] in a Jupyter Notebook.[21] After establishing a workflow to extract metadata descriptions from the HC Archives' catalog, we again met HC Archives' employees to discuss the challenges that biased language poses to their work and to their visitors. This meeting helped us add to our data statements, identify stakeholders in our research, and begin describing the stakeholders' power relations. Moreover, the meeting confirmed the value of an NLP system that detects and classifies bias, as the HC Archives does not currently have a computational approach to measuring bias in its catalog documentation.

[17] Digital Humanities is characterized by collaborations between technologists and humanists that often analyze data sources with historical language (Champion, 2016).
[18] See the pilot code at: github.com/thegoose20/eula41.
[19] docs.python.org/3/library/xml.etree.elementtree.html
[20] www.nltk.org
[21] jupyter.org

4.1.6.1 Applying Human-Centered Research Methods: Researcher and Archives

Stakeholder Identification In our execution of the Bias-Aware Methodology, we study power relations among five stakeholders:

1. Us (the authors) as researchers,
2. The HC Archives' employees,
3. The HC Archives (as an institution),
4. People represented in the archival documentation, and
5. The HC Archives' visitors.

Literature on power relations in Archives and the wider GLAM sector (Adler, 2017; Caswell and Cifor, 2019; Hauswedell et al., 2020; McPherson, 2012; Risam, 2015) informed our identification of these stakeholders. We recorded our understanding of their power relations in our data statement (Appendix A) and power relations document (Appendix B).[22]

Stakeholder Collaboration In line with PAR, we collaborate with HC Archives employees to learn about their perception of biased language in archival documentation, and challenges and potential approaches to addressing it. We facilitated a group discussion with stakeholders who had a range of roles, including technical, curatorial, administrative, servicing, and documenting responsibilities; and a range of GLAM work experience, from one year to over 20 years. The discussion informs our understanding of the range of attitudes towards bias and neutrality in archival documentation. Results of the discussion enabled us to answer the power relations questions (Appendix B).

Unavailable Stakeholders Our stakeholders include people who documented HC Archives collections but no longer work there, and people who are written about in HC Archives documentation, which describes material dating back to the 1st century AD. To study power relations among these unavailable stakeholders, we wrote a data biography (Appendix C) for HC Archives documentation. The data biography informs our understanding of the power

[22]The data statement and power relations document included in this thesis' appendices are updated versions of those documents as originally published with this paper in 2020.

relations at play in our research, which in turn informs our data statement and technical decisions about NLP methods to apply.

4.1.6.2 Explaining the Bias of Focus: Contextual Gender Bias

Our NLP research focuses on identifying types of contextual gender bias from archival documentation, complementing Hitti et al.'s (2019) focus on identifying structural gender bias. We build upon their taxonomy of gender bias, which has two subtypes of contextual bias: behavioral stereotypes and societal stereotypes. We expand their definitions and subtypes of contextual bias (Chapter 5) to simplistic, hyperbolic language in metadata descriptions that indicates the presence of stereotypes, because historical text often contains spellings and syntax (among other linguistic characteristics) different to the modern text on which NLP tools have been developed (Casey et al., 2021). In the context of the HC Archives, gender biases may cause representational harms, because the HC Archives supports information access, circulating ideas documented in its catalog when users search it online. Societal and behavioral stereotypes present in HC Archives documentation may negatively impact perceptions of people represented in the descriptions. We are researching the types of gender bias in the descriptions, and ways to measure such biases, in an effort to support the HC Archives in mitigating harms from biased documentation.

4.1.6.3 Applying NLP Methods: Information Extraction for Classification

Information Extraction Methods The archival documentation we use as this case study's language source are from the HC Archives' public, online catalog. Using the programming language Python version 3.8.10,[23] and the libraries ElementTree,[24] pandas (The pandas development team, 2023), and NLTK (Loper and Bird, 2002), we obtained descriptive metadata fields as Extensible Markup Language (XML) data using the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH)[25] and then filtered the metadata for descriptive fields relevant to our research, and then removed duplicate

[23] www.python.org
[24] docs.python.org/3.8/library/xml.etree.elementtree.html
[25] www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

by field	Biographical / Hist.	Scope and Cont.	Processing Info.	all fields
total words	801,893	208,190	11,016	966,763
total sentences	11,323	55,434	1,691	68,448
by collection	minimum	maximum	mean	std. dev.
words	7	156,747	1,036.200	7,784.500

Table 4.2: **Dataset Summary Statistics.** Total, minimum, maximum, mean, and standard deviation (std. dev.) for words and sentences in the descriptions from the Biographical / Historical (Biographical / Hist.), Scope and Contents (Scope and Cont.), and Processing Information (Processing Info.) metadata fields. The descriptions were gathered from all 1,231 collections in the HC Archives' catalog in April 2020. Tokens and sentences were calculated using the Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002); words were estimated by calculating the number of alphabetic tokens.

descriptions. Table 4.2 summarizes the resulting corpus.[26] HC Archives organizes metadata hierarchically, creating metadata for collections (the archival term is *fonds*), subcollections (the archival terms are *sub-fonds*, *series*, and *sub-series*), and *items*). We grouped all subcollection and item descriptions within their overarching collection.

Annotations to Inform Classification With our case study, we created a dataset annotated for gender biased language, on which we then trained classification models to identify types of gender bias in text. Due to ethical concerns regarding the use of crowdsourcing platforms (Gleibs, 2017; Irani, 2015, 2016; Taylor, 2018), people employed to contribute to the annotation work were paid above minimum wage. To guide the annotation process and ensure the reproducibility of our research, we documented instructions we follow to annotate contextual gender bias (Appendix D). We collaborated with the HC Archives and a gender studies expert to write these instructions. As we published the results of our research (Havens et al., 2020, 2022), we provided documentation of the annotation instructions, data statements, data biography, and power relations questions. After creating a dataset annotated for contextual gender bias (Chapter 5), we trained discriminative classifiers on the dataset using supervised learning (Chapter 6). We then experimented with

[26]Further PAR with the HC Archives led to the addition of descriptions from the "Title" metadata field after this paper's publication, a change reflected in Chapter 5.

and evaluated how the classifiers differentiate between types of contextual gender bias in archival documentation, and will report openly on the results (chapters 6 and 7) in future publications.

4.1.7 Conclusion

We introduced the Bias-Aware Methodology for NLP research to mitigate harms from biased NLP systems, which serves as the Methodology for this Ph.D. thesis. The Methodology integrates practices and methods from NLP, ML, Data Science, Gender Studies, Linguistics, and Design. Due to the numerous types of bias, the intersectional nature of oppression, and the possibility of direct and indirect harms from bias, detecting and measuring bias is a complex process. Our Methodology encourages NLP researchers to situate their work in case studies, explicitly describing the context of and stakeholders in their research. We advise NLP researchers to build the time and resources needed to undertake such work into project plans, and to put consideration of data and model biases at the center of their research. Documenting instances of bias and their associated power relations will enable the NLP community to look for patterns across different contexts that use NLP systems. Amassing case studies in order to look for such patterns will guide NLP research towards generalizable approaches to bias mitigation, approaches that do not unintentionally minoritize people whose perspectives were unknowingly excluded.

4.2 Comments on the Paper

4.2.1 Generalizing the Methodology to ML

Although the paper focuses on why *NLP* needs a Bias-Aware Methodology, the same rationale applies to *ML* more broadly. Even if data are not language-based, be it text or audio, they still reflect and reinforce societal power relations (D'Ignazio and Klein, 2020; Noble, 2018; O'Neil, 2016; Perez, 2019; Rogers, 2021). Moreover, the design of ML systems using that data further engineer particular perspectives, or biases, into those systems (Benjamin, 2019; Crawford, 2021; Hicks, 2018; Suresh and Guttag, 2021), as do the metrics chosen to evaluate those technologies (Paullada et al., 2021; Raji et al., 2021) and the way those technologies are reported (Birhane et al., 2022; Raji et al., 2021).

4.2.2 Recalibrations with the Methodology

Applying the Bias-Aware Methodology throughout my Ph.D. research, I demonstrate how the Methodology can facilitate a recalibration of ML research for social biases, prioritizing:

- ***Quality* over quantity** by defining a specific bias of focus (§4.1.6, §5.1.5), and identifying specific groups of stakeholders who impact and are impacted by the research (§4.1.6.1).
- ***Accuracy* over efficiency** by engaging with identified stakeholders in the ML model creation process end to end, from problem definition (§4.1.6.1)[27] to dataset curation (§5.1.6) to model evaluation (Chapter 7).
- ***Representativeness* over convenience** by collaborating with stakeholders through PAR (§4.1.6.1, Chapter 7), and reflecting upon how the perspectives of stakeholders unavailable for collaboration can be incorporated into the research, at minimum communicating the omission of those perspectives as a limitation of the research (§4.1.6.1).

[27]The HC Archives was already dedicating resources to addressing gender bias in their collections when I began my Ph.D. research, so the aims of my thesis aligned with that team's priorities from the start.

- ***Situated thinking* over universal thinking** by positioning the case study in a social context where the perspectives of research participants and an ML model's audience are described in terms of time period, location, and communities of people (§4.1.6).

The next three chapters describe how I applied the Bias-Aware Methodology, executing its three activities in parallel to produce a Taxonomy of Gendered and Gender Biased Language (Chapter 5), text datasets annotated for gender biases (Chapter 5), gender biased text classification models (Chapter 6), and a participatory approach to data and model evaluation (Chapter 7).

Chapter 5

Annotated Data Creation

> Bias everyone with as many perspectives as possible.
>
> –Raghava KK, *Coloring Outside the Lines*, 2013

The research question investigated with this chapter is: ***What types of gender bias are present in the language of archival documentation?*** This chapter contributes (1) a Taxonomy of Gendered and Gender Biased Language and (2) datasets of archival documentation annotated according to the Taxonomy, which together characterize how gender biases may manifest in archival documentation. I was motivated to create a bespoke annotation taxonomy and annotated dataset to train my classification models due to limitations with existing taxonomies and datasets (Cao and Daumé, 2021; Dinan, Fan, Wu, et al., 2020; Doughman et al., 2021; Hitti et al., 2019), as discussed in §5.1.5. My Taxonomy of Gendered and Gender Biased Language is the first taxonomy of gender biased language applied to text classification models that is inclusive of trans and gender diverse identities, and that accounts for uncertainty in gender identification. The Taxonomy is also the first taxonomy of biased language developed for ML-GLAM collaborations.

The annotated datasets are the first datasets created from GLAM documentation for the purpose of training ML models to detect biases. Although research on bias in GLAM catalogs is not new, most of this research focuses on classification schemes and controlled vocabularies (Adler, 2016, 2017; Adler and Harper, 2018; Drabinski, 2013; Furner, 2007; Junginger and

Dörk, 2021; Olson, 2001). More recently, however, resources have begun to be dedicated to studying the descriptive language in GLAM catalogs' metadata fields, though approaches to this work are largely manual (Caswell, 2022; Caswell and Cifor, 2016, 2019; Tai, 2021; Wetli, 2019). Geraci (2019), Baker and Salway (2020), and Salway and Baker (2020) have published computational approaches to addressing and analyzing biased language in metadata descriptions, however they do not provide reusable resources (i.e. datasets, code, or models) for future research. My datasets are the first datasets of GLAM documentation to be augmented with annotations, providing a publicly-available resource for future analyses of biased language and for ML model creation.[1]

§5.1 presents the Taxonomy of Gendered and Gender Biased Language, including the rationale behind the labels in the Taxonomy and the datasets annotated for gender biased language, created through the application of the Taxonomy to archival documentation. §5.2 discusses the generalizability of the Taxonomy and datasets to future research on bias in ML, and explains how my creation of the Taxonomy and datasets further demonstrates a recalibration of ML for social biases.

Publication Note: I originally wrote §5.1 as a paper titled *Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text*. The paper was published in the *Proceedings of the Fourth Workshop on Gender Bias for NLP* (Havens, Terras, et al., 2022), as part of the North American Chapter of the Association for Computational Linguistics Conference. I wrote the paper as lead author, with my supervisors providing feedback as I wrote to guide my revisions. My responsibilities as lead author included executing all PAR sessions; recruiting, hiring, and training annotators; conducting annotations as lead annotator (A0); and performing all data transformations and analysis. For the paper as presented in the next section (§5.1), I made small changes to the original publication to keep my terminology and formatting consistent across chapters, and added two subsections with additional detail about the annotated data to ensure reproducibility (§5.1.6.1 and §5.1.6.2). I also included additional figures and tables that did not appear in the original publication.

[1] github.com/thegoose20/annot

5.1 The Taxonomy and Datasets

5.1.1 Introduction

The need to mitigate bias in data has become urgent as evidence of harms from such data grows (Birhane and Prabhu, 2021; Noble, 2018; O'Neil, 2016; Perez, 2019; L. Sweeney, 2013; Vainapel et al., 2015). Due to the complexities of bias often overlooked in ML bias research, including NLP (Devinney et al., 2022; Stańczak and Augenstein, 2021), Blodgett et al. (2020), Leavy (2018), and Crawford (2017) call for greater interdisciplinary engagement and stakeholder collaboration. The GLAM sector has made similar calls for interdisciplinary engagement, looking to applications of data science and ML to better understand and mitigate bias in GLAM collections (Geraci, 2019; Padilla, 2017, 2019). Supporting the NLP and GLAM communities' shared aim of mitigating the minoritization of certain people that biased language causes, we provide a taxonomy of gender biased language and demonstrate its application in a case study with GLAM documentation.

We use *GLAM documentation* to refer to the descriptions of heritage items written in GLAM catalogs. Adapting our previously published definition, we use *gender biased language* to refer to "language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their [gender] identity; and privileging other people through words or phrases that favor their [gender] identity" (Havens et al., 2020, p. 108). We focus on gender bias due to the contextual nature of gender and bias (they vary across time, location, culture, and people), as well as the existing efforts of our partner institution, the HC Archives, to mitigate gender bias in its documentation.

GLAM documentation provides a unique benefit compared to many text sources: it contains historical and contemporary language. GLAM continually acquire and describe heritage items to enable the items' discoverability. In Archives, heritage items include photographs, handwritten documents, instruments, and tweets, among other materials. Heritage items and the language that describes them influence society's understanding of the past, the present, and the direction society is moving into the future (Benjamin, 2019; Cook, 2011; Duff and Harris, 2002; Smith, 2006; Welsh, 2016; Yale, 2015).

Through research with GLAM documentation, variations in biased language could be better understood. Should diachronic patterns emerge, the NLP community could train models to identify newly-emerging, previously unseen types of bias.

This paper presents an annotation taxonomy, the Taxonomy of Gendered and Gender Biased Language (§5.1.5), to label gender biased language inclusive of trans and gender diverse identities, as well as a dataset of historical and contemporary language from British English archival documentation annotated according to the Taxonomy. Linguistics, Gender Studies, GLAM, and NLP literature inform the Taxonomy's categorization of gender biased language. As a result, the Taxonomy holds relevance beyond the GLAM sector in which we situate our work. The Taxonomy may be applied when creating NLP datasets or models, or when measuring varieties of gender bias in language, because the Taxonomy's definitions of types of gender biases are rooted in the language of text, rather than an abstracted representation of text. Uniquely, our Taxonomy includes labels that record uncertainty about a person's gender.

As we situate our work in the GLAM sector, this paper provides a case study (§5.1.6) demonstrating how the annotation Taxonomy was applied to create an annotated dataset of archival documentation. For future NLP work, the resulting dataset of historical and contemporary language annotated for gender biases provides a corpus to analyze gender biased language for diachronic patterns, to analyze correlations between types of gender biases, and to develop gender bias classification models. Specific to the GLAM sector, gender bias classification models could enhance collection reviews. A model's ability to automatically identify descriptions of heritage items that contain gender biases would enable efficient prioritization of the additions and revisions needed on outdated, incorrect, and otherwise harmful descriptions in GLAM documentation.

5.1.2 Bias Statement

This paper adopts our previously published definition of biased language (Havens et al., 2020), narrowing the focus to gender bias as written in §5.1.1. Gender biased language may cause representational or allocative

harms to a person of any gender (Blodgett et al., 2020; Crawford, 2017). The Taxonomy created in this paper considers a person's gender to be self-described and changeable, rather than being limited to the binary and static conceptualization of gender as either a man or woman since birth (Keyes, 2018; Scheuerman, Spiel, et al., 2020). Recognizing that a person's gender may be impossible to determine from the information available about them, the Taxonomy also allows annotators to record uncertainty (Shopland, 2020). Furthermore, the paper acknowledges that characteristics other than gender, such as racialized ethnicity and economic class, influence experiences of power and oppression (Crenshaw, 1989, 1991). Drawing on Archival Science and feminist theories, the paper considers knowledge derived from language as situated in a particular perspective and, as a result, incomplete (Haraway, 1988; Harding, 1995; Tanselle, 2002).

To communicate this paper's perspective, we as authors report our identification as three women and one man; and our nationalities, as American, German, and Scots. Annotators identify as women (one specifying queer woman and one, cis woman); they are of American, British, Hungarian, and Scots nationalities. Though annotators do not represent great gender diversity,[2] the annotation process still contributes to the advancement of gender equity. As women, the annotators identify as a minoritized gender. The evolution of British English demonstrates the historical dominance of the perspective of the heteronormative man, and the pejoration of terms for women (Lakoff, 1989; Schulz, 2000; Spencer, 2000).[3] Creating a women-produced dataset challenges the dominant gender perspective by explicitly labeling where minoritized genders' perspectives are missing (D'Ignazio and Klein, 2020; Fairclough, 2003; Smith, 2006).

5.1.3 Related Work

Evidence of bias in ML data and models abound regarding gender (Kurita et al., 2019; Zhao et al., 2019), disability (B. Hutchinson et al., 2020), racialized ethnicities (Sap et al., 2019), politics and economics (Elejalde et al., 2017),

[2]The availability of people who responded to the annotator application and the annotation timeline limited the gender diversity that could be achieved among annotators.

[3]In the 16th century, grammarians instructed writers to write "men" or "man" before "women" or "woman." In the 18th century, "man" and "he" began to be employed as universal terms, rather than "human" and "they" (Spencer, 2000).

and, for an intersectional approach (Crenshaw, 1989, 1991), a combination of characteristics (Jiang and Fellbaum, 2020; C. Sweeney and Najafian, 2019; Tan and Celis, 2019). Harms from such biases are also well documented (Birhane and Prabhu, 2021; Costanza-Chock, 2018; Noble, 2018; L. Sweeney, 2013; Vainapel et al., 2015). Despite numerous bias mitigation approaches put forth (Cao and Daumé III, 2020; Dinan, Fan, Williams, et al., 2020; Hube and Fetahu, 2019; Webster et al., 2018; Zhao, Wang, Yatskar, Ordonez, et al., 2018), many have limited efficacy, failing to address the complexity of biased language (Blodgett, Lopez, et al., 2021; Gonen and Goldberg, 2019; Stańczak and Augenstein, 2021).

Methods of removing bias tend to be mathematically focused (e.g. Basta et al., 2020; Borkan et al., 2019). As McCradden et al. (2020) state, typical ML bias mitigation approaches assume biases' harms can be mathematically represented, though no evidence of the relevance of proposed bias metrics to the real world exists. On the contrary, Goldfarb-Tarrant et al. (2021) found no correlation between a commonly used intrinsic bias metric, Word Embedding Association Test, and extrinsic metrics in the downstream tasks of coreference resolution and hate speech detection. Due to the misalignment between abstract representations of bias and the presence and impact of bias, this paper presents an annotation taxonomy to measure biased language at its foundation: words.

Limitations to bias mitigation efforts also result from overly simplistic conceptualizations of bias (Blodgett et al., 2020; Devinney et al., 2022; Stańczak and Augenstein, 2021). NLP gender bias work, for example, often uses a binary gender framework either in its conceptualization (such as Webster et al., 2018) or application (such as Dinan, Fan, Wu, et al., 2020), and tends to focus on one variety of gender bias, stereotypes (Bolukbasi et al., 2016; Doughman et al., 2021; Stańczak and Augenstein, 2021). NLP bias work more generally often asserts a single ground truth (Basile et al., 2021; Davani et al., 2022; Sang and Stanton, 2022). Despite evidence that bias varies across domains (Basta et al., 2020), approaches to mitigating bias have yet to address the contextual nature of biased language, such as how it varies across time, location, and culture (Bjorkman, 2017; Bucholtz, 1999; Corbett, 1990). This paper adopts a data feminist (D'Ignazio and Klein, 2020) and perspectivist (Basile, 2022) approach to situate identification and measurement of bias in a

particular context.

Data feminism views data as situated and partial, drawing on feminist theories' view of knowledge as particular to a time, place, and people (Crenshaw, 1989, 1991; Haraway, 1988; Harding, 1995). Similarly, the Perspectivist Data Manifesto encourages disaggregated publication of annotated data, recognizing that conflicting annotations may all be valid (Basile, 2022). Indigenous epistemologies, such as the Lakota's concept of "waȟkàŋ," further the notion of the impossibility of a universal truth. Translated as "that which cannot be understood," waȟkàŋ communicates that knowledge may come from a place beyond what we can imagine (J. E. Lewis et al., 2018). Our Taxonomy thus permits annotations to overlap and record uncertainty, and our aggregated dataset incorporates all annotators' perspectives.

Encouraging greater transparency in dataset creation, Bender and Gebru et al. (2021) and Jo and Gebru (2020) caution against creating datasets too large to be adequately interrogated. Hutchinson et al. (2021), Mitchell et al. (2019), and Bender and Friedman (2018) propose new documentation methods to facilitate critical interrogation of data and the models trained on them. Our appendices include a data statement documenting the creation of the annotated data presented in this paper (Appendix E). To maximize the transparency of our data documentation, we will publish the data only after further interrogation of its gender bias annotations, including collaborative analysis with the HC Archives team.

5.1.4 Methodology

To practically apply theories and approaches from NLP, data feminism, and indigenous epistemologies, we apply the case study method, common to Social Science and Design research. Case studies use a combination of data and information gathering approaches to study particular phenomena in context (Martin and Hanington, 2012a), suitable for annotating gender biased language because gender and bias vary across time, location, and culture. Furthermore, case studies report and reflect upon outliers discovered in the research process (ibid.), supporting our effort to create space for the perspectives of people minoritized due to their gender identity. After first developing the annotation Taxonomy through an interdisciplinary

literature review and Participatory Action Research (PAR) with the University of Edinburgh's Heritage Collections (HC) team (§5.1.5), we applied the Taxonomy in a case study to create datasets annotated for gender bias (§5.1.6).

Adopting our previously published Bias-Aware Methodology (Havens et al., 2020), we employed PAR (Reid and Frisby, 2008; Swantz, 2008), collaborating with the institution that manages our data source: the HC team. Due to validity (Welty et al., 2019) and ethical concerns (Gleibs, 2017) with crowdsourcing, we hired annotators with expertise in Archives (the domain area of the case study's data) and Gender Studies (the focus area of this paper's bias mitigation) to apply the Taxonomy in a case study. Hiring a small number of annotators enables us to publish disaggregated versions of the annotated data, implementing data perspectivism (Basile, 2022; Basile et al., 2021).

Following the approach of Smith (2006) to heritage, we consider heritage to be a process of engaging with the past, present, and future. Annotators in this paper's case study visited, interpreted, and negotiated with heritage (Smith, 2006) in the form of archival documentation. Annotating archival documentation with labels that mark specific text spans as gender biased transforms the documentation, challenging the authorized heritage discourse (Smith, 2006) of the heteronormative man. We aim such explicit labeling to recontextualize the archival documentation, transforming its language by placing it in a new social context (Fairclough, 2003): the 21st century UK, with gender conceptualized as a self-defined, changeable identity characteristic. We aim this negotiation-through-annotation to guide the NLP models we will create with the data in the future towards more equitable representations of gender (Chapter 6).

5.1.5 Annotation Taxonomy

The Taxonomy of Gendered and Gender Biased Language organizes labels (lettered) into three categories (numbered). Category names are in ***bold italics*** and label names are in *italics*. Each label's listing includes a definition and example. Examples are provided for each label, with the annotated text underlined. For every label, annotators could annotate a single word or multiple words. Examples come from the archival documentation summarized in §5.1.6 except for 1(a), *Non-binary*, and 3(d), *Empowering*, because

annotators did not find text relevant to their definitions (the "Fonds ID," or collection identifier, indicates where in the HC Archives catalog the example descriptions may be found). §5.1.7 further explains the rationale for the Taxonomy's labels, and how they facilitate analysis and measurement of gender biased language.

1. ***Person Name***: the name of a person, including any pre-nominal titles (e.g. Professor, Mrs., Sir, Queen), when the person is the primary entity being described (rather than a location named after a person, for example)

    (a) *Non-binary*: the pronouns, titles, or roles of the named person are non-binary
    Example: Francis McDonald went to the University of Edinburgh where they studied law.

    (b) *Feminine*: the pronouns, titles, or roles of the named person are feminine
    Example: "Jewel took an active interest in her husband's work..." (Fonds ID: Coll-1036)

    (c) *Masculine*: the pronouns, titles, or roles of the named person are masculine
    Example: "Martin Luther, the man and his work." (Fonds ID: BAI)

    (d) *Unknown*: any pronouns, titles, or roles of the named person are gender neutral, or none are provided
    Example: "Testimonials and additional testimonials in favour of Niecks, candidacy for the Chair of Music, 1891." (Fonds ID: Coll-1086)

2. ***Linguistic***: gender marked in the way a word or words reference a person or people, assigning them a grammatical gender

    (a) *Generalization*: use of a gender-specific term (e.g. roles, titles) to refer to a group of people that could identify as more than the specified gender
    Example: "His classes included Anatomy, Practical Anatomy...Midwifery and Diseases of Women, Therapeutics, Neurology...Public Health, and Diseases of the Skin." (Fonds ID: Coll-1118)

(b) *Gendered Role*: use of a word denoting a person's role that marks either a non-binary, feminine, or masculine grammatical gender
Example: "New map of Scotland for Ladies Needlework, 1797" (Fonds ID: Coll-1111)

(c) *Gendered Pronoun*: marking a person or people's grammatical gender with gendered pronouns (e.g. "she," "he," "ey," "xe," or "they" as a singular pronoun)
Example: "He obtained surgical qualifications from Edinburgh University in 1873" (Fonds ID: Coll-1096)

3. ***Contextual***: expectations about a gender or genders that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (e.g. sentence structure or word choice)

(a) *Stereotype*: a word or words that communicate an expectation of a person or people's behaviors or preferences that does not reflect the extent of their possible behaviors or preferences; or that focus on a single aspect of a person that doesn't represent that person holistically
Example: "The engraving depicts a walking figure (female) set against sunlight, and holding/releasing a bird." (Fonds ID: Coll-1116)

(b) *Omission*: focusing on the presence, responsibility, or contribution of one gender in a situation where more than one gender has a presence, responsibility or contribution; or defining a person in terms of their relation to another person
Example: "This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version in 1909 entitled *Apollinaire and his friends*." (Fonds ID: Coll-1090).

(c) *Occupation*: a word or words that refer to a person or people's job title for which the person or people received payment, excluding occupations in pre-nominal titles (for example, "Colonel Sir Thomas" should not have an *Occupation* label)
Example: "He became a surgeon with the Indian Medical Service." (Fonds ID: Coll-1096).

(d) *Empowering*: reclaiming derogatory words as positive
Example: a person describing themself as queer in a self-affirming manner.

We chose to build on the gender bias taxonomy of Hitti et al. (2019) because the authors grounded their definitions of types of gender bias in Gender Studies and Linguistics, and focused on identifying gender bias at the word level, aligning with our approach. Though Dinan et al. (2020) also provide a framework for defining types of gender bias, their framework focuses on relationships between people in a conversation, identifying "bias when speaking ABOUT someone, bias when speaking TO someone, and bias from speaking AS someone" (p. 316). The nature of our corpus makes these gender bias dimensions irrelevant to our work: GLAM documentation contains descriptions that only contain text written *about* a person or people (or other topics); it does not contain text that provides gender information about who is speaking or who is being spoken to. Additionally, despite writing of four gender values (unknown, neutral, feminine, and masculine), the dataset and classifiers of Dinan et al. (2020) are limited to "*masculine* and *feminine* classes" (p. 317). The authors also do not explain how they define "bias," limiting our ability to draw on their research.

Doughman et al. (2021) provide another gender bias taxonomy that builds on that of Hitti et al. (2019), resulting in overlaps between our taxonomies. However, Doughman et al. (2020) focus on gender stereotypes, while our Taxonomy considers other types of gender biases. Though less explicit in the names of our Taxonomy's labels, we also looked to the descriptions of gender and gender bias from Cao and Daumé III (2021), who point out the limited gender information available in language. The aim of our dataset creation differs from that of Cao and Daumé III (2021), though. They created data that represents trans and gender diverse identities in order to evaluate models' gender biases, specifically looking at where coreference resolution fails on trans and non-binary referents. By contrast, we aim to create a dataset that documents biased representations of gender, with the aim of creating models that are able to identify types of gender bias in language (Chapter 6).

	Title	Biographical / Hist.	Scope & Cont.	Processing Info.	Total
Count	4,834	576	6198	280	11,888
Words	51,904	75,032	269,892	3,129	399,957
Sentences	5,932	3,829	14,412	301	24,474

Table 5.1: **Total counts, words, and sentences for metadata fields' descriptions in the aggregated dataset.** Descriptions are from the "Title," "Biographical / Historical" (Biographical / Hist.), "Scope & Contents" (Scope & Cont.), and "Processing Information" (Processing Info.) metadata fields. Calculations were made using Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002).

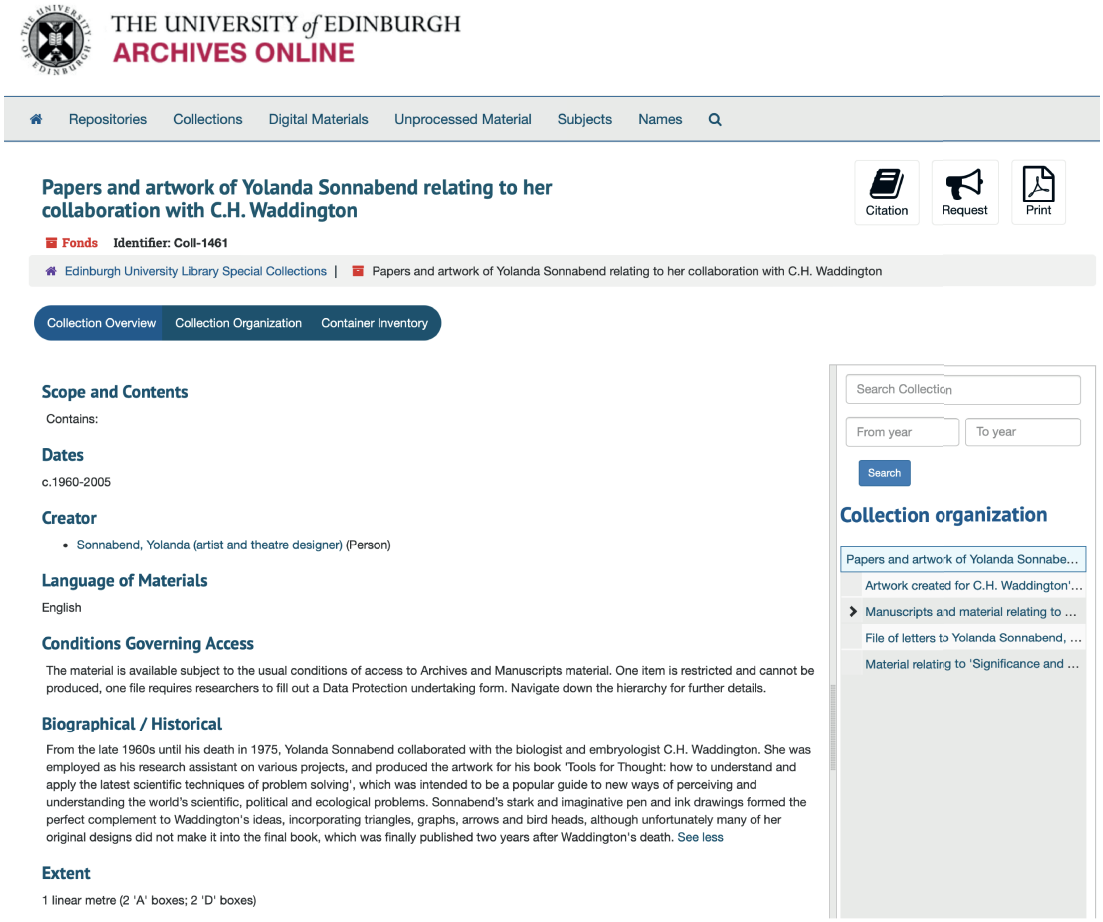


Figure 5.1: **An example of GLAM documentation from the archival catalog of Heritage Collections at the University of Edinburgh (Heritage Collections, 2018).** Metadata field names are in bold, blue text and their descriptions are in regular, black text. The "Title" field's description, however, is bolded in blue at the top ("Papers and artwork of...").

5.1.6 Case Study

To demonstrate the application of the Taxonomy, we present a case study situated in the UK in the 21st century, annotating collection documentation written in British English from the HC Archives' catalog. This paper thus takes the first step in building a collection of case studies that situate NLP gender bias research in a specific context. A collection of case studies would enable the NLP community to determine which aspects of gender bias mitigation approaches generalize across time, location, culture, people, and identity characteristics.

The HC Archives' documentation served as a suitable data source because the documentation adheres to an international standard for organizing archival metadata, ISAD(G) (ICA, 2011), the HC team had found gender bias in the documentation's language, and the HC team were already engaged in efforts to mitigate gender bias in the archival documentation. Furthermore, the documentation is in the public domain: the catalog and its metadata descriptions can be browsed at archives.collections.ed.ac.uk. Figure 5.1 displays an example of the documentation of a collection on this website. The documentation describes a variety of heritage collections and items, such as letters, journals, photographs, degree certificates, and drawings; on a variety of topics, such as religion, research, teaching, architecture, and town planning. Employees at the HC Archives describe themselves as activists changing archival practices to more accurately represent the diverse groups of people that the archival collections are intended to serve.

I created the annotation corpus using the programming language Python version 3.8.10,[4] as well as the Python programming libraries pandas (The pandas development team, 2023) and Natural Language Toolkit (NLTK) (Loper and Bird, 2002). The annotation corpus consists of 24,474 sentences and 399,957 words, selected from the first 20% of the entire corpus of documentation from the HC Archives' catalog (see Appendix A for more on this corpus). Table 5.1 provides a breakdown of the size of the annotation corpus by metadata field. 90% of the annotation corpus (circa 22,027 sentences and 359,961 words) was doubly annotated with all labels, and 10% of the annotation corpus (circa 2,447 sentences and 39,996 words) was triply annotated with all labels. In total, the annotation process amounted

[4] www.python.org

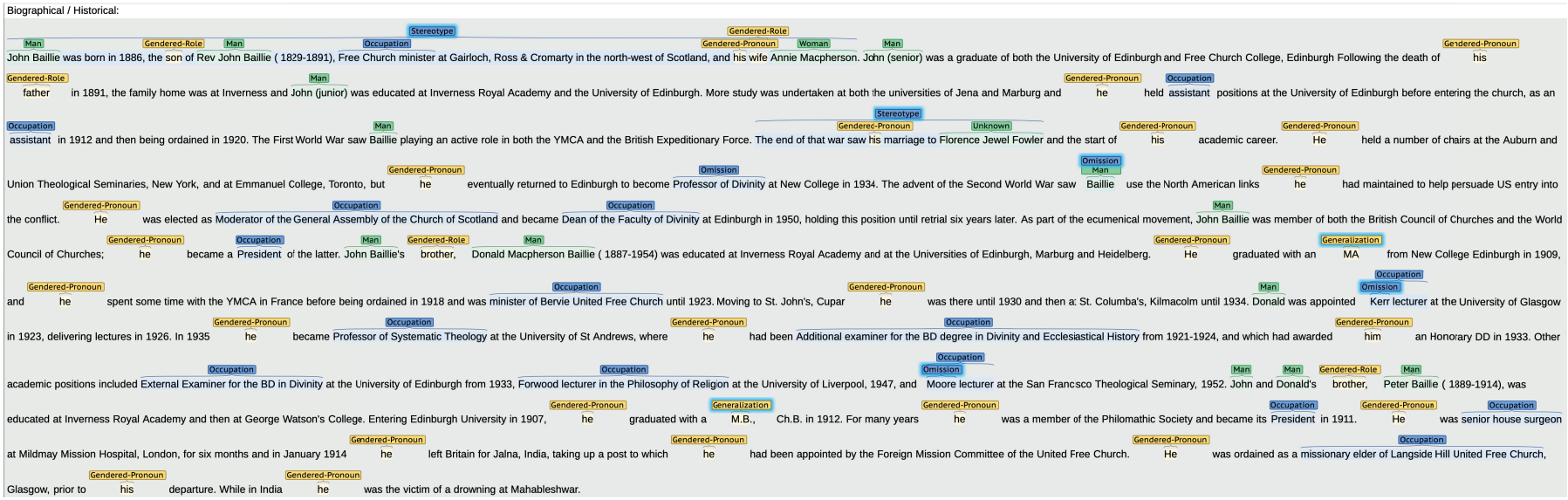


Figure 5.2: **An example of an annotated description.** A "Biographical / Historical" metadata field's description annotated with all labels from the Taxonomy, displayed in brat, the online platform used to perform the annotation work (Stenetorp et al., 2012).

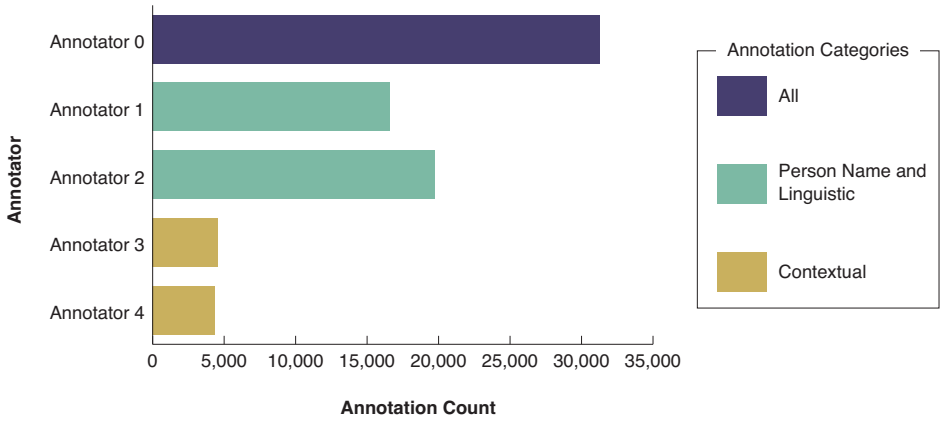


Figure 5.3: **Total Annotations Per Annotator.** The number of annotations each annotator applied to their given subset of archival metadata description. Bars are color-coded based on the Taxonomy categories with which annotators labeled descriptions. In total, annotators made 198,520 annotations.

to circa 400 hours of work and £5,333.76, funded by a variety of internal institutional funds. Each of the four hired annotators worked for 72 hours over eight weeks at £18.52 per hour (minimum wage at the time was £9.50 per hour (Gov.uk, 2022)). The hired annotators were Ph.D. students selected for their experience in Gender Studies or Archives, with three of the annotators having experience in both. As lead annotator (A0), I worked for 86 hours over 16 weeks. Figure 5.2 displays an example of a description in the annotation platform the annotators used to apply the Taxonomy's labels, brat (Stenetorp et al., 2012).[5]

The categories of labels in the Taxonomy were divided among annotators according to the textual relations the labels record. Hired annotators 1 and 2 (A1 and A2) labeled internal relations of the text with ***Person Name*** and ***Linguistic*** categories, hired annotators 3 and 4 (A3 and A4) labeled external relations of the text with the ***Contextual*** category, and the lead annotator (A0) labeled both relations with all categories. A1 and A3 labeled the same subset of archival documentation, and A2 and A4 labeled the same subset of archival documentation, ensuring every description had labels from all categories. The lead annotator labeled the same descriptions as A1 and

[5] brat.nlplab.org/index.html

A3, and a subset of the descriptions that A2 and A4 labeled (due to time constraints, A0 could not label all the same descriptions). Prior to beginning annotation, *Gendered Pronoun*, *Gendered Role*, and *Occupation* labels were automatically applied. The annotators corrected mistakes from this automated process during their manual annotation. Throughout the annotation process, annotators communicated with one another in a Microsoft Teams discussion thread regarding their interpretation of the Taxonomy's labels and the types of linguistic patterns they were annotating. This communication guided iterative adjustments to the annotation instructions and to annotators' labeling, similar to the agile approach described in Alex et al. (2010). Figure 5.3 visualizes the total number of annotations per annotator across their five datasets.

We produced six instances of the annotation corpus: one each with the annotations of A0, A1, A2, A3, and A4, and one aggregated dataset. The aggregated dataset combines annotations from all five annotators, totaling 76,543 annotations with duplicates and 55,260 annotations after deduplication. Manual and programmatic analysis of each annotator's dataset (§5.1.6.1) informed the aggregation approach, which also involved a combination of programmatic and manual steps (§5.1.6.2). The data statement in Appendix E documents the deduplicated, aggregated dataset ("the aggregated dataset") and Appendix F contains additional annotation tables and figures. In line with data perspectivism (Basile, 2022), the individual annotators' datasets will be published alongside the aggregated dataset, enabling researchers to interrogate patterns of agreement and disagreement, and enabling future work to compare the performance of classifiers trained on disaggregated and aggregated datasets. I analyzed and aggregated the annotated data using Python version 3.9.13,[6] pandas (The pandas development team, 2023), and intervaltree.[7]

5.1.6.1 Annotated Data Analysis

After each annotator, including myself, completed their annotations of the archival documentation in brat, I exported the annotation data, which brat provides as ".ann" files, with one file per ".txt" file of archival documentation that I had uploaded. For each annotator, I combined the data across files and

[6] www.python.org
[7] pypi.org/project/intervaltree

	Title		Biog. / Hist.		Scope & Cont.		Proc. Info.	
	count	ratio	count	ratio	count	ratio	count	ratio
Generalization	462	0.224	402	0.195	1193	0.579	4	0.002
Omission	1472	0.194	631	0.083	5480	0.722	3	0.000
Stereotype	421	0.159	482	0.182	1745	0.659	0	0.000

Table 5.2: **Gender biased language annotations by metadata field.** The "count" and "ratio" columns display the number and proportion (e.g. 0.244 = 24.4% of Generalization annotations are in the "Title" metadata field) of annotations across all annotator datasets that occur in each metadata field, either "Title," "Biographical / Historical" (Biog. / Hist.) "Scope and Contents" (Scope & Cont.) or "Processing Information" (Proc. Info.).

transformed them into a tabular format (Table 5.3). To understand the type of language the annotators had labeled as gender biased according to the Taxonomy, specifically with the *Generalization*, *Omission*, and *Stereotype* labels, I conducted analysis on the annotated text spans and annotators' notes for these labels.

To begin, I calculated the occurrences gender biased language annotations. For the calculations by metadata field (Table 5.2), most gender biased language annotations occur in the "Scope and Contents" metadata field and almost none occur in the "Processing Information" field. For the calculations by text span-label pairs (tables 5.4 and 5.5), the results show that gendered language (i.e. text spans that also had *Gendered Pronoun* or *Gendered Role* labels) was commonly annotated across all three types of gender biased language. For *Omission*, however, proper names were also common, calling attention to the way in which men are commonly referenced by their last name only, whereas women are commonly referenced by a feminine title and last name (Table 5.5b). This practice indicates the assumption of the default man, where people of other genders need a qualifier to show that a position or person is held by someone other than a man (e.g. an artist who is a woman being described as a "woman artist" and a man simply being described as an "artist;" the woman is given the qualifier "woman" but the man is not (Hessel, 2023a, 2023c).

While the gender bias of the text spans annotated as *Generalization* or *Omission* are simple to understand from the grammatical gender of the text spans themselves, the gender bias behind the *Stereotype* labels are not clear from the text spans alone. Consequently, I conducted further analysis on the

annotations with this label, reviewing the text spans alongside the comments annotators made about their rationale for each annotation with the *Stereotype* label. First, I combined and deduplicated the annotators' comments into one table with columns for the comment and associated text spans, label, and annotation IDs. Next, I grouped similar comments. For example, I grouped the annotator comments "man associated with technology, operating machinery" and "man associated with male-dominated disciplines" under a category I named "Association of men with specific disciplines, professions, subjects." This resulted in 23 categories of stereotypes covered with the Taxonomy's *Stereotype* label. Table 5.6 displays the categories with examples of annotator comments for each category, as well as the annotated text span associated with each comment.

My analysis of gender biased language that annotators labeled in HC Archives documentation provides a starting point for collection reviews in the GLAM sector and for detecting gender bias in text corpora in the NLP community. The 23 categories offer the HC Archives[8] and wider GLAM sector an understanding of the types of language that may be considered stereotypical, informing collection reviews and guidelines for descriptive practices. For the NLP community, the categories offer a starting point for defining types of gender stereotypes that can inform future efforts to annotate text corpora for gender biased language. I recommend that annotators be given a set of categories to use to explain the rationale for their labels while also being given the opportunity to define their own categories. The gender stereotypes in my dataset are not comprehensive of all gender stereotypes that may occur in English text, so annotators should be permitted to define categories in response to the text they read and based on their own experiences.

[8]During the evaluation workshop reported in Chapter 7, participants' questions communicated an interest in gaining this understanding.

file	entity	label	start	end	text	note
Coll-1326_00100.ann	T13	Stereotype	1555	1599	President of the Royal College of Physicians	honour or achievement held by man
Coll-1320_01900.ann	T1	Occupation	657	670	Embryologists	
Coll-1320_01900.ann	T2	Occupation	1836	1857	Chief Medical Officer	
Coll-1143_00100.ann	T5	Stereotype	1359	1391	Faculty of Law Class Merit Lists	honour or achievement held by man
Coll-1287_00100.ann	T2	Occupation	172	180	Clergyman	
Coll-1287_00100.ann	T3	Omission	539	551	Prince Reuss	person not named - referred to by gendered role (male)
Coll-1287_00100.ann	T4	Omission	590	609	General Montesquieu	only title (gender neutral) and family name given
Coll-1287_00100.ann	T7	Omission	687	695	von Berg	only family name given

Table 5.3: **Sample Annotator Data.** Sample of the output annotation data from the annotation platform brat, converted from the ".ann" file format to a tabular format. The "file" column refers to the file of annotations exported from brat (Stenetorp et al., 2012), "entity" is a brat identifier (unique per file), "label" is the label from the Taxonomy the annotator applied, "start" and "end" refer to the starting and ending positions of the text span that was annotated, "text" is the annotated text span from the archival documentation, and "note" is the annotator-written comments about the annotation.

text	occurrence	label
Thomson	981	Omission
man	566	Generalization
man	429	Stereotype
Ledermann	351	Omission
men	342	Stereotype
a man	286	Omission
woman	246	Generalization
a man	223	Stereotype
Beale	220	Omission
Beatty	146	Omission

Table 5.4: **Top ten text spans annotated as gender biased.** The "text" column lists the top ten text spans annotated with a *Generalization*, *Omission*, or *Stereotype* label in descending order. The "occurrence" column lists the total count of each text-label pair across the five annotators' datasets. The "label" column lists the labels that annotators applied to the text spans.

text	occurrence
man	566
woman	246
boy	41
he	36
Thomson	35
his	34
boys	34
Midwifery	31
MA	25
Empress	23

(a) *Generalization*

text	occurrence
Thomson	981
Ledermann	351
a man	286
Beale	220
Beatty	146
Lady Thomson	79
two men	77
men	77
group of men	43
Thurstone	38

(b) *Omission*

text	occurrence
man	451
men	357
a man	223
a woman	108
woman	67
women	61
two men	54
a group of men	32
female	24
boys	21

(c) *Stereotype*

Table 5.5: **Top ten text spans annotated per gender biased language label.** From left to right, the top ten text spans ("text" column) and their occurrence with the associated label ("label" column) for the *Generalization*, *Omission*, and *Stereotype* labels. The counts in the "occurrence" columns are based on text span counts that are not case sensitive (e.g. "man" and "Man" are listed together in a row for "man").

Table 5.6: **Stereotype Examples from the Annotated Data.** The "category" column lists the categories of stereotypes I defined after reviewing the text spans manually annotated with the *Stereotype* label; there are 23 total categories (this table continues onto the next page). The "annotated text example" column gives an example of an annotated text span from the archival metadata descriptions representing that row's *Stereotype* category. The "annotator note" column displays the comments the annotator who labeled the example text span provided with their annotation.

ID	category	annotated text example	annotator note
1	Association of men with specific disciplines, professions, subjects	This series contains cuttings from British newspapers that relate, in the widest possible sense, to the life and works of Sir Walter Scott	materials described as relating to walter scott's life in widest sense, no mention of domestic or familial matters, only work
2	Association of women with domestic life, not professional career	As a young married woman with two small children - Kathleen and Margaret - she found it difficult to keep her medical career going	blaming difficulty keeping career going on being a young woman with children
3	Demeaning, derogatory language	Woman Hater	derogatory language
4	Devaluing women's labor	Jewel took an active interest in her husband's work, accompanying when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more	female labour on behalf of husband goes unpaid and not properly credited
5	Devaluing women's needs	Young was the first Professor of Midwifery at Edinburgh to actually lecture on the subject of obstetrics.	obstetrics (pregnancy, childrearing) had been skipped over even in subject entirely about giving birth - what on earth had midwifery covered before???
6	Distracting from, doubting, or minimizing women's achievements	acted as a	a woman "acted as" rather than was a clinical clerk, despite all the qualifications listed for her

Continued on next page

Table 5.6 – continued from previous page

ID	category	annotated text example	annotator note
7	Excessive feminization	Instructresses	excessive feminization, assumption of male defaults
8	Expectation of or judgment on woman's interests	not normally associated	expectation of what a young lady of this period would read
9	Imbalanced attitudes for terminology associated with woman/man	Bad boy of biology	likened to a "bad" child has positive connotations for men (a young and successful man who does things in a way that is very different from the usual way), while for women, it has negative connotations - sexual impropriety
10	Infantilization of women	girls	graduates described as "girls" instead of "women"
11	Man active, woman passive	Koestler's marriage to Mamaine	husband represented as the active party in marrying. alternative: mamaine and arthur's marriage
12	Man as default	Diseases of Women	assumption of male as default - women's diseases need their own subject, but not men's diseases
13	Man as leader	leading Scottish Theologian	man associated with leadership role
14	Man as owner	workman's home	house as property of man alone
15	Man's value described without accompanying evidence	the most influential anatomy professor in the English speaking world	very positive claim made about a man without any evidence being provided to back it up
16	Men commemorated with prestigious positions, honors	who founded the southern African territory of Rhodesia, which was named after him	honour or achievement held by man
17	Men listed before women	John Baillie and Florence Jewel Fowler	man listed first before woman

Continued on next page

Table 5.6 – continued from previous page

ID	category	annotated text example	annotator note
18	Not inclusive of trans and gender diverse identities	two unidentified men	photograph description assumes unknown person's gender (man/men/boy)
19	Primacy of men in family, society	had been put forward by a group of prominent Scotsmen in March 1812. The fund would protect 'widows, sisters and other females' from poverty	men providing for women
20	Wife playing supporting role to husband	with whom he had another son	mother given a subordinate role in reproduction. non-biased alternative: "they had a son"
21	Woman as property or vessel	not only intend to give me a church but a wife also	woman described as object
22	Woman reduced to physical appearance, sexuality, reproductive capability	Virgin	woman referenced by sexuality - "virgin" (biblical)
23	Gendered language as explicitly noted as from archival material	The certificate was designed for a male student and the word 'he' has not been amended	gendered language noted by archivist

5.1.6.2 Annotated Data Aggregation

I aggregated the individual annotator datasets through a combination of automated and manual reviews.[9] Data perspectivism (Basile, 2022) inspired my nuanced approach to data aggregation. Recognizing that my Taxonomy and annotation instructions could be interpreted in multiple ways, I wanted to ensure that I only excluded annotations from the aggregated dataset if they directly contradicted the annotation instructions. In this way, the aggregated data represents all annotators' perspectives on gendered and gender biased language.

To aggregate the data, first I manually reviewed annotations for mistakes, such as a labeled text span that accidentally excluded the first letter of a labeled word, or a labeled text span that accidentally included the ending punctuation of a sentence. I created lists of valid text spans for the *Gendered Pronoun* and *Occupation* labels to determine which annotations with these labels should be included in the aggregated dataset. The lists of valid text spans were created after generating lists of unique text spans to which annotators applied one of these labels. I manually corrected the lists of unique text spans, removing mistaken or incorrect text spans (for example, a person's name being labeled as a *Gendered Pronoun*, or the word "lecturing" being labeled as an *Occupation*). I compared the annotators' text spans annotated with *Gendered Pronoun* and *Occupation* labels against these lists; only text spans that matched a value in the lists were added to the aggregated dataset.[10]

Next, I manually reviewed the 97,861 disagreeing annotations (Figure 5.4), defined as annotations with the same or overlapping text spans but different labels, and added the correct annotations to the aggregated dataset. This was the most time-consuming task of the entire manual review process. For the *Stereotype* and *Omission* labels, I added all annotators' annotations to the aggregated dataset. For the remaining labels in the Taxonomy, I deemed only one annotation correct based on the annotation instructions (see Appendix D), and then added that annotation to the aggregated dataset. The *Gendered Role* and *Generalization* labels proved particularly difficult to distinguish, as all three

[9] See the code for aggregating the data at: github.com/thegoose20/annot/tree/main/notebooks/aggregating_data.

[10] The code for correcting annotation mistakes is available at: github.com/thegoose20/annot/blob/main/notebooks/aggregating_data/MergeData2_GPronouns_Occupations.ipynb.

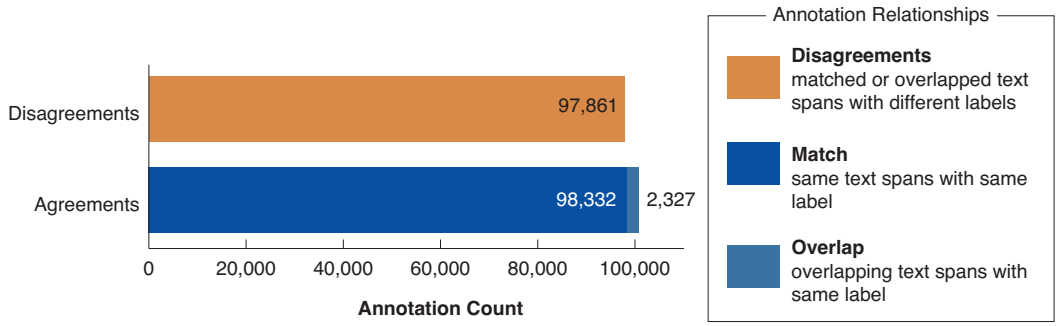


Figure 5.4: **Total Annotations Manually Reviewed.** The number of disagreeing and agreeing annotations ("Disagreements" and "Agreements," respectively) across the five individual annotator datasets that I manually reviewed to determine which annotations to keep in the aggregated dataset. Agreeing annotations are subdivided into annotations from different annotators that label the exact same text spans ("Match") and annotations from different annotators that label different but overlapping text spans ("Overlap") with the same label.

annotators who used this label applied it inconsistently. Consequently, I wrote a new definition of *Generalization* and applied during the aggregation process to more clearly distinguish it from the *Gendered Role* label.[11]

I added the 100,659 agreeing annotations (Figure 5.4) to the aggregated dataset next. Due to the greater importance of recognizing the presence of gendered or gender biased language in a description compared to labeling the exact same text span as gendered or gender biased, I considered overlapping annotations with the same label to be in agreement, in addition to exact matches, where an annotation had the same text span and same label. Among the 2,327 overlapping annotations, I chose the annotation with the longest text span in each group of overlaps automatically (using text mining approaches with intervaltree and pandas) to add to the aggregated dataset. Then, I automatically identified and added the 98,332 matching annotations (using text mining approaches with intervaltree and pandas) to the aggregated dataset. All remaining annotations were then added to the aggregated dataset, with the exception of one hired annotator's ***Person Name*** labels. That annotator's ***Person Name*** labels were applied inconsistently and thus were excluded from the aggregated dataset, unless they matched another

[11]The code for reviewing disagreeing annotations is available at: github.com/thegoose20/annot/blob/main/notebooks/aggregating_data/MergeData1_Disagreements.ipynb.

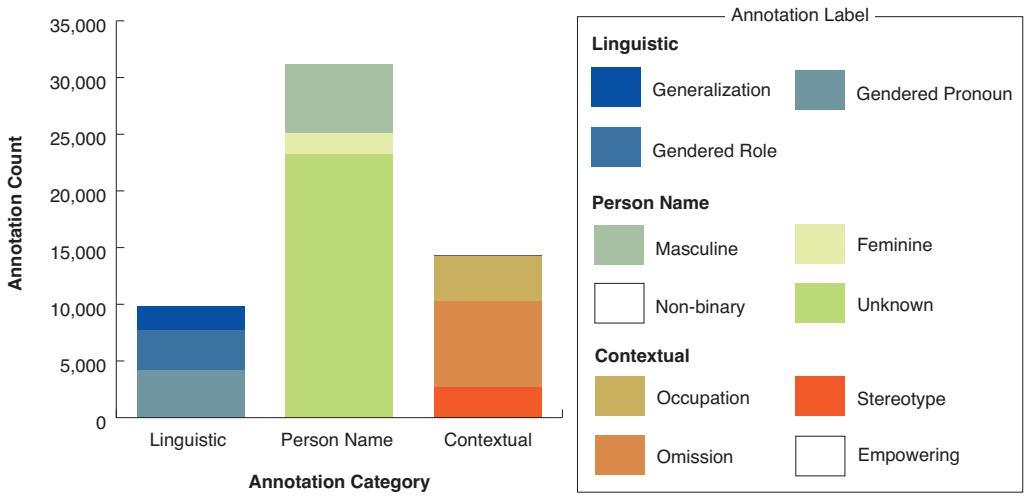


Figure 5.5: **Total Annotations Per Label in the Aggregated Dataset.** The stacked bar chart groups annotation labels into bars by category. Across all three categories, there are 55,260 annotations in the aggregated dataset. *Non-binary* (a ***Person Name*** label) and *Empowering* (a ***Contextual*** label) both have a count of zero.

annotator's annotations. I deduplicated the final aggregated dataset so that any matching annotations from different annotators appear only once in the dataset.[12,13] Figure 5.5 visualizes the number of annotations in the aggregated dataset by Taxonomy category and label (see Table E.1 for precise counts).

5.1.6.3 Inter-Annotator Agreement

Due to the greater importance of recognizing the presence of gendered or gender biased language in a description compared to labeling the exact same words as gendered or gender biased, identifying strictly matching text spans that annotators labeled was deemed less important than the presence of a label in a description. Consequently, Inter-Annotator Agreement (IAA) calculations consider annotations with the same label to agree if their text spans match or overlap. The IAA metrics, calculated between pairs of annotators (one designated as "expected" and the other as "predicted"), are:

[12]The code for reviewing agreeing annotations is available at: github.com/thegoose20/annot/blob/main/notebooks/aggregating_data/MergeData3_AgreementsAndRemaining.ipynb.

[13]The aggregated dataset with all annotators' labels and notes, including duplicates, and the final, deduplicated aggregated dataset are available at: github.com/thegoose20/annot

- **True Positive count**, or True Positives (TP), which records the total number of times the expected labels agree with predicted labels
- **False Positive count**, or False Positives (FP), which records the total number of predicted labels that were not in the expected labels
- **False Negative count**, or False Negatives (FN), which records the total number of expected labels that were not in the predicted labels
- **Precision**, which is calculated as the ratio of TP to the sum of TP and FP, measuring the proportion of correctly predicted labels among all predicted labels

$$\frac{tp}{tp+fp} \tag{5.1}$$

- **Recall**, which is calculated as the ratio of TP to the sum of TP and FN, measuring the proportion of correctly predicted labels among all expected labels

$$\frac{tp}{tp+fn} \tag{5.2}$$

- **F_1 score** (van Rijsbergen, 1979), which combines precision and recall into one metric, calculated as their harmonic mean

$$2 \times \frac{precision \times recall}{precision + recall} \tag{5.3}$$

I selected these metrics as they are standard metrics for text classification that can be used to measure agreement between manual annotators *and* to measure the performance of a classification model relative to manual annotators' labels (Eisenstein, 2018; Jurafsky and Martin, 2023). Figure 5.6 visualizes three archival metadata descriptions with all five of the annotators' labels, illustrating example agreements and disagreements. Figures 5.7 and 5.8 display F_1 scores for each label, with the aggregated dataset as having the predicted labels and the individual annotators' datasets as having the expected labels. In Appendix F, Tables F.1 and F.4 list TP, FP, and FN, as well as precision, recall, and F_1 scores, for IAA among the annotators and with the aggregated dataset.

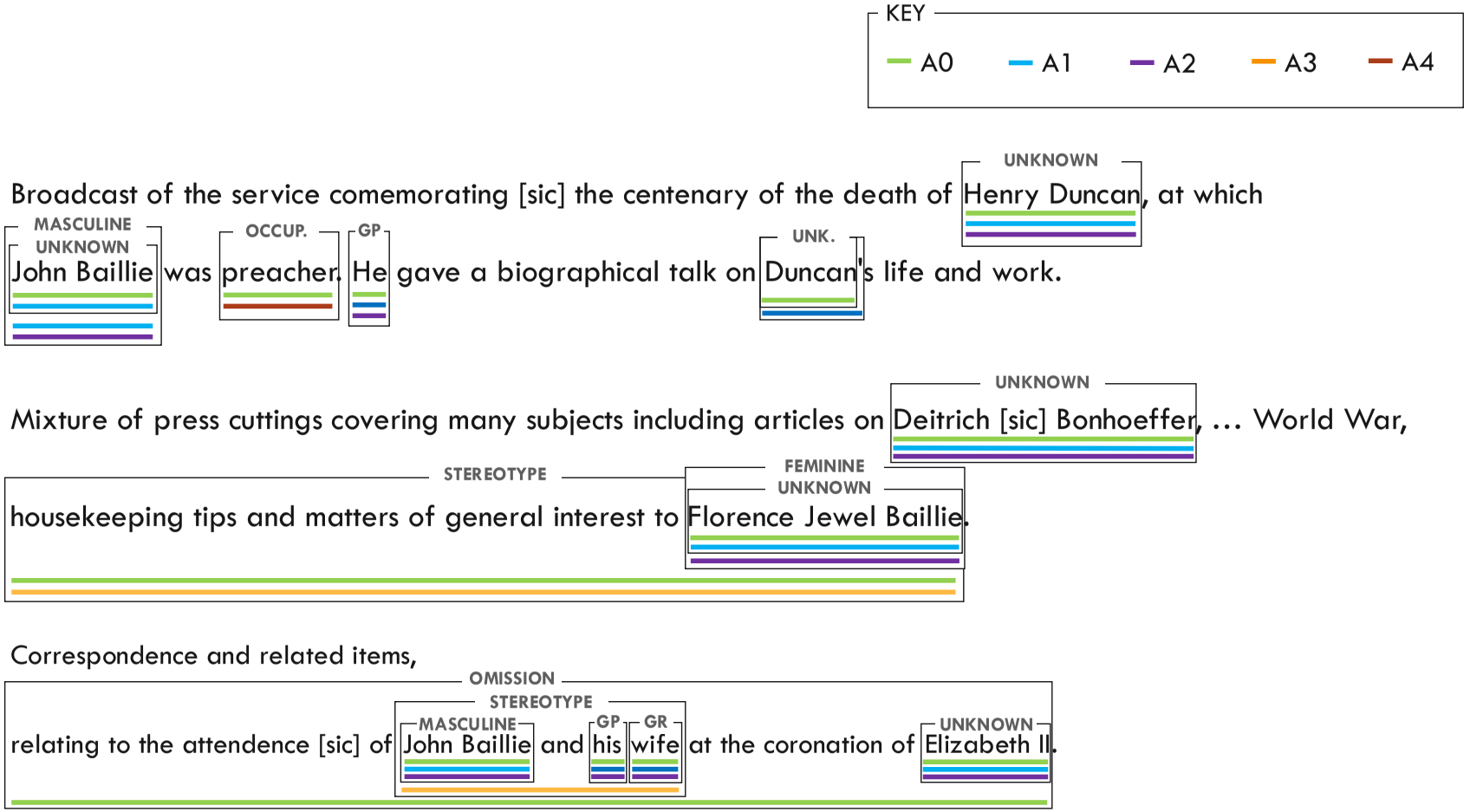


Figure 5.6: **Visualization of Annotator Agreements and Disagreements with Example Descriptions.** Manual annotators' labels on three example descriptions are visualized using color-coded underlining to represent each annotator, and labeled boxes to indicate the span of the annotation. I as annotator 0 (A0) annotated with all the Taxonomy's labels, annotators 1 and 2 (A1 and A2) were instructed to annotate with the Taxonomy's ***Linguistic*** and ***Person Name*** labels only, and annotators 3 and 4 (A3 and A4) were instructed to annotate with the Taxonomy's ***Contextual*** labels only.

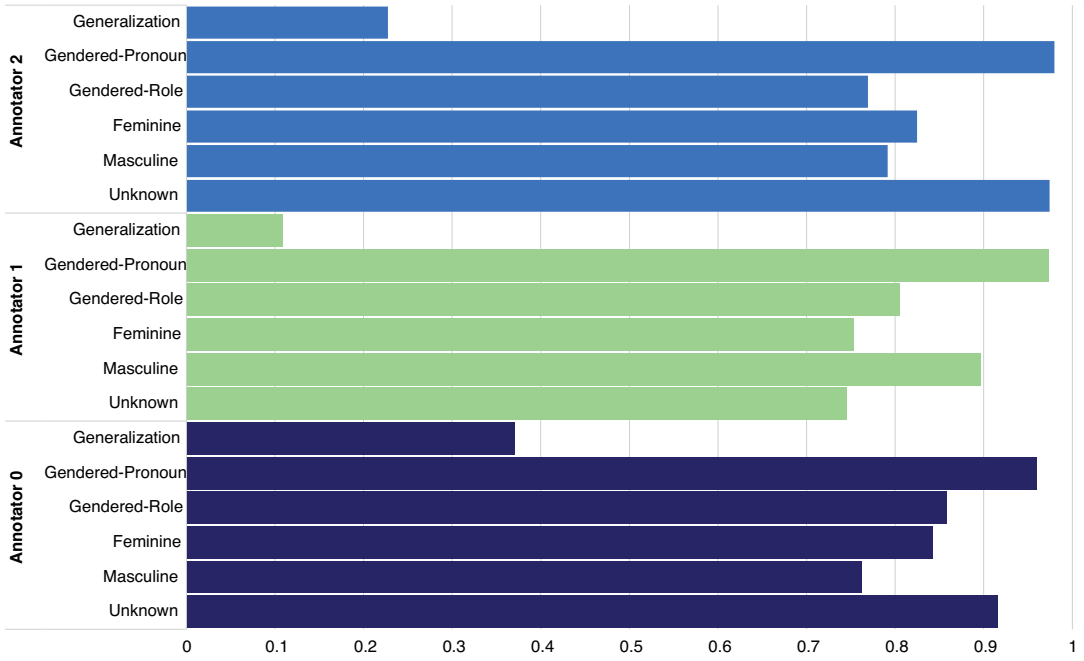


Figure 5.7: ***Person Name* and *Linguistic* F_1 scores for annotators' agreement with the aggregated dataset.** F_1 scores (X axis) are calculated with the aggregated dataset's labels as expected labels and the annotators' (Y axis) labels as predicted labels. Annotators did not use the *Non-binary* label (from the ***Person Name*** category) so it is not in the aggregated dataset.

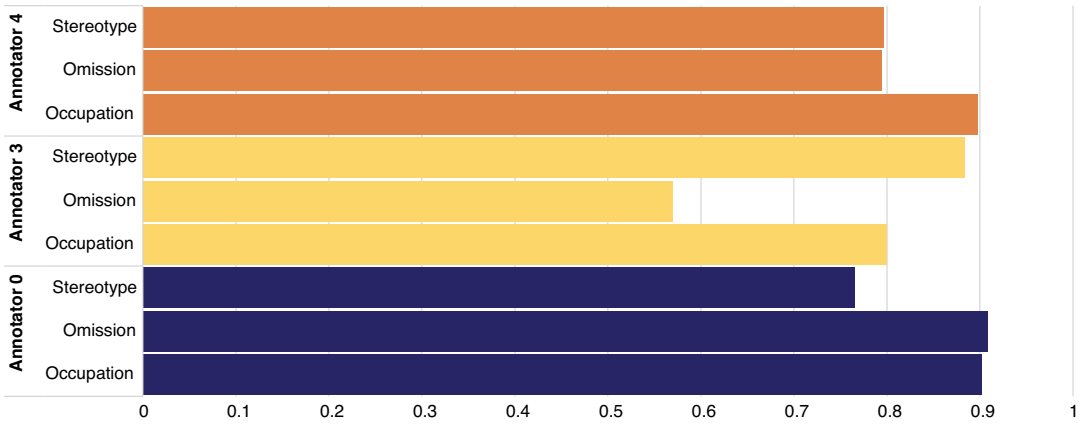


Figure 5.8: ***Contextual* labels' F_1 scores for annotators' agreement with the aggregated dataset.** F_1 scores (X axis) are calculated with the aggregated dataset's labels as the expected labels and each annotator's (Y axis) labels as predicted labels. Annotators did not use the *Empowering* label as defined in the annotation instructions, so it is not in the aggregated dataset.

IAA calculations reflect the subjectivity of gender bias in language. F_1 scores for the gendered language labels *Gendered Role* and *Gendered Pronoun* fall between 0.71 and 0.99. F_1 scores for annotating gender biased language are relatively low, with the greatest agreement on the *Generalization* label at only 0.56, on the *Omission* label at 0.48, and on the *Stereotype* label at 0.57. For ***Person Name*** labels, A0 and A2 agree more than A1: A0 and A2's F_1 scores for all ***Person Name*** labels are between 0.82 and 0.86, while A1's scores with either A0 or A2 are between 0.42 and 0.64. A1 has a particularly high false negative rate for the *Unknown* label compared to A0.

After creating the aggregated dataset, we calculated IAA between each annotator and the aggregated dataset. F_1 scores for all ***Person Name*** and ***Linguistic*** labels except *Generalization* are similarly high (0.74 to 0.98). *Generalization* proved particularly difficult to label. Annotators used *Generalization* and *Gendered Role* inconsistently. As a result, during the aggregation process, we revised the definition of *Generalization* to more clearly distinguish it from *Gendered Role*. Consequently the IAA between annotators and the aggregated dataset for this label is particularly low (0.1 to 0.4).

For ***Contextual*** labels, F_1 scores with the aggregated dataset as "expected" and an annotator as "predicted" increased more dramatically than the ***Person Name*** and ***Linguistic*** labels' F_1 scores. Besides *Omission* with A3, all F_1 scores are between 0.76 and 0.91. For *Stereotype*, A3 agreed more strongly with the aggregated dataset than A0 and A4. The reverse is true for *Omission* and *Occupation*, with A0 and A4 agreeing more strongly with the aggregated dataset than A3. A3's notes explain that she did not annotate an incomplete version of a person's name as an *Omission* if the complete version was provided elsewhere in the collection's descriptions, whereas A0 and A4 annotated incomplete versions of people's names as *Omission* unless the complete version appeared in the same description.

Two labels were not applied according to the Taxonomy's definitions: *Empowering* and *Non-binary*. A3 used *Empowering* according to a different definition than that of the Taxonomy (Appendix E). As only 80 instances of the label exist in A3's dataset, though, there are likely to be insufficient examples for effectively training classifiers on this label. No annotators used the *Non-binary* label. That being said, this does not mean there were not people who would identify as non-binary represented in the text of the

annotation corpus. Additional linguistic and historical research may identify people who were likely to identify as non-binary in the corpus of archival documentation, as well as more specific gender identities for people whose names were annotated as *Masculine* or *Feminine*. Metadata entries for people in the HC Archives' catalog may also provide more information relevant to gender identities. Shopland (2020) finds that focusing on actions that people were described doing can help to locate people of minoritized genders (and sexualities) in historical texts. However, Shopland also cautions researchers against assuming too much: a full understanding of a person's gender often remains unattainable from the documentation that exists about them.

As Figure 5.5 displays, *Unknown* is the most prevalent label in the ***Person Name*** category, because each annotation of a person's name was informed by words within the description in which that name appears. Consequently, for people named in more than one description, there may be different ***Person Name*** labels applied to their name across those descriptions. The rationale for this approach comes from the aim to train document classification models on the annotated data where each description serves as a document. Should a person change their gender during their lifetime, and archival documentation exists that describes them as different genders, the person may wish a model to use the most recent description of a person to determine their gender, or not use any gender information about the person, in case obviating their change of gender leads to safety concerns (Dunsire, 2018). Furthermore, many GLAM content management systems do not have versioning control, so dates of descriptions may not exist to determine the most recent description of a person's gender. ***Person Name*** labels are thus based on the description in which a name appears to minimize the risk of misgendering (Scheuerman, Spiel, et al., 2020).

5.1.7 Discussion

Our annotation Taxonomy builds on biased language research from NLP, GLAM, Gender Studies, and Linguistics literature. The gender bias taxonomy of Hitti et al. (2019), which categorizes gender biases based on whether the bias comes from the sentence structure or the context (e.g. people, relationships, time period, location) of the language, served as a foundation. We adopted four

labels from that taxonomy: *Gendered Pronoun*, *Gendered Role*, *Generalization*, and *Stereotype* (merging Hitti et al.'s Societal Stereotype and Behavioral Stereotype categories). Drawing on Archival Science and critical discourse analysis, and guided by PAR with archivists (e.g. interviews, workshops), we added to and restructured Hitti et al.'s taxonomy. The ***Person Name*** labels were added so that the representation of people of different genders in the archival documentation could be estimated. Annotators chose which label to apply to a person's name based on gendered pronouns or roles that refer to that person in the description in which their name appears. For example, "they" as singular for *Non-binary*, "his" for *Masculine*, and "she" for *Feminine*; or "Mx." for *Non-binary*, "Lady" for *Feminine*, or "son" for *Masculine*. The *Unknown*, *Feminine*, and *Masculine* labels distinguish our approach from previous NLP gender bias work that has not allowed for uncertainty.

Guessing a person's gender risks misgendering (Scheuerman, Spiel, et al., 2020), a representational harm (Blodgett et al., 2020; Crawford, 2017), and fails to acknowledge that sufficient information often is not available to determine a person's gender with certainty (Shopland, 2020). This led us to replace the initial labels of *Woman* and *Man* with *Feminine* and *Masculine*, recognizing that pronouns and roles are insufficient for determining how people define their gender. Each ***Person Name*** label encompasses multiple genders. For instance, a person who identifies as a transwoman, as genderfluid, or as a cis woman may use feminine pronouns, such as "she," or feminine roles, such as "wife." Though we aimed to create a taxonomy inclusive of all genders, we acknowledge this may not have been achieved, and welcome feedback on how to represent any genders inadvertently excluded.

We also added three labels to the ***Contextual*** category: *Occupation*, *Omission*, and *Empowering*. *Occupation* was added because, when combined with historical employment statistics, *Occupation*-labeled text spans could inform estimates of the representation of particular genders within the HC Archives' collections, as well as contribute to studies of occupational gender stereotypes (M. Lewis and Lupyan, 2020). Furthermore, ***Person Name*** annotations combined with *Occupation*s could guide researchers to material beyond the HC Archives that may provide information about those people's gender identity. *Omission* was added because, during a PAR workshop with 11

members of the HC team,[14] participants described finding gender bias through the lack of information provided about women relative to the detail provided about men. *Empowering* was added to account for how communities reclaim certain derogatory terms, such as "queer," in a positive, self-affirming manner (Bucholtz, 1999).

Figure 5.5 displays how prevalent *Omission* was in the annotated data: this label is the most commonly applied label from the ***Contextual*** category. Such prevalence demonstrates the value of interdisciplinary collaboration and stakeholder engagement, carried out in our PAR with domain experts. Had HC employees not been consulted, we would not have known how relevant omitted information regarding gender identities would be to identifying and measuring gender bias in archival documentation. Omissions hold relevance to not only to Archives, but also to other GLAM institutions (Hessel, 2023b; Ortolja-Baird and Nyhan, 2022).

The final Taxonomy includes labels for gendered language (specifically, *Gendered Role*, *Gendered Pronoun*, and all labels in the ***Person Name*** category), rather than only explicitly gender biased language (specifically, *Generalization*, *Stereotype*, and *Omission*), because measuring the use of gendered words across an entire archival collection or catalog provides information about gender bias at the overall collections level. For example, using the gendered pronoun "he" is not inherently biased, but if the use of this masculine gendered pronoun far outnumbers the use of other grammatically gendered pronouns in our dataset, we can observe that the masculine is over-represented, indicating a masculine bias in the HC Archives' collections overall. Labeling gender biased language focuses on the individual description level. For example, the stereotype of a wife playing a supporting role to her husband comes through in this description: "Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more."[15]

Instructions for applying the Taxonomy permitted labels to overlap as each annotator saw fit, and asked annotators to annotate from their contemporary perspective (Appendix D). Approaching the archival documentation as

[14]The workshop received ethical approval from the School of Informatics at the University of Edinburgh (reference 2019/81479).

[15]archives.collections.ed.ac.uk/repositories/2/archival_objects/2115

discourse (meaning language as representations of the material, mental, and social worlds (Fairclough, 2003)), the Taxonomy of labels represents the "internal relations" and "external relations" of the descriptions (ibid., 37). The ***Person Name*** and ***Linguistic*** categories annotate internal relations, meaning the "vocabulary (or 'lexical') relations" (ibid., 37) of the descriptions. To apply their labels, annotators looked for the presence of particular words and phrases (e.g. gendered pronouns, gendered titles, familial roles).

The ***Contextual*** category annotates external relations: relations with "social events...social practices and social structures" (Fairclough, 2003, p. 36). To apply ***Contextual*** labels, annotators reflected on the production and reception of the language in the archival documentation. For instance, to apply the *Stereotype* label, annotators considered the relationship between a description's language with social hierarchies in 21st century British society, determining whether the term or phase adequately represented the possible gender diversity of people being described.

5.1.8 Conclusion

This paper has presented a Taxonomy of Gendered and Gender Biased Language with a case study to support clarity and alignment in NLP gender bias research. Recognizing the value of clearly defined metrics for advancing bias mitigation, the Taxonomy provides a structure for identifying types of gender biased language at the level they originate (words and phrases), rather than at a level of abstraction (i.e. vector spaces). Still, the case study presented in this paper demonstrates the difficulty of determining people's gender with certainty. While recognizing the value of NLP systems for mitigating harms from gender biased language at large scale, we contend that conceptualizations of gender must extend to trans and diverse gender expressions if NLP systems are to empower minoritized gender communities.

Future work will include a publication of the case study's datasets[16] with classification models trained on the datasets.[17] The datasets will include each individual annotator's dataset and two aggregated datasets, one with duplicates across different annotators, and one deduplicated to

[16]The disaggregated and aggregated datasets are now available on the University of Edinburgh's DataShare platform at: https://doi.org/10.7488/ds/7540.

[17]See Chapter 6.

exclude matching and overlapping annotations from different annotators. The evaluation of models trained on the datasets will be informed by PAR, incorporating stakeholders' perspectives on the models' annotations.[18] The datasets will be made available in the same location as the code written to create the corpus of archival documentation and the annotated datasets.[19] The Taxonomy and forthcoming datasets aim to guide NLP systems towards measurable and inclusive conceptualizations of gender.

[18] See Chapter 7.
[19] github.com/thegoose20/annot

5.2 Comments on the Paper

5.2.1 Generalizing the Taxonomy and Datasets

In addition to informing the HC Archives team's understanding of the language of their descriptive metadata, the Taxonomy and annotated data hold value for the wider NLP and GLAM communities. The Taxonomy can be applied to other text-based data, though variations in conceptualizations of gender bias between and within cultures (Awad et al., 2018) may lead people to *apply* the Taxonomy's labels differently than my hired annotators and I did. The annotated data offers insight on how gender biases manifest in catalog metadata descriptions written in British English, as well as providing insight on the ways in which biased language can be communicated in English text more broadly. However, the annotated data should not be seen as a comprehensive representation of all varieties of gender biases that may be found in English, or even British English, GLAM catalogs or other text corpora. Rather, the annotated data provides a starting point for developing an understanding of the ways in which English text communicates biases.

5.2.2 Recalibrations with the Taxonomy and Datasets

My approach to creating the Taxonomy and annotated datasets prioritized:

- ***Quality* over quantity** by employing a small number of annotators with domain expertise to label subsets of archival documentation, rather than hiring hundreds of crowdworkers to label larger subsets (or all) of the data. Hiring only four annotators enables careful training and evaluation of annotators' understanding of the Taxonomy, as well as iterative refinement of my bias of focus through dialogue with the hired annotators.

- ***Accuracy* over efficiency** by creating a manually annotated dataset for training models through a supervised learning approach, rather than using an unsupervised learning approach. Manually annotating words and phrases in the archival documentation ensures that the context in which the descriptions are read is taken into account to guide its subsequent annotation.

- ***Representativeness* over convenience** by creating a bespoke taxonomy and datasets for my case study, rather than using existing taxonomies of gender biased language or existing datasets of English text. Creating my own Taxonomy and datasets took more resources (i.e. time, funding, people) but ensured that definitions of types of gender bias in the Taxonomy, as well as the instances of gender biased language annotated according to the Taxonomy, are relevant to the case study in which I situate my research.

- ***Situated thinking* over universal thinking** by presenting the annotated datasets as the output of a case study, specifically a case study of the HC Archives documentation, which are written in British English and annotated for gender biases in the 21st century in the UK. This presentation enables researchers to approach the Taxonomy and datasets critically, considering for how their research context compares and contrasts with mine to determine how best to apply or build upon the Taxonomy and datasets.

Chapter 6 uses the aggregated dataset contributed in this chapter as training, validation, and test data to create gender biased text classification models.

Chapter 6

Gender Biased Text Classification

> I do not wish to ban or cancel images that show women being exploited...I am interested that we notice them.
>
> –Mary Beard, interview with Katy Hessel (2022)

Making use of the aggregated dataset from Chapter 5, this chapter describes my next contribution: text classification models trained to identify gendered and gender biased language, where the classes that the models assign to text from archival documentation are the labels from the Taxonomy (§6.5). I report on experiments conducted with text classification models, as well as sequential combinations, or *cascades*, of models that detect gender biased language (§6.6). The research question investigated with this chapter is: ***Can gender biased language be reliably annotated by domain experts to train a classification model to automatically annotate gender biased language?*** The performance of the models, measured using precision, recall, and F_1 scores, indicates that text classification models can be trained to annotate gender biased language, though performance evaluations indicate that different model setups are best for different types of gender biases. Chapter 7 contributes an additional, human-centered approach to model evaluation.

The datasets used to train, validate, and test the classification models reported in this chapter can be downloaded from the University of Edinburgh's DataShare platform at: doi.org/10.7488/ds/7539. The code

for creating the classification models is available at: github.com/thegoose20/gender-bias-models.

Publication Note: A shorter version of the classification work reported in this chapter was first published and presented at the Digital Humanities Conference (Havens et al., 2023; Appendix K). I wrote the publication as lead author, with my supervisors and Rachel Hosker (University Archivist and Research Collections Manager at the University of Edinburgh) providing feedback to inform my revisions to the paper. I conducted the research (i.e. the literature review, data analysis and transformation, and model training, testing, and evaluation) reported in that paper and in this chapter.

6.1 Introduction

Two gaps in existing research on bias in Natural Language Processing (NLP) motivate this chapter. The first gap is the lack of adequately complex conceptualizations of bias. The NLP community has not reached consensus on how gender biases manifest in language, yet many NLP research publications report on the minimization or removal of bias. Most often, efforts to minimize or remove biased language focus on word embeddings (Bordia and Bowman, 2019; Husse and Spitz, 2022; Sun et al., 2019), which abstract language to numeric vectors that represent a word's meaning based on its surrounding words (Jurafsky and Martin, 2023). However, research has not proven that minimizing or removing bias in word embeddings results in the minimization or removal of bias in downstream tasks, such as text classification (Goldfarb-Tarrant et al., 2021; Jin et al., 2021; Steed et al., 2022). Moreover, word embeddings only model "internal relations" of a text: the relationship between words explicitly written or spoken, and how the ordering of those words create meaningful utterances (Fairclough, 2003). Approaching language through the lens of critical discourse analysis (§3.3) requires a consideration of more than these internal relations to study the meaning of a text.

The meaning of a text also comes from "external relations," which are not represented in word embeddings. External relations refer to the context of language, such as the relationship between people producing and receiving the language, the time period and location in which the text is produced, and the cultural and political environment in which the language is produced (Fairclough, 2003). Though approaches to bias mitigation in NLP that do not focus on word embeddings exist (e.g. model fine-tuning (Orgad et al., 2022), or augmenting datasets with counter narratives (Sahoo et al., 2022) and anti-stereotypical text (Zhao, Wang, Yatskar, Ordonez, et al., 2018)), few approaches incorporate the external relations of text (van den Berg and Markert, 2020). Situating my classification work in a case study with the HC Archives (Chapter 4) enables me to examine how gender biases manifest in language, through a consideration of the external relations defined in the case study as well as a consideration of the internal relations, or sequences of words, of HC Archives documentation.

Authors	Dataset Name	English Text Data Source
Wang et al., 2018	General Language Understanding Evaluation (GLUE)	Wikipedia, news, movie reviews, fiction books, social QA questions, miscellaneous
Socher et al., 2013	Stanford Sentiment Treebank (SST)	Movie reviews
Williams et al., 2018	Multi-genre Natural Language Inference (MNLI)	Magazine articles, speeches, transcriptions of conversations, press releases, reports, letters, speeches, non-fiction and fiction books, travel guides
Maas et al., 2011	IMDb Movie Reviews	Movie reviews from Internet Movie Database (IMDb)
Bowman et al., 2015	Stanford Natural Language Inference (SNLI)	Human-generated Flickr image captions
Wang et al., 2018	Question-answering Natural Language Inference (QNLI)	Wikipedia, human annotators
Kwiatkowski et al., 2019	Natural Questions	Google queries, Wikipedia
Bajaj et al., 2018	Microsoft Machine Reading Comprehension (MS MARCO)	Bing queries, human- and machine-generated answers
Merity et al., 2016	WikiText-2	Wikipedia
X. Zhang et al., 2015	AG News	News articles

Table 6.1: **Top ten datasets of English text from the Papers with Code platform.** The table reports the top ten most cited datasets of English text on the Papers with Code platform as of July 17, 2023, with columns from left to right displaying the authors, dataset name, and dataset source. If a dataset contains text in more than one language, only sources for the English text are reported here. *This table is an updated version of Table 1 from Havens et al., 2024.*

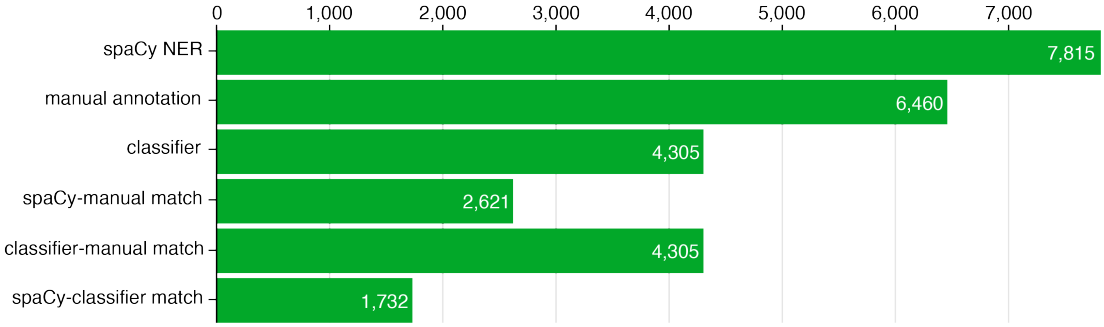


Figure 6.1: **Manual vs. Automated Annotation of Person Names.** A comparison of the person names labeled during the manual annotation process, the person names labeled automatically with an out-of-the-box Named Entity Recognition (NER) model provided with the Python programming library spaCy, and the person names labeled using the Baseline Person Name and Occupation Classifier (PNOC; see §6.6.3). From the top bar down: the count of names labeled with the spaCy NER model, the count of names labeled manually by human annotators, the count of names labeled with the Baseline PNOC, and, for the three remaining bars, the count of names labeled by both annotation methods named in the bar's label, strictly evaluated (names must exactly match).

The second gap in existing research on bias in NLP is the lack of NLP resources developed for the GLAM sector. Although NLP models are often published as widely applicable, without intended use cases specified (Raji et al., 2021), the data on which the models are trained come from limited domain areas. Table 6.1 shows the dataset names and English language data sources from the top ten most cited datasets on the Papers with Code platform:[1] the most common data sources are Wikipedia, movie reviews, and online news or magazine articles. The content and writing style of text from these domain areas differs from GLAM catalogs' documentation (Cordell, 2020), which are highly structured according to international, national, and institutional standards and practices (Dunsire and Willer, 2014; Thomassen, 2002; Welsh and Batley, 2009). The differences are evidenced in the poor performance of a spaCy Named Entity Recognition (NER) model[2] trained on online English text[3] on the corpus of archival documentation this thesis uses as data. Figure 6.1 shows that out of the 7,815 unique person names that the spaCy NER model labeled, only 2,621 (40%) were correct when strictly evaluated (meaning

[1] paperswithcode.com

[2] Named Entity Recognition (NER) models classify named entities in text, such as people, places, and organizations.

[3] The documentation of the training data, `en_core_web_sm`, is available at: github.com/explosion/spacy-models/releases/tag/en_core_web_sm-3.6.0.

names must match exactly) to the 6,460 names manual annotators' labeled. In contrast, the bespoke classifier I created (§6.6.3) labels 4,305 (60%) of the unique manually-labeled person names when strictly evaluated. Moreover, the fact that all the unique names the classifier labeled (third bar from top) equals the exact matches between the classifier and manually-annotated labelled (second bar from bottom) indicates that the classifier's false positive labels were, in fact, person names, but they were missed by the manual annotators.

Drawing on feminist theories that view knowledge and data as partial and situated (Crenshaw, 1989, 1991; D'Ignazio and Klein, 2020; Haraway, 1988; Harding, 1995; Hill Collins, 2000), I aim to detect gender biases in language so that those biases can be communicated transparently to visitors. The task defined for the models reported in this chapter is to classify types of gender biases in the dataset of archival documentation at two levels:

1. At a high level, looking at linguistically gendered terminology (i.e. the grammatical and lexical gender of words) in a text corpus as one approach to detecting bias for and against particular genders; for example, a lack of neopronouns (e.g. xe, xir, ey, eir) in a corpus indicates the exclusion of non-binary communities of people; and

2. At a low level, looking at how people are portrayed in specific spans of text can communicate stereotypes or omissions; for example, if women are frequently referred to as "his wife;' rather than their own name.

In Machine Learning (ML), classification tasks are often undertaken with a *supervised learning* approach, where a dataset is manually annotated with a predefined set of labels and a model is trained to perform that annotation automatically (Jurafsky and Martin, 2023). For text classification, this means annotating characters, words, phrases, sentences, paragraphs, or documents with those predefined labels. As described in Chapter 5, human annotators manually labeled words and phrases in HC Archives documentation using the Taxonomy of Gendered and Gender Biased Language. Here I report on training models with the aggregated dataset created from those annotators' work.

This chapter offers a starting point for classifying gender biases in text that considers the internal and external relations of the text. I do not aim to determine the best algorithm, best model parameters, or best model features. Instead, the experiments (§6.5) and results (§6.6) I report provide:

- Evidence of the potential for automatically detecting gender biases in text using NLP models, focusing on making biased language transparent rather than trying to remove biased language;
- Cascades of text classification models that identify potential gender biases in language, offering a starting point for the GLAM and NLP communities to further analyze gender biased language and optimize the models; and
- A demonstration of an alternative approach to creating ML models that prioritizes quality over quantity, representativeness over convenience, accuracy over efficiency, and situated thinking over universal thinking.

The next subsection defines key terms for this chapter. Then, I summarize related work (§6.2), describe my classification methods (§6.3), summarize experiments with and final setups for text classification models (§6.5), report the most promising model setups for detecting gendered and gender biased language (§6.6), discuss implications and limitations of this chapter's work for future research (§6.7), and then conclude the work (§6.8).

6.1.1 Definitions

To ensure clarity of communication, key terms for this chapter are defined below, building on those in §1.3.

A **token** is a term in NLP that refers to a single character or sequence of multiple characters that should be treated as a group (S. Bird et al., 2019). Tokens can contain letters, numbers, punctuation, or other symbols. Examples of tokens are "was," "n't," "2023," "!," and "#metoo."

A **model** in this thesis refers to an ML model, which consists of an algorithm, parameters, and data. The Oxford English Dictionary defines a model as, "A simplified or idealized description or conception of a particular system, situation, or process, often in mathematical terms, that is put forward as a basis for theoretical or empirical understanding, or for calculations, predictions, etc.; a conceptual or mental representation of something" (OED, 2023). When training an ML model, an algorithm applies mathematical (largely statistical)

techniques to data with the aim of finding meaningful patterns in the data. The term **Learning** in Machine Learning refers to this process of finding meaningful patterns and being able to do so repeatedly (Deisenroth et al., 2020). Once the model is trained to find patterns in one dataset, it can be run on other datasets.

Deep learning models rely on a neural network, a many-layered network of computing units, where each unit outputs one value based on a vector of values input to the unit. A deep learning model trained on one dataset is often referred to as a **pre-trained model** or *foundation model* (Bommasani et al., 2021). The process of customizing a pre-trained model to other data is called fine-tuning, resulting in a **fine-tuned model** (Jurafsky and Martin, 2023).

If data are not numeric, such as in NLP when data are text, they must be converted to a numeric format in order for the algorithm to run mathematical techniques on them (one such technique is word embeddings, defined below). In NLP, a **language model**, such as the GPT-4 model behind ChatGPT (OpenAI, 2023), estimates the probability of the occurrence of a particular sequence of tokens (Goldwater, 2021). Language models such as this are **generative**, because they produce sequences of tokens likely to occur following a given sequence of tokens. This chapter reports on experiments with **discriminative** models, specifically **discriminative classification models**. Discriminative models aim to learn how to differentiate between data in meaningful ways. For classification, this means differentiating between the different labels applied to the dataset that is input to a model (Jurafsky and Martin, 2023).

An **algorithm** refers to a mathematical equation designed to provide a generalizable, computational method for extracting meaningful patterns, or information, from data (Deisenroth et al., 2020; Eisenstein, 2018). Algorithms aim to map outputs to inputs, or predictions to observations (Jurafsky and Martin, 2023), such as mapping my Taxonomy's labels to an archival metadata description.

Parameters are values input into algorithms along with training data (i.e. the data from which the model is meant to learn). Essentially, they are predefined variables in mathematical equations. Model parameters' values are often determined based on the model's training data. Unless otherwise stated, models in this chapter use default parameters supplied by the Python programming

libraries scikit-learn (Pedregosa et al., 2011), scikit-multilearn (Szymański and Kajdanowicz, 2018), and sklearn-crfsuite (Korobov, 2015).

Features refer to the representation of data input into an ML model, so they must be provided to a model in a numeric representation (Jurafsky and Martin, 2023). They provide a way of encoding prior knowledge for a model, so the model can combine that knowledge with the training data to more effectively learn what patterns in the data are meaningful. Features can be manually created (e.g. manually defining labels and annotating data with those labels) or obtained automatically from another model (e.g. using a part-of-speech tagger to assign parts of speech to tokens in a text corpus) (Eisenstein, 2018).

Word embeddings are numeric vectors that represent the meaning of words based on their distribution in a text corpus (Eisenstein, 2018; Jurafsky and Martin, 2023). The idea behind word embeddings is rooted in the distributional hypothesis from Linguistics: words' meanings come from their surrounding words, so synonymous words are likely to appear in similar locations in relation to other words (Firth, 1957; Z. S. Harris, 1954; Joos, 1950).

6.2 Related Work

While access to ML models, as well as the data and computing power they require, has improved, the risks such models pose remain. Platforms such as Amazon SageMaker[4] and Google Colab[5] have increased access to computing power for training and using deep learning models; Hugging Face[6] has increased access to pre-trained models; and Zooniverse[7] and Amazon Mechanical Turk[8] have increased the scale at which data can be annotated. Thanks to platforms such as these, programmers can more easily generate large datasets for models, fine-tune models, and run models. That being said, concerns of ethics and quality regarding the use of such platforms

[4] aws.amazon.com/sagemaker
[5] colab.google
[6] huggingface.co
[7] zooniverse.org
[8] mturk.com

remain. Training ML models is expensive not only financially, but also environmentally (Strubell et al., 2019; Bender and Gebru et al., 2021); pre-trained models have social biases engineered into them (Jentzsch and Turan, 2022; Jiang and Fellbaum, 2020; Lu et al., 2020; Martinková et al., 2023; Tan and Celis, 2019), which fine-tuning approaches have not removed (de Vassimon Manela et al., 2021; Jin et al., 2021); ethical compensation for crowdworkers is not always enforced (Crawford, 2021; Gray and Suri, 2019; Irani, 2015), and datasets created with crowdworkers' labor have been found to contain inaccuracies large enough to jeopardize those datasets' validity (Kreutzer et al., 2022). While certain tasks may be minimally affected by these risks, my task of classifying gender biased language is not.

To classify types of gender biases across a large-scale text corpus, one must be able to distinguish between biases that originate in the text corpus and biases that originate in the model performing the classification. No approach to making such a distinction has been established in NLP or, for non-textual data, in ML more broadly. Investigations of bias injection from a pre-trained model to a fine-tuned model have yielded mixed results (Jin et al., 2021; Steed et al., 2022); similar to the inadequacy of debiasing word embeddings (Goldfarb-Tarrant et al., 2021; Gonen and Goldberg, 2019), minimizing models' biases through fine-tuning has proved insufficient for mitigating ML systems' biases. Steed et al. (2022) encourage greater emphasis be placed on data quality and the context of the research, sentiments that echo the recommendations of Crawford (2017), D'Ignazio and Klein (2020), Ciora, Iren, and Alikhani (2021), and Aragon et al. (2022). Looking to the Digital Humanities, collaborations across ML and GLAM demonstrate how ML research could more effectively incorporate considerations of data quality and context.

Within the GLAM sector, discussions of the capabilities of data differ from those in ML. In GLAM, data, in the form of cultural heritage collections and documentation of those collections (Padilla, 2017, 2019), are viewed as serving a particular narrative in the interest of certain communities and at the expense of other communities (Adler, 2017; Hauswedell et al., 2020; Ortolja-Baird and Nyhan, 2022; Yale, 2015). In a collaboration between the British Library and Alan Turing Institute, Beelen et al. (2023, 2022) demonstrate one approach to measuring data biases: they investigated the representativeness of a dataset in relation to the historical context of their research project. The authors introduce

the "environmental scan" as an approach to estimating the representativeness of a dataset, estimating how well the political views in a collection of digitized, 19th century, British newspapers reflect the variety of political views that existed in all newspapers throughout Great Britain in the 19th century. Historians could then adjust their analysis based on an understanding of the political skew in the the available newspapers relative to the unavailable newspapers, meaning newspapers omitted from the British Library's digitized collection.

Baker and Salway (2020) and Salway and Baker (2020) provide, at the time of writing, the only known research that applies computational methods to study bias in GLAM documentation. Applications of ML to GLAM have focused on creating or adding to catalog metadata (Berardi et al., 2012; Cordell, 2020; Padilla, 2019), digitizing cultural heritage items (Hosseini et al., 2022; Nockels et al., 2022; Pal et al., 2016), and analyzing the content of cultural heritage collections (Ames and Havens, 2022; Beelen et al., 2023; Beelen et al., 2022). However, Baker and Salway (2020) and Salway and Baker (2020) applied computational linguistic and qualitative analysis techniques to study biases in the descriptive language of a historical cataloger: that of Mary Dorothy George in a British Museum catalog. Salway and Baker's (2020) application of computational linguistic techniques demonstrates that, "aspects of curatorial voice do manifest in linguistic features" (p. 165), and those linguistic features (e.g. adverbs that communicate an interpretation of an action) can be identified at a large scale computationally.

To fully understand the implications of these computationally-detected linguistic features, though, the authors note the need for qualitative analysis through close reading. Baker and Salway (2020) report on this close reading, analyzing the linguistic features of George's documentation in relation to "the institutional culture of the British Museum Department of Prints and Drawings, the expectations of cataloguing and of writing for an academic press, and the labour conditions produced both by global conflict and by class and gender dynamics deeply rooted in British society" (p. 778). Through analysis of the documentation relative to its context and alongside cultural heritage material the documentation describes, the authors identify social biases in the association of certain terminology to particular communities of people, and in the details omitted from the documentation (Baker and Salway, 2020). Similar to Baker and Salway (2020), I combine computational and

qualitative approaches in my research of bias in GLAM documentation, with this chapter reporting on the computational approach and Chapter 7 reporting on the qualitative approach.

Unlike Baker and Salway, however, the GLAM documentation I study in this thesis has been written over a longer time period (circa 18th century to 21st century) and by numerous (an unknown number of) catalogers. Rather than trying to understand all social biases in HC Archives documentation, I focus on gender biases, an active area of research in many computational disciplines due to the ways in which sexism is built into datasets (Perez, 2019) and technology systems (Hicks, 2021; Noble, 2018). In NLP gender bias research, gender is often reduced to a binary and gender bias is often vaguely defined (Blodgett et al., 2020; Devinney et al., 2022; Stańczak and Augenstein, 2021). Too often researchers make assumptions about gender that encode their own biases into NLP systems. For example, Dinan et al. (2020) reduce gender to a binary in their design decisions for a gender biased text classification model, even though their data collection contained more gender categories; and Garimella et al. (2019) found performance differences between part-of-speech tagging models' application to news articles written by men and women, though even the authors' approach to encoded gender biases into their output dataset by assigning authors' genders based on their names. In reality, text datasets often do not provide enough information to determine a person's gender with certainty (Shopland, 2020; Spiel et al., 2019).

This chapter reports on my execution of an alternative approach to gender bias research in NLP that allows for uncertainty and subjectivity. Utilizing the types of gendered and gender biased language from the Taxonomy introduced in Chapter 5, I create gender biased text classification models that allow for uncertainty in gender categorizations. By training models on the aggregated dataset from Chapter 5, which contains a combination of annotators' interpretation of gender biases in text (namely *Omission* and *Stereotype*), I also allow for subjectivity in the models. Prioritizing representativeness over convenience, I use traditional ML approaches to create the models reported in this chapter (§6.5, §6.6). This approach enables me to better anticipate the potential harms from my data and models, and avoid the risk of pre-trained models injecting or amplifying biases from their training datasets into my case study. Using traditional ML methods also increases the accessibility of the

models as tools for the GLAM sector, as training deep learning models from scratch requires expensive computing power (Bender et al., 2021; Bommasani et al., 2021), an already-identified barrier to large-scale computational analysis of cultural heritage in GLAM (Terras et al., 2018).

6.3 Methods

The models in this chapter use Python version 3.19.13[9] and several programming libraries built on this language. Specifically, I use scikit-learn (Pedregosa et al., 2011) version 1.2.1, a well-documented library for traditional ML models, scikit-multilearn (Szymański and Kajdanowicz, 2018) version 0.2.0, a library built on top of scikit-learn specifically for multinomial classification tasks,[10] and sklearn-crfsuite (Korobov, 2015) version 0.3.6, a library built on top of scikit-learn specifically for sequence classification tasks. Using this selection of libraries facilitated my creation of model cascades, or sequential combinations of models, because the same data structures, functions, and methods can be used across the three libraries. Unless otherwise stated, the models use the default parameters provided by these libraries. Except for the word representation algorithm, fastText, the algorithms I employ are applied in a supervised ML setting, with the aggregated data from Chapter 5 providing the labeled dataset on which to train, validate, and test classification models. The remainder of this section details my classification models' algorithms (§6.3.1), quantitative evaluation metrics (§6.3.2), and word representations (§6.3.3).

6.3.1 Algorithms

Experiments were conducted with several algorithms for each type of model (multilabel token, multiclass sequence, and multilabel document classifiers) to determine which would provide the strongest foundation for gender biased text classification models. All algorithms are multinomial, either multilabel, meaning they aim to classify text with zero or more labels from the Taxonomy,

[9] www.python.org

[10] Multinomial classification tasks, as opposed to binary classification tasks, have more than two classes, or labels, with which a model can classify data.

or multiclass, meaning they aim to classify text with up to one label from the Taxonomy. These algorithms include:

- **Logistic Regression**, which is the baseline supervised ML algorithm for text classification. In other fields, Logistic Regression may be referred to as Logit Analysis (Cramer, 2010a). Logistic Regression is a linear classifier, meaning it predicts labels based on a linear function of the features input into the model in log space. As a discriminative model, Logistic Regression aims to learn what distinguishes one label from another, directly computing the probability that a text has a particular label. The history of Logistic Regression as a statistical method begins in the 19th century and has independent origins in several disciplines (Cramer, 2010b). I use Logistic Regression[11] with liblinear regularization[12] (Fan et al., 2008), as scikit-learn recommends this regularization approach for smaller datasets (scikit-learn developers, 2023c), and with a one-vs.-rest setup, also called the binary relevance method, where one classifier is fit per label,[13] for multilabel document classification.

- **Conditional Random Field** (CRF), an algorithm designed for sequence classification (e.g. NER) that is built on Logistic Regression (Eisenstein, 2018; Lafferty et al., 2001). CRFs deal with unknown words more effectively than other algorithms for sequence classification (Jurafsky and Martin, 2023), so I use CRFs[14] for multiclass sequence classification.

- **Support Vector Machines** (SVM), a linear, discriminative model which has been shown to perform well on document classification tasks for multiple domain areas compared to deep learning models (Adhikari et al., 2019) and is suggested in the *scikit-learn algorithm cheat sheet* for supervised text classification tasks (scikit-learn developers, 2023b). Originally for binary settings, the SVM algorithm we use makes use of Zadrozny and Charles' (2002) extension of SVM for multinomial settings. I use SVM with Stochastic Gradient Descent,[15] a simple yet

[11] scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[12] Regularization mitigates overfitting a model to its training data (Jurafsky and Martin, 2023).
[13] scikit-learn.org/stable/modules/multiclass.html
[14] sklearn-crfsuite.readthedocs.io/en/latest/api.html
[15] scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html

efficient optimization technique (scikit-learn developers, 2023d), in a one-vs.-rest setup for multilabel document classification.

- **Random Forest**, a discriminative model which has been shown to perform well in experiments with multilabel classification tasks (Madjarov et al., 2012) and is suggested in the *scikit-learn algorithm cheat sheet* for supervised text classification tasks (scikit-learn developers, 2023b). Random Forests implement an *ensemble* method (scikit-learn developers, 2023a; Szymański and Kajdanowicz, 2018), combining multiple algorithms, in this case, Decision Trees (Breiman et al., 1984), for improved generalizability (Breiman, 2001). I use Random Forest[16] for multilabel token and document classification experiments.

- **Passive Aggressive** is a linear, online learning algorithm, meaning an algorithm that processes data in sequential order, that can be applied in binary and multinomial classification tasks on non-separable data (Crammer et al., 2006). In this chapter I report on experiments using Passive Aggressive[17] with a Classifier Chain model for multilabel token classification, and with a CRF model for multiclass sequence classification.

- **Adaptive Regularization of Weight Vectors** (AROW), an online learning algorithm that assumes data are non-separable (Crammer et al., 2013). As such, AROW is well-suited to sequence classification. This algorithm is also better suited for datasets with noisy labels compared to the Passive Aggressive algorithm (ibid.). I use AROW with a CRF model[18] for multiclass sequence classification.

- **Classifier Chain**, which takes a one-vs.-rest approach, treating every label in a multilabel classification setup as a binary classification task (either text has or does not have the label) (Read et al., 2009). Classifier Chains provide a scalable and simple-to-implement (no parameter

[16] scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[17] scikit-learn.org/stable/modules/generated/sklearn.linear_model.PassiveAggressiveClassifier.html
[18] sklearn-crfsuite.readthedocs.io/en/latest/api.html

configuration is required) algorithm(ibid.). This chapter uses a Classifier Chain[19] of Random Forests for multiclass sequence classification.

- **fastText**, which provides an unsupervised ML algorithm for creating representations of word meanings, called word embeddings (Bojanowski et al., 2017). Instead of representing entire words as vectors, typical of word representation algorithms, fastText creates vector representations of character n-grams, enabling unseen words to be represented as vectors. I used fastText to represent words of HC Archives documentation as features for multilabel token and multiclass sequence classification.

- Lastly, **Term Frequency-Inverse Document Frequency** (TF-IDF), an approach to encode text documents numerically as matrices, so the documents can be input into ML models. *Term frequency* refers to the total occurrences of a term in a document (Luhn, 1957) and *document frequency* refers to the total number of documents in which that term occurs (Sparck Jones, 1972). Each token in a document's TF-IDF matrix comes from multiplying the token's term frequency by its inverse document frequency. As a result, TF-IDF assigns rare words that tend to carry more meaning than common words (e.g. "and," "the") that tend to carry less meaning. TF-IDF is a simple algorithm and thus a suitable baseline for representing documents (Jurafsky and Martin, 2023). I used TF-IDF[20] to represent descriptions from HC Archives documentation in multilabel document classification.

6.3.2 Evaluation

I evaluated all the models with the same metrics as the manual annotation process (§5.1.6.3), calculated for a held-out test subset of the data. A "correct" label refers to a label the model made on a text span that human annotators also made on that same text span. The evaluation metrics are:

- **True Positive count**, or True Positives (TP), which records the total number of times a model correctly predicts a label;

[19] scikit.ml/api/skmultilearn.problem_transform.cc.html

[20] scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

- **False Positive count**, or False Positives (FP), which records the total number of times a model predicts a label where there should not be one;

- **False Negative count**, or False Negatives (FN), which records the number of times a model did not predict a label when it should have;

- **Precision**, which is calculated as the ratio of TP to the sum of TP and FP, measuring how many of the predicted labels (meaning the labels the model made) are correct (see equation 5.1 in 5.1.6.3);

- **Recall**, which is calculated as the ratio of TP to the sum of TP and FN, measuring how many of the expected, or correct, labels were predicted by the model (see equation 5.2 in 5.1.6.3); and

- **F_1 score**, which combines precision and recall into one metric, calculated as the harmonic mean of precision and recall (van Rijsbergen, 1979; see equation 5.3 in 5.1.6.3).

In addition to reporting the above metrics per label, I report **macro** and **micro** precision, recall, and F_1 scores when evaluating this chapter's models. Macro scores are the average of each of those scores per label, providing a balanced measure of a model's ability to recognize the labels (Eisenstein, 2018). Micro scores are computed from the sums of TP, FP, and FN across all labels, providing a measure of a model's ability to recognize the labels weighted by each labels' frequency (Eisenstein, 2018). I selected these metrics because they are standard for text classification evaluation, as they are suitable for imbalanced samples per label in a corpus (Eisenstein, 2018; Jurafsky and Martin, 2023).

When comparing the performance of models in experiments (§6.5) to choose the best model setup for cascades (§6.6), I focus on models' F_1 scores. Optimizing for F_1 score in this way balances considerations of both precision and recall, and is standard practice in NLP. This means that the aim is to create models that are both (a) highly precise, so when a model makes an annotation it is likely to be correct, and (b) highly robust, meaning a model is unlikely to miss making annotations that should have been made.

6.3.3 Word Representations

To numerically represent the token data, I created custom word embeddings with fastText (Bojanowski et al., 2017), which is available through the Python programming library Gensim,[21] to train custom word embeddings on HC Archives documentation. I used embeddings because this vector representation is the standard approach to modeling word meanings in NLP (Eisenstein, 2018; Firth, 1957; Z. S. Harris, 1954; Joos, 1950; Jurafsky and Martin, 2023). I used fastText to create embeddings because this model can represent unseen words using a combination of embeddings for character n-grams (parts of words), making it more generalizable than embedding models that represent entire words (Bojanowski et al., 2017). As with all models in this chapter, I used the default parameters and training architecture (Continuous Bag of Words) for the fastText model, creating 50- and 100-dimension embeddings.[22] Though pre-trained word embeddings (e.g. spaCy's pre-trained sense2vec embeddings[23] (Trask et al., 2015) or pre-trained GloVe embeddings[24] (Pennington et al., 2014)) rely on datasets much larger than HC Archives documentation to represent words' meanings, thus they may better represent the complexity of word meanings, using pre-trained embeddings risks bias injection from those larger datasets (Jin et al., 2021; Steed et al., 2022). As a result, this chapter's models use the custom fastText embeddings to ensure any identified gender biases originate in my dataset of HC Archives documentation.

6.4 Data Preparation

Using the aggregated dataset and description dataset from Chapter 5,[25] I created three versions of the data to input into classification models. My code uses Python and the Python libraries Natural Language Toolkit (NLTK) (Loper and Bird, 2002) and pandas (The pandas development team, 2023) to

[21] radimrehurek.com/gensim/models/fasttext.html
[22] github.com/thegoose20/gender-bias-models/blob/main/word_embeddings/WordEmbeddings.ipynb.
[23] spacy.io/universe/project/sense2vec
[24] github.com/stanfordnlp/GloVe
[25] Available at: https://datashare.ed.ac.uk/handle/10283/8563.

perform these data transformations, and is available on GitHub.[26,27,28] NLTK is a widely-used and well-documented library for NLP programming in Python, and pandas is a widely-used and well-documented library for data science programming in Python; both are built to work with scikit-learn (Pedregosa et al., 2011), the ML programming library also used in this chapter (e.g. input data for scikit-learn models can be represented as a pandas DataFrame, a tabular representation of data). Specifically, my data transformations involved:

1. *Tokenization:* Using NLTK's `word_tokenize`[29] and `sent_tokenize`[30] methods, I separated the text in the descriptions dataset into word and punctuation tokens, and into sentences.

2. *Description to Annotation Linking:* I associated every annotation from the aggregated dataset to a description in the description dataset, adding annotation identifier and annotation label columns to the description dataset, resulting in a new CSV dataset of annotated descriptions.

3. *Sentence Offsets:* using description offsets (according to the brat standoff format, where the start offset is the index position of the first character and the end offset is one index position after the last character for a file of descriptions[31]) in the description dataset, I calculated each sentence's offsets, creating a new CSV dataset of sentences with columns for sentence and description identifiers, sentences, and sentence offsets.

4. *Token Offsets:* Using the sentence offsets, I calculated each token's offsets, creating a new CSV dataset of tokens with columns for token, sentence, and description identifiers; and tokens and token offsets.

5. *Token to Annotation Linking:* I associated each token to an annotation identifier and annotation label from the aggregated dataset, or to the integer value `99999` if the token was not in an annotated text span.

[26] github.com/thegoose20/gender-bias-models/blob/main/analysis/Analysis_LengthsAndOffsets.ipynb
[27] github.com/thegoose20/gender-bias-models/blob/main/analysis/Analysis_TokenBIOTags.ipynb
[28] github.com/thegoose20/gender-bias-models/blob/main/document_classification/SplitData_DocumentClassification.ipynb
[29] www.nltk.org/api/nltk.tokenize.word_tokenize.html
[30] www.nltk.org/api/nltk.tokenize.sent_tokenize.html
[31] brat.nlplab.org/standoff.html

6. *BIO Tagging:* Following the standard Beginning, Inside, Outside (BIO) tagging scheme for NER (Jurafsky and Martin, 2023), I associated every annotated token with a `B-[LABELNAME]` tag or `I-[LABELNAME]` tag, where `[LABELNAME]` was the name of the label from the annotation Taxonomy with which the token was annotated (e.g. `B-Gendered-Pronoun`, `I-Gendered-Pronoun`). I assigned `B-[LABELNAME]` tags to the first token in an annotation and `I-[LABELNAME]` to the remaining tokens of an annotation. I gave every remaining, unannotated token an `O` tag. This resulted in a new CSV dataset of tagged tokens.

7. *Token Labeling*: I generalized tokens' `BI` tags to their corresponding label names (e.g. `B-Gendered-Pronoun` became `Gendered-Pronoun`), keeping the `O` tags to indicate unlabeled tokens. This resulted in a new CSV dataset of labeled tokens.

The labeled token dataset (Table 6.2) served as input data for multilabel token classification (§6.5.1), the tagged token dataset (Table 6.3) served as input data for multiclass sequence classification (§6.5.2), and the annotated description dataset (Table 6.4) served as input data for multilabel document classification. For my initial Experiments (§6.5), I split these datasets into training, validation, and test subsets in proportions of 60%, 20%, and 20%, respectively, balancing the type of metadata field ("Title," "Scope and Contents," "Biographical / Historical," "Processing Information") across the subsets. I ensured that no sentence was divided across multiple splits in the token and sequence classification input, and I ensured that no description was divided across multiple splits in the document classification input. Tables 6.5, 6.6, and 6.7 display the labels' occurrences across the subsets.

For the resulting model cascades (§6.6), which consist of two to three models, I performed five-fold cross-validation, creating five folds, or subsets, of 20%, again balancing the type of metadata field across each fold. Then, I iteratively selected four folds (80% of the input dataset) for the training data and used the remaining fold (20% of the input dataset) for the test data, training and testing five instances of each model in the cascades. Cross-validation is a standard approach in ML for training and evaluating models. This approach enabled me to generate predictions for the entire

dataset, one 20% test subset at a time, which I then input into a subsequent model of a cascade as features. Tables 6.8 through 6.10 display the labels' occurrences across all five folds.

Tables 6.11 and 6.12 display the occurrence of different combinations of labels in the labeled token and description datasets, respectively. The datasets for classification experiments and cascades can be downloaded from the University of Edinburgh's DataShare platform at: https://doi.org/10.7488/ds/7539.

description_id	sentence_id	ann_id	token_id	token	token_offsets	label
1	1	99999	5	Papers	(24, 30)	O
1	1	99999	6	of	(31, 33)	O
1	1	14384	7	The	(34, 37)	Unknown
1	1	24275	7	The	(34, 37)	Masculine
1	1	52952	7	The	(34, 37)	Stereotype

Table 6.2: **Labeled Token Dataset.** A sample of the data input to multilabel token classification models. The token with identifier 7, "The," received more than one label, so it appears in three rows, each with a unique annotation identifier. Displayed tokens are from the sentence, "Papers of The Very Rev Prof James Whyte (1920-2005)."

description_id	sentence_id	ann_id	token_id	token	token_offsets	tag
1	1	99999	5	Papers	(24, 30)	O
1	1	99999	6	of	(31, 33)	O
1	1	14384	7	The	(34, 37)	B-Unknown
1	1	24275	7	The	(34, 37)	B-Masculine
1	1	52952	7	The	(34, 37)	B-Stereotype

Table 6.3: **Tagged Token Dataset.** A sample of the data that served as input for multiclass sequence classification models. Displayed tokens are from the sentence, "Papers of The Very Rev Prof James Whyte (1920-2005)."

desc_id	start	end	field	description	label
4699	1853	2066	Biographical / Historical	Labelled Apparently some chapters, amounting t...	[Omission]
8942	384	540	Biographical / Historical	James Aikman of Perth signed his name to a vol...	[]
5440	5692	5850	Biographical / Historical	This piece was published in 'Milk Production i...	[]
3474	3608	8549	Biographical / Historical	Margaret Winifred Bartholomew was born...	[Omission, Stereotype]
4769	2378	2576	Biographical / Historical	Blacker and Thomson became close friends...	[Omission]

Table 6.4: **Description Dataset.** A sample of the data that served as input for multilabel document classification models. The columns from left to right are the description's identifier, the brat start and end offsets, metadata field name, sample text, and label.

	training	validation	test	all
Gendered Pronoun	2210	759	762	3731
Gendered Role	1954	693	607	3254
Generalization	1203	385	430	2018

Table 6.5: ***Linguistic* Labels per Data Subset in Experiments.** Count of descriptions with a ***Linguistic*** label across the training, validation, and test subsets of the labeled token dataset used in classifier experiments (§6.5).

	training	validation	test	all
B-Feminine	840	323	298	1461
I-Feminine	1827	846	696	3369
B-Masculine	3390	1024	1096	5510
I-Masculine	4693	1378	1366	7437
B-Unknown	6233	2060	2024	10317
I-Unknown	10210	3506	3235	16951
B-Occupation	1827	655	474	2956
I-Occupation	2156	781	565	3502

Table 6.6: ***Person Name* and *Occupation* Labels per Data Subset in Experiments.** Count of descriptions labeled with a ***Person Name*** or an *Occupation* label across the training, validation, and test subsets of the tagged token dataset used in classifier experiments (§6.5).

	training	validation	test	all
Omission	2400	804	828	4032
Stereotype	957	315	329	1601

Table 6.7: ***Omission* and *Stereotype* Labels per Data Subset in Experiments.** Count of descriptions labeled with *Omission* and *Stereotype* across the training, validation, and test subsets of the description dataset used in classifier experiments (§6.5).

	1	2	3	4	5	all
Gendered Pronoun	728	689	793	759	762	3731
Gendered Role	588	638	728	693	607	3254
Generalization	373	406	424	385	430	2018

Table 6.8: ***Linguistic* Labels per Data Subset in Cascades.** Count of descriptions labeled with a ***Linguistic*** label across the five folds of the labeled token dataset used in classifier cascades (§6.6). Columns "1" through "5" indicate the fold number; "all" indicates the total across all folds.

	1	2	3	4	5	all
B-Feminine	264	302	274	323	298	1461
I-Feminine	582	661	584	846	696	3369
B-Masculine	1037	1098	1255	1024	1096	5510
I-Masculine	1371	1512	1810	1378	1366	7437
B-Unknown	1996	2143	2094	2060	2024	10317
I-Unknown	3179	3510	3521	3506	3235	16951
B-Occupation	587	590	650	655	474	2956
I-Occupation	683	688	785	781	565	3502

Table 6.9: ***Person Name* and *Occupation* Labels per Data Subset in Cascades.** Count of tokens with a ***Person Name*** or an *Occupation* label across the five folds of the tagged token dataset used in classifier cascades (§6.6). Columns "1" through "5" indicate the fold number; "all" indicates the total across all folds.

	1	2	3	4	5	all
Omission	798	749	834	813	838	4032
Stereotype	341	290	325	302	343	1602

Table 6.10: ***Omission* and *Stereotype* Labels per Data Subset in Model Cascades.** Count of descriptions labeled with *Omission* and *Stereotype* across the five folds of the description dataset used in classifier cascades (§6.6). Columns "1" through "5" indicate the fold number; "all" indicates the total across all folds.

labels	tokens
none	744728
Gendered Pronoun	3624
Gendered Role	3151
Generalization	1808
Gendered Pronoun, Generalization	107
Generalization, Gendered Role	103

Table 6.11: ***Linguistic* Label Combinations.** Total tokens with no label ("none"), one of the ***Linguistic*** labels, or multiple ***Linguistic*** labels in the labeled token dataset used in classification experiments (§6.5) and cascades (§6.6).

labels	descriptions
none	22747
Omission	2964
Omission, Stereotype	1068
Stereotype	533

Table 6.12: ***Omission* and *Stereotype* Combinations.** Total descriptions with no label ("none"), an *Omission* label, a *Stereotype* label, or both labels in the description dataset used in classification experiments (§6.5) and cascades (§6.6).

6.4.1 Preprocessing for Linguistic Classifiers

For the classification of ***Linguistic*** labels (*Gendered Pronoun*, *Gendered Role*, *Generalization*), I used the Python programming libraries pandas (McKinney, 2010; The pandas development team, 2023), scikit-multilearn (Szymański and Kajdanowicz, 2018), and scikit-learn (Pedregosa et al., 2011) to preprocess the token data (Table 6.2), train Linguistic Classifiers, and evaluate the Classifiers. Scikit-multilearn provides functions and methods built on top of the scikit-learn library to create models for multilabel classification tasks, where a model can classify a single word with more than one label. Thus, scikit-multilearn was suitable for my task of classifying tokens with ***Linguistic*** labels: a single token could have any combination of this category's labels.

Data preprocessing included the standard practices of lowercasing tokens and representing tokens numerically with word embeddings; I did not remove punctuation, numbers, or stop words. I converted the tokens' labels (i.e. either

one or more ***Linguistic*** labels, or an `O` to indicate no label) to a binary representation, with one number per label, where `1` indicated a token had a label and `0` indicated a token did not have that label. For example, the annotation of a token with a *Gendered Pronoun* label would be represented with the sequence `1 O O O`, and a token with no labels, `O O O 1`. These binarized labels were passed to the Classifiers as targets, meaning the classes with which the model should annotate text.

6.4.2 Preprocessing for Person Name and Occupation Classifiers

For the classification of ***Person Name*** (*Feminine*, *Masculine*, *Unknown*) and *Occupation* labels, I used the Python programming libraries pandas (McKinney, 2010; The pandas development team, 2023), sklearn-crfsuite (Korobov, 2015), and scikit-learn (Pedregosa et al., 2011) to preprocess the tagged token data (Table 6.3), train Person Name and Occupation Classifiers, and evaluate the Classifiers. Sklearn-crfsuite provides functions and methods built on top of scikit-learn to create models for sequence classification, where a model considers a sequence of text (in this chapter, a sentence) when classifying tokens. Sequence classification is a common task for NER, which includes classifying people's names, as I aimed to do. In sequence classification, a token can have either zero or no classes. Thus, sklearn-crfsuite was suitable for my task of classifying tokens with ***Person Name*** and *Occupation* labels: a single token should have at most one tag (either `B-[LABELNAME]`, `I-[LABELNAME]`, or `O`).

Data preprocessing included the standard practices of lowercasing tokens and representing the tokens numerically with word embeddings; I did not remove punctuation, numbers, or stop words. Then, I converted the tokens' labels (i.e. either one or more of the ***Person Name*** or *Occupation* labels, or an `O` to indicate no label) to a binary representation, with one number per label, where `1` indicated a token had a label and `0` indicated that a token did not have that label. Next, I grouped the tagged token data by token, ensuring each token had only one label (e.g. a token should not have one of the Taxonomy's labels and an `O`). I then further grouped the data by sentence and created `START` and `END` boolean values for each token, with `START=True` indicating

that a token was the first token of a sentence and `END=True` indicating that a token was the last token of a sentence; all other tokens had `START=False` and `END=False`. Lastly, I passed the tokens' embeddings, labels, `START` and `END` booleans, and bias values of 1.0 (as recommended in the sklearn-crfsuite documentation (Korobov, 2015) to the Classifiers as features.

6.4.3 Preprocessing for Omission and Stereotype Classifiers

I used the Python programming libraries pandas (McKinney, 2010; The pandas development team, 2023), scikit-learn (Pedregosa et al., 2011), and SciPy (Virtanen et al., 2020) to preprocess the description data (Table 6.4), train Omission and Stereotype Classifiers, and evaluate the Classifiers. In a multilabel document classification task, a model can classify a document with any combination of labels, considering the text of the entire document when determining how to classify it. In this chapter, one description from the HC Archives' catalog serves as one document (from either the "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information" metadata field), which could be classified with no label, either an *Omission* or *Stereotype* label, or both labels. Thus multilabel document classification suited my task of annotating descriptions with *Omission* and *Stereotype* labels.

Scikit-learn provides a straightforward approach to multilabel document classification with a small number of labels, so I used this library rather than scikit-multilearn, which better facilitates multilabel classification of tokens with a larger number of labels. Moreover, as sckikit-learn is the foundational library of scikit-multilearn and sklearn-crfsuite, the data input and output formats are similar, so I could easily combine the previous classifiers with an Omission and Stereotype Classifier (§6.6). SciPy provides methods for representing sparse matrices, providing a memory-efficient approach to representing documents that scikit-learn models use.

Data preprocessing included the standard practice of representing HC Archives' descriptions numerically with matrices. First, I transformed the descriptions into TF-IDF matrices and binarized the descriptions' *Omission* and *Stereotype* labels. Then, when including previous classifiers' labels as features (§6.6), I binarized those labels and concatenated them to the document matrices. The matrices served as features to input into Omission and

Stereotype Classifiers; the binarized labels served as targets for the Classifiers.

The next section (§6.5) further explains the rationale for the token, sequence, and document classification setups for the models reported in this chapter: Linguistic Classification (§6.5.1), Person Name and Occupation Classification (§6.5.2), and Omission and Stereotype Classification (§6.5.3).

6.5 Experiments

Initially, I created a multilabel document classifier to annotate descriptions with all the Taxonomy's labels (tables G.7, G.8, and G.9). Upon conducting error analysis on the resulting predictions, however, I realized that while annotated text spans labeled as *Omission* or *Stereotype* need the context of the description surrounding them to understand what has been omitted or represented stereotypically, annotated text spans with the other labels were actually harder to interpret when applied to an entire description. Consequently, I experimented with token classification models for ***Linguistic***, ***Person Name***, and *Occupation* labels; and with document classification models for *Omission* and *Stereotype* labels only. Those experiments, summarized in this section, inform the construction of the cascades (§6.6): I used the model setups that yielded the best performance, primarily based on F_1 scores, for the cascades' models. For each experiment, performance scores were based on strict agreement measures on the validation subsets of data, as this is more computationally efficient than a loose evaluation. Strict agreement means that to be a TP (correct), a model's annotation must exactly match the manual annotation of a token with *labels* for multilabel token classification, the manual annotation of a token with *tags* (`B-[LABELNAME]`, `I-[LABELNAME]`, `O`) for multiclass sequence classification, and the manual annotation of a description with *labels* for multilabel document classification.

6.5.1 Linguistic Classification

I experimented with multilabel token classification model setups to automatically annotate tokens with the Taxonomy's ***Linguistic*** category of labels: *Gendered Pronoun*, *Gendered Role*, and *Generalization*. According to

the annotation instructions (Appendix D), a token could be annotated with multiple ***Linguistic*** labels, making the classification of these labels intuitively appropriate for a multilabel task, where each token can be given more than one label. Thus my model setup experiments included word representation experiments (§6.5.1.1) and algorithm experiments (§6.5.1.2) with multilabel classifiers. Due to inconsistencies in manual annotation with the *Generalization* label, models' predictions with this label require further error analysis to determine how well the model annotates with the *Generalization* label relative to the annotation instructions. As a result, this chapter frames the multilabel token classification task as detecting gendered language, because while *Generalization* annotates gender biased language, only the *Gendered Pronoun* and *Gendered Role* labels were known to be reliably annotated. I created models to classify text with ***Linguistic*** labels to provide an approach to measure gender biases at a high level, because comparing the quantities of grammatically feminine and masculine terminology[32] provides an indication of the prevalence of certain genders relative to others across a text corpus. For example, a larger number of masculine pronouns compared to feminine pronouns in a corpus indicates the privileging of masculine perspectives.

6.5.1.1 Word Representation Experiment

As mentioned previously, word embeddings are a standard approach in NLP to representing word meanings in models. In order to confirm the value of using word embeddings for this chapter's models, I conducted an experiment to compare the performance of multilabel token classification models with and without the custom fastText embeddings of 100 dimensions (§6.3.3). Both models classify tokens with the ***Linguistic*** labels using a Classifier Chain with the Random Forest algorithm (with parameter `random_state` set to `22`).

The model representing the tokens of HC Archives documentation with custom fastText embeddings yielded better performance scores overall than model without embeddings (Table 6.13). The macro F_1 score of the model with custom embeddings was 0.607, an improvement of 0.044 over the model without embeddings. The micro F_1 score of the model with custom embeddings was 0.715, an improvement of 0.093 over the model without embeddings.

[32] Additional annotated data are needed to consider grammatically non-binary terminology.

model	macro prec.	macro rec.	macro F1	micro prec.	micro rec.	micro F1
None	0.564	0.561	0.563	0.624	0.619	0.622
fastText	**0.712**	**0.589**	**0.607**	**0.763**	**0.673**	**0.715**

Table 6.13: **Comparing word representations in multilabel token classification of *Linguistic* labels.** Macro and micro precision (prec.), recall (rec.), and F_1 scores for multilabel token classification with no word embeddings and custom fastText word embeddings of 100 dimensions. Both models are a Classifier Chain with the Random Forest algorithm (`random_state = 22`; CC-RF), trained to classify tokens with ***Linguistic*** labels (*Gendered Pronoun*, *Gendered Role*, *Generalization*. The highest scores per metric are in bold. Scores are calculated on the validation subset of the token dataset strictly.

Appendix G reports strict evaluation measures per label for the multilabel token classification models without any embeddings (Table G.1) and with fastText embeddings (Table G.2). For the *Gendered Pronoun* label, the model without any embeddings yielded a higher F_1 score by 0.005. For *Gendered Role*, the model with custom embeddings has an F_1 score 0.149 higher than the model without embeddings for this label. For *Generalization*, the model without embeddings has the highest F_1 score at 0.277, while the model with custom embeddings has an F_1 score of 0.267. Nonetheless, the model with custom embeddings has the highest macro and micro precision and recall scores, in addition to the highest macro and micro F_1 scores.

Due to the results above indicating the stronger performance of a model with custom word embeddings, combined with this thesis' prioritization of representativeness over convenience, all subsequent models for classifying ***Linguistic*** labels use the custom fastText embeddings as word representations.

6.5.1.2 Algorithm Experiment

To determine which algorithm would be best for the multilabel token classification task, I experimented with two algorithms available for use with the Classifier Chain model using scikit-multilearn (Szymański and Kajdanowicz, 2018). I chose the Classifier Chain because the scikit-multilearn documentation stated that this model can generalize beyond the combinations of labels provided in training data and considers relationships between labels. Furthermore, Madjarov et al.'s (2012) comparisons of 12 multilabel learning methods showed the Classifier Chain to be a relatively simple and

model	macro prec.	macro rec.	macro F_1	micro prec.	micro rec.	micro F_1
CC-RF	**0.712**	**0.589**	**0.607**	**0.763**	**0.673**	**0.715**
CC-PA	0.433	0.279	0.337	0.617	0.336	0.435

Table 6.14: **Comparison of algorithms for *Linguistic* labels, strictly evaluated.** Macro and micro precision (prec.), recall, and F_1 scores of Classifier Chain models with Passive Aggressive (CC-PA) and Random Forest (CC-RF) algorithms for annotating ***Linguistic*** labels (*Gendered Pronoun*, *Gendered Role*, *Generalization*). The highest score per column is in bold.

high-performing approach to multilabel classification.

I experimented with two algorithms in combination with the Classifier Chain method: Random Forest and Passive Aggressive, both using 100-dimension custom fastText embeddings. Random Forest was also a strong performer in Madjarov et al.'s (2012) comparisons. Passive Aggressive performed well on Person Name and Occupation labels in my sequence classification experiments with the CRF model (Table G.6), and as an online learning algorithm that processes data sequentially, it provides a contrast to Random Forest, which repeatedly creates decision trees based on subsets of data.

The experiment's results show that Random Forest yielded better performance than Passive Aggressive across macro and micro precision, recall, and F_1 scores by a range of 0.337 to 0.146 (Table 6.14). For scores per label, see tables G.3 and G.2. Due to these results, I use Random Forest in combination with the Classifier Chain method for all subsequent Linguistic Classifiers in this chapter.

6.5.2 Person Name and Occupation Classification

I experimented with multiclass sequence classification model setups to automatically annotate tokens with *Feminine*, *Masculine*, and *Unknown* labels from the Taxonomy's ***Person Name*** category,[33] and the *Occupation* label from the Taxonomy's ***Contextual*** category. Intuitively, ***Person Name*** and *Occupation* labels suit sequence classification because the text spans annotated with these labels ranged in length from one to ten words, due to lengthy titles, such as "The Very Rev. Andrew N. Nisbet D.D." from the fonds titled *The Papers of Andrew Nisbet Bogle* (Identifier: Coll-1004), or highly specific job names, such

[33] As described in the previous chapter, the *Non-binary* label was not applied during manual annotation though it was included in the Taxonomy's ***Person Name*** category.

as "chair of practical theology and Christian ethics" from the fonds titled *Papers of The Very Rev Prof James Whyte (1920-2005)* (Identifier: AA5). Additionally, each occurrence of a person's name should have only one label from the ***Person Name*** category, and a token should not have both an *Occupation* label and a ***Person Name*** label, aligning with the multiclass classification task, where each token can be annotated with at most one label.

The classification task with ***Person Name*** labels aims to detect gendered language, focusing on the grammatical gender of terms that refer to a person. Thus, as with the ***Linguistic*** labels, this provides an approach to analyzing gender bias at a high level, quantifying the presence of *Feminine*- and *Masculine*-labeled names to compare the prevalence of grammatical genders across the archival documentation. The *Occupation* label provides the opportunity to analyze associations between job titles represented in HC Archives documentation and gender biased language found through classification with *Omission* and *Stereotype* labels (such as Garg et al. (2018) and Lewis and Lupyan's (2020) investigation of correlations between occupations and stereotypes using word embeddings). The model setup experiments for ***Person Name*** and *Occupation* labels included word representation experiments (§6.5.2.1) and algorithm experiments (§6.5.2.2).

6.5.2.1 Word Representation Experiment

As with the classification of ***Linguistic*** labels, I compared multiclass sequence classification models using no embeddings and custom fastText embeddings of 100 dimensions for ***Person Name*** and *Occupation* labels. Both models were CRF models (Lafferty et al., 2001) with the AROW algorithm (Crammer et al., 2013) (the next section, §6.5.2.2, explains this choice of algorithm).

When looking at the macro and micro scores across all ***Person Name*** and *Occupation* labels' `B-[LABELNAME]` and `I-[LABELNAME]` tags, the model with custom fastText embeddings performed best (Table 6.15). The model with custom fastText embeddings outperformed the model without embeddings across all metrics by a range of 0.019 to 0.041. When looking at precision, recall, and F_1 scores of the models per label, the model with fastText word embeddings yielded better scores for all labels except *Occupation* (Table 6.16). The model without embeddings yielded precision and F_1 scores for *Occupation* (calculated as the average of those scores for the `B-Occupation`

model	macro prec.	macro recall	macro F_1	micro prec.	micro recall	micro F_1
None	0.890	0.826	0.856	0.887	0.810	0.847
fastText	**0.913**	**0.867**	**0.889**	**0.906**	**0.842**	**0.873**

Table 6.15: **Comparison of word representations for *Person Name* and *Occupation* classification, strictly evaluated.** Macro and micro precision (prec.), recall (rec.), and F_1 scores of CRF models with the AROW algorithm (`variance = 1`) using no word embeddings (None) and custom word embeddings (fastText) to annotate with ***Person Name*** (*Feminine*, *Masculine*, *Unknown*) and *Occupation* labels. The highest score per metric is in bold. Scores are calculated on the validation subset of the tagged token dataset strictly.

	none			fastText		
label	precision	recall	F_1	precision	recall	F_1
Feminine	0.909	0.872	0.889	**0.930**	**0.902**	**0.915**
Masculine	0.789	0.652	0.714	**0.842**	**0.769**	**0.804**
Unknown	0.891	0.819	0.853	**0.912**	**0.833**	**0.871**
Occupation	**0.973**	0.961	**0.967**	0.967	**0.966**	0.966

Table 6.16: **Comparison of word representations for *Person Name* and *Occupation* classification per label, strictly evaluated.** Precision (prec.), recall, and F_1 scores averaged across the ***Person Name*** (*Feminine*, *Masculine*, *Unknown*) and *Occupation* labels for CRF models with the AROW algorithm (`variance = 1`) without word embeddings ("none") and with custom fastText embeddings ("fastText"). For each label, the highest score per metric is in bold. Per metric, each label's score is the average of that label's `B-[LABELNAME]` and `I-[LABELNAME]` tags' scores. Scores are calculated on the validation subset of the tagged token dataset strictly.

and `I-Occupation` tags) that are 0.006 and 0.001 higher than those of the model with embeddings. Appendix G reports strict evaluation measures per `B-[LABELNAME]` and `I-[LABELNAME]` tag for the multiclass sequence classification models without any embeddings (Table G.4) and with custom fastText embeddings (Table G.5).

When looking at the macro and micro scores across all ***Person Name*** and *Occupation* labels' `B-[LABELNAME]` and `I-[LABELNAME]` tags, the model with custom fastText embeddings performed best (Table 6.15). The model with custom fastText embeddings outperformed the model without embeddings across all metrics by a range of 0.019 to 0.041. When looking at precision, recall, and F_1 scores of the models per label, the model with fastText word

embeddings yielded better scores for all labels except *Occupation* (Table 6.16). The model without embeddings yielded precision and F_1 scores for *Occupation* (calculated as the average of those scores for the `B-Occupation` and `I-Occupation` tags) that are 0.006 and 0.001 higher than those of the model with embeddings. Appendix G details the strict evaluation measures per `B-[LABELNAME]` and `I-[LABELNAME]` tag for the multiclass sequence classification models without any embeddings (Table G.4) and with fastText embeddings (Table G.5).

Considering these results, which confirm the value of using word embeddings to represent word meanings, alongside my prioritization of representativeness over convenience, all subsequent models for classifying with ***Person Name*** and *Occupation* labels use custom fastText embeddings.

6.5.2.2 Algorithm Experiment

The Python library sklearn-crfsuite provides four algorithms with suggested parameter values that I experimented with to determine the best model setup for sequence classification of ***Person Name*** and *Occupation* labels. These algorithms are: gradient descent using the L-BFGS method (LBFGS), Stochastic Gradient Descent with an L2 regularization term (L2SGD), Averaged Perceptron (AP), Passive Aggressive (PA), and Adaptive Regularization of Weight Vector (AROW) (§6.3.1). I ran each algorithm using one to three different parameters, choosing parameter values based on guidance in the sklearn-crfsuite documentation (Korobov, 2015).

To determine the highest-performing algorithm and parameter combination, I ran 11 models total. For computational efficiency, I ran each model with a maximum of 50 iterations and with custom fastText embeddings of 50 dimensions, rather than 100 iterations and 100 dimensions. The AROW algorithm with the parameter `variance` set to `1` yielded the best performance when measuring by macro and micro F_1 scores, with scores of 0.513 and 0.492, respectively (Table G.6). As a result, I chose to use the AROW algorithm with `variance` set to `1` for all remaining CRF models classifying ***Person Name*** and *Occupation* labels in this chapter.

To align with the previously reported model setups, next I ran the CRF model with the AROW algorithm (`variance = 1`) using the default value for the `max_iterations` parameter, `100`, and using custom fastText embeddings of

tag	FN	FP	TP	precision	recall	F_1
B-Feminine	44	48	474	0.908	0.915	0.912
I-Feminine	124	51	990	0.951	0.889	0.919
B-Masculine	296	179	1042	0.853	0.779	0.814
I-Masculine	441	282	1392	0.832	0.759	0.794
B-Unknown	404	205	1994	0.907	0.832	0.868
I-Unknown	644	295	3246	0.917	0.834	0.874
B-Occupation	22	29	738	0.962	0.971	0.967
I-Occupation	35	25	846	0.971	0.960	0.966
macro				0.913	0.867	0.889
micro				0.906	0.842	0.873

Table 6.17: **CRF model performance with `algorithm = AROW, variance = 1` for *Person Name* and *Occupation* classification, per tag, strictly evaluated.** Precision, recall, and F_1 scores are reported for `B-[LABELNAME]` and `I-[LABELNAME]` tags.

label	precision	recall	F_1
Feminine	0.930	0.902	0.915
Masculine	0.843	0.769	0.804
Unknown	0.912	0.833	0.871
Occupation	0.967	0.966	0.966

Table 6.18: **CRF model performance with `algorithm = AROW, variance = 1` for *Person Name* and *Occupation* classification, per label, strictly evaluated.** Precision, recall, and F_1 scores from Table 6.17 are averaged across the `B-[LABELNAME]` and `I-[LABELNAME]` tags for each label.

100 dimensions. The F_1 scores per `B-[LABELNAME]` and `I-[LABELNAME]` tags range from 0.794 for `I-Masculine` to 0.967 for `B-Occupation` (Table 6.17). Macro averaging tags by their associated label, this model yielded the best performance for the *Occupation* label, secondly the *Feminine* label, thirdly the *Unknown* label, and lastly the *Masculine* label (Table 6.18). I used this same model setup for all subsequent Person Name and Occupation Classifiers reported in this chapter.

6.5.3 Omission and Stereotype Classification

I experimented with multilabel document classification setups to automatically annotate tokens with the gender biased language labels from the Taxonomy's

algorithm	macro prec.	macro rec.	macro F_1	micro prec.	micro rec.	micro F_1
LR	**0.918**	0.549	0.687	**0.897**	0.524	0.661
RF	0.417	0.003	0.006	0.833	0.004	0.009
SVM	0.888	**0.624**	**0.732**	0.873	**0.592**	**0.705**

Table 6.19: **Comparison of algorithms for *Omission* and *Stereotype* labels.** Macro and micro precision (prec.), recall (rec.), and F_1 scores for multilabel document classifiers annotating *Omission* and *Stereotype* labels using Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) algorithms. The highest scores per metric are in bold.

Contextual category: *Omission* and *Stereotype*. These labels were chosen for document, classification due to their contextual nature. Little could be learned from token classification with these labels, because the information surrounding the annotated text spans provided the impetus for annotators to label those text spans. Each document is one description from a "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information" metadata field in HC Archives documentation. As a multilabel task, a model may classify a description with zero, one, or both labels. The model setup experiment for classifying with *Omission* and *Stereotype* labels is an experiment with three algorithms.

6.5.3.1 Algorithm Experiment

To determine the most suitable algorithm for multilabel document classification, I experimented with Logistic Regression, Random Forest, and SVM (§6.3.1). I chose Logistic Regression because it is the baseline algorithm for classification, and I chose Random Forest and SVM because they have been shown to perform well relative to deep learning models on document classification in multiple domain areas (Adhikari et al., 2019; Madjarov et al., 2012). Furthermore, scikit-learn recommends these three algorithms for supervised text classification (scikit-learn developers, 2023b). For all models, I represented descriptions as TF-IDF matrices (§6.3.1).

Overall, SVM yielded the highest performance with a macro F_1 of 0.732 and micro F_1 of 0.705 (Table 6.19). SVM also yielded the highest macro and micro recall scores, while Logistic Regression yielded the highest macro and micro precision scores. Per label (Table 6.20), SVM yielded the highest F_1 scores for *Omission* and *Stereotype*, 0.667 and 0.797, respectively, as well as the highest

labels	algorithm	FN	FP	TP	precision	recall	F_1
Omission	LR	409	61	395	**0.866**	0.491	0.627
Omission	RF	799	1	5	0.833	0.006	0.012
Omission	SVM	362	79	442	0.848	**0.550**	**0.667**
Stereotype	LR	124	6	191	**0.970**	0.606	0.746
Stereotype	RF	315	0	0	0.000	0.000	0.000
Stereotype	SVM	95	17	220	0.928	**0.698**	**0.797**

Table 6.20: **Comparison of algorithms for *Omission* and *Stereotype* labels, per label.** False Negative (FN), False Positive (FP), and True Positive (TP) counts and precision, recall, and F_1 scores for multilabel document classifiers annotating *Omission* and *Stereotype* labels using Logistic Regression (LR), Random Forest (RF), and Support Vector Machines (SVM) algorithms. The highest precision, recall, and F_1 scores per label are in bold.

recall scores, 0.550 and 0.698, respectively. Logistic Regression yielded the highest precision scores, though, at 0.866 for *Omission* and 0.970 for *Stereotype*.

Based on the above results showing the SVM model yielding the highest overall scores, subsequent document classification models in this chapter use SVM for classifying descriptions with *Omission* and *Stereotype* labels. That being said, if one wished to optimize for precision, rather than F_1 score (a combination of precision and recall), Logistic Regression would be the most suitable choice, as this algorithm yielded the highest precision scores for *Omission* (0.866) and *Stereotype* (0.970). Higher precision scores than recall scores indicate that a model is more likely to miss an annotation than make a mistaken annotation; in other words, when a model makes an annotation, it is highly likely to be correct. Nonetheless, I aim to optimize for F_1 score in this chapter, so the next section's document classification models use SVM.

6.6 Results: Model Cascades

Next, I used the highest-performing model setups from the experiments above to create cascades, or sequential combinations, of models. The cascades enable me to investigate the extent to which (1) annotating grammatically and lexically gendered language (***Linguistic*** labels) informs the annotation of people's associated gender group (***Person Name*** labels) and people's occupations (*Occupation* labels), and (2) annotating gendered language

(***Linguistic*** and ***Person Name*** labels) and occupations informs the annotation of gender biased language (*Stereotype* and *Omission* labels). I created three cascades:

- **Cascade 1: Linguistic Classifier to Person Name and Occupation Classifier to Omission and Stereotype Classifier (LC > PNOC > OSC):** I ran the Linguistic Classifier and then passed its predictions (*Gendered Pronoun*, *Gendered Role*, and *Generalization* annotations) to the Person Name and Occupation Classifier as features. Then, I ran the Person Name and Occupation Classifier and passed its predictions (*Feminine*, *Masculine*, *Unknown*, and *Occupation* annotations) and the previous model's predictions (*Gendered Pronoun*, *Gendered Role*, and *Generalization* annotations) to the Omission and Stereotype Classifier as features.

- **Cascade 2: Linguistic Classifier to Omission and Stereotype Classifier (LC > OSC):** I ran the Linguistic Classifier and then passed its predictions to the Omission and Stereotype Classifier as features.

- **Cascade 3: Person Name and Occupation Classifier to Omission and Stereotype Classifier (PNOC > OSC):** I ran the Person Name and Occupation Classifier and then passed its predictions to the Omission and Stereotype Classifier as features.

In §6.6.1, §6.6.2, and §6.6.3, I report and compare the performance scores of each cascade for classifying descriptions as *Stereotype* and *Omission*, as well as reporting and comparing the performance scores of the individual classifiers across cascades, and the classifiers' baselines, meaning models that do not have any of the Taxonomy's labels passed to them as features. The scores I report for the model cascades are based on model predictions over the entire aggregated dataset using five-fold cross-validation. As explained in §6.4, five-fold cross validation involves running five instances of a model to generate predictions (model-made annotations) for the entire dataset, providing a more robust comparison of automated (model-made) and manual annotation performances. Performance scores are calculated both strictly, meaning a model's annotation must exactly match the text span that was manually annotated with the same label to be a TP, and loosely, meaning a model's annotation can exactly match

or overlap a text span that was manually annotated with the same label to be a TP. As described in Chapter 5, I consider the loose evaluation to be more important than the strict evaluation due to the subjectivity of bias classification and the greater importance of identifying the presence of biased language over the precise tokens that communicate bias. I thus report loose evaluations in this chapter and strict evaluations in the appendices.

Figures 6.2, 6.3, and 6.4 illustrate annotations from Cascade 1, Cascade 2, and Cascade 3, respectively, using three example descriptions from the HC Archives documentation. ***Linguistic*** annotations are yellow, ***Person Name*** annotations are green, and ***Contextual*** annotations are blue. The *Omission* and *Stereotype* labels are made at the description level and the remainder of the Taxonomy's labels are made at the word level, with *Person Name* and *Occupation* `B-[LABELNAME]` and `I-[LABELNAME]` tags generalized to their label.

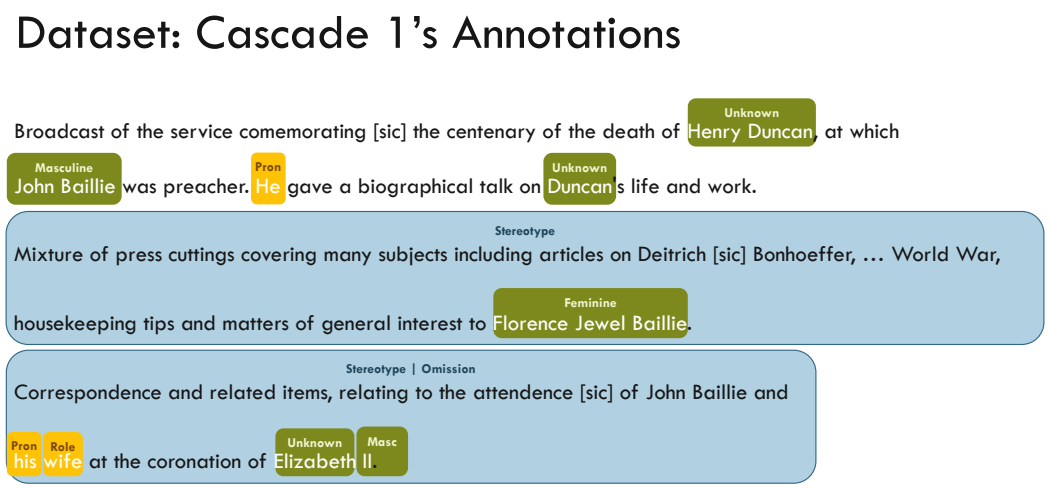


Figure 6.2: **Annotations with Cascade 1's Classifiers.** In these example descriptions, one label is incorrect: *Masculine* on "II" should be *Unknown*, and three labels are missing: *Occupation* for "preacher," *Unknown* for "Deitrich [sic] Bonhoeffer," and *Unknown* for "John Baillie."

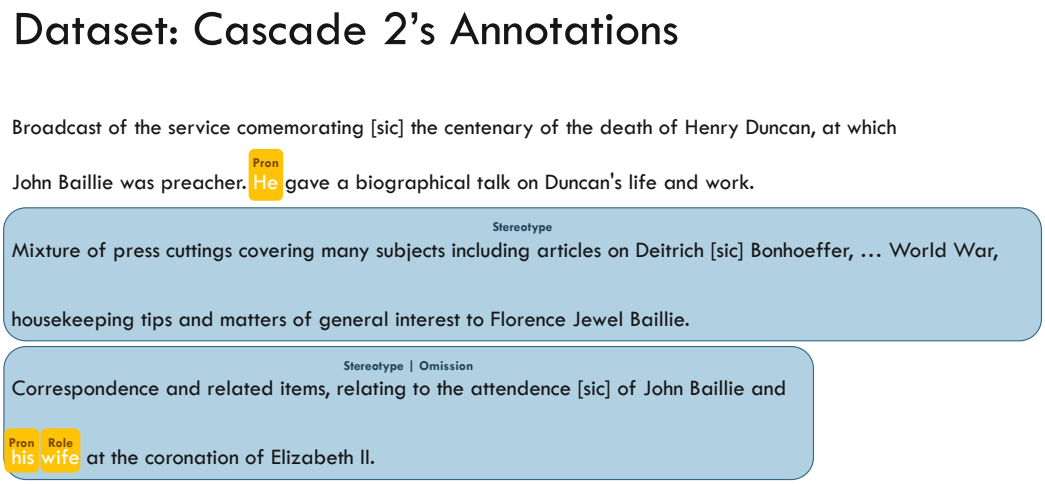


Figure 6.3: **Annotations with Cascade 2's Classifiers.** In these example descriptions, all the provided labels are correct and no labels are missing.

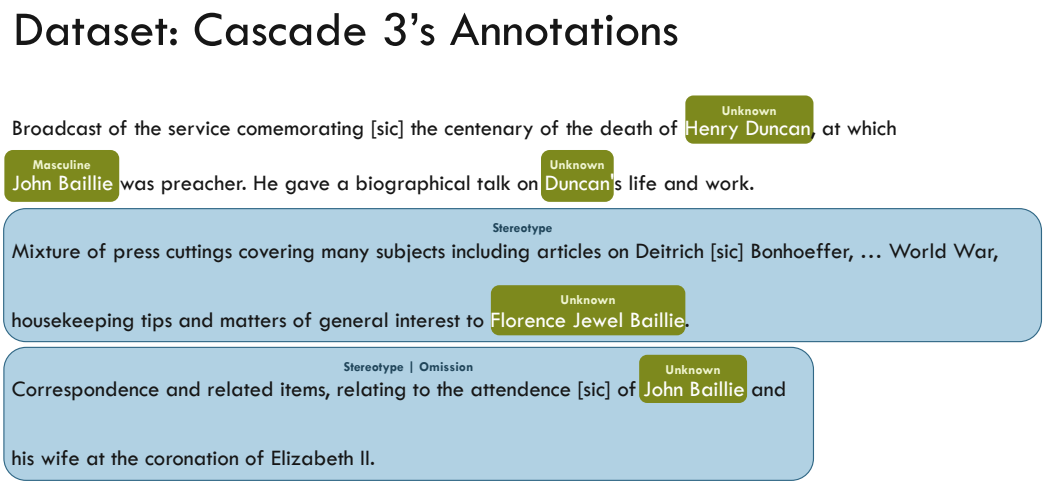


Figure 6.4: **Annotations with Cascade 3's Classifiers.** In these example descriptions, all provided labels are correct and three labels are missing: *Occupation* for "preacher," *Unknown* for "Dietrich Bonhoeffer," and *Unknown* for "Elizabeth II."

6.6.1 Cascade 1: LC > PNOC > OSC

6.6.1.1 Linguistic Classifier

The Linguistic Classifier is a multilabel token classification model trained to annotate tokens with the Taxonomy's ***Linguistic*** category of labels: *Gendered Pronoun*, *Gendered Role*, and *Generalization*. I preprocessed the token data for this Classifier as described in §6.4.1. For training the Classifier, the training data's embeddings were passed to the model as features and the binarized labels were passed to the model as targets, meaning the classes with which the model should learn to annotate text. The Linguistic Classifier is a Classifier Chain model using a Random Forest algorithm (`random_state = 22`) with scikit-learn's default parameters, the highest-performing model setup from earlier experiments (§6.5.1.2). I used five-fold cross-validation to train five instances of the Classifier on four folds (80%) of the data and test on one fold (20%), rotating which folds were in the training and test sets for each instance of the Classifier. This way, each fold was a test set for one model instance, providing me with predictions, meaning tokens classified with Linguistic labels, for 100% of the data. I saved the final, fifth instance of the model for reuse, using Joblib[34] as recommended in scikit-learn's documentation (Pedregosa et al., 2011).

For testing, I compared the Classifier's predictions (across all five model instances, or 100% of the data) to the manual annotations in the aggregated dataset at the token and annotation level. Predictions are evaluated strictly at the token level (Table G.10) and loosely at the annotation level (Table 6.21). As with the manual annotation process and the experiments described earlier (§6.5), to evaluate the Linguistic Classifier, I report FN, FP, and TP, as well as precision, recall, and F_1 scores.

The Classifier's annotation of ***Linguistic*** labels performs strongly overall, with macro and micro scores well above 0.500 (or 50%, as 1 is the highest score, so better than random chance), ranging from 0.677 to 0.798 (Table 6.21). The results show that annotating gendered language proved reliable, with a macro average F_1 score across the *Gendered Pronoun* and *Gendered Role* labels of 0.833; while annotating gender biased language in the form of *Generalization* proved unreliable, with an F_1 score well below 0.500 at 0.341.

[34] joblib.readthedocs.io

label	FN	FP	TP	precision	recall	F1
Gendered Pronoun	29	851	3654	0.811	0.992	0.893
Gendered Role	535	791	2255	0.740	0.808	0.773
Generalization	1010	167	305	0.646	0.232	0.341
macro				0.733	0.677	0.669
micro				0.775	0.798	0.786

Table 6.21: **Cascades 1 and 2, *Linguistic* Classifier performance, loosely evaluated.** Performance scores for multilabel token classification with ***Linguistic*** labels. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across labels. Scores are calculated loosely, meaning a model's annotation is considered correct if it matches or overlaps with a manual annotation of the same label.

	between annotators			annotators vs. aggregated		
label	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
Gendered Pronoun	0.961	0.954	0.957	0.986	0.956	0.971
Gendered Role	0.784	0.804	0.780	0.803	0.833	0.811
Generalization	0.381	0.295	0.277	0.981	0.138	0.236

Table 6.22: **Inter-Annotator Agreement (IAA) for manual annotation with Linguistic labels, loosely evaluated.** In the "between annotators" columns, IAA scores between annotators 0 and 1, 0 and 2, and 1 and 2 were averaged to get the precision, recall, and F_1 scores displayed. In the "annotators vs. aggregated" columns, IAA scores between annotator 0 and the aggregated dataset, annotator 1 and the aggregated dataset, and annotator 2 and the aggregated dataset were averaged to get the precision, recall, and F_1 scores displayed. Scores are calculated loosely, meaning one annotation agrees with another annotation if it exactly matches or overlaps that other annotation, and both annotations have the same label.

Classification of tokens with the *Gendered Pronoun* label is highest, with precision, recall, and F_1 scores ranging from 0.811 to 0.992 (Table 6.21). The higher recall score (0.992) than precision score (0.811) indicates that the model is more robust, or more sensitive, than it is precise. Error analysis should be conducted to determine what types of manually-annotated *Gendered Pronouns* the model is missing; then, additional examples of those types could be added to the training data in an effort to improve the Classifier's precision score. Relative to the annotators' agreement amongst themselves and with the aggregated dataset (Table 6.22), the Classifier's *Gendered Pronoun* performance

has lower precision and F_1 scores but a higher recall score. The Classifier's high performance is unsurprising given the high Inter-Annotator Agreement (IAA) scores for manually annotating *Gendered Pronouns*, which were higher than all other Taxonomy labels, indicating that the aggregated data would provide clear annotation patterns for training a model to annotate *Gendered Pronouns*.

Classification of tokens with the *Gendered Role* label performs second-best among the ***Linguistic*** labels, with precision, recall, and F_1 scores ranging from 0.740 to 0.808 (Table 6.21). As with *Gendered Pronoun*, the higher recall score (0.808) than precision score (0.740) for *Gendered Role* indicates that the model is more robust than it is precise, and error analysis should be conducted to determine what types of manually-annotated *Gendered Roles* could be added to augment the model training data. Relative to the IAA scores (Table 6.22), the Classifier's *Gendered Role* performance scores are lower, except the Classifier's recall score, which falls between the annotators' agreement amongst themselves and the annotators' agreement with the aggregated dataset. Nonetheless, the Classifier performs well, with scores well above 0.500. Again, this performance is unsurprising given the high IAA scores for *Gendered Role* which, though less than those for *Gendered Pronoun*, were still high enough to indicate consistency in manual annotation that provides clear training data for a model.

Classification of tokens with the *Generalization* label shows the worst performance, with precision, recall, and F_1 scores of 0.646, 0.232, and 0.341, respectively. Nonetheless, the Classifier's precision score is fairly high, indicating that most (64.6%) of its annotations were correct. Classification with the *Generalization* label reflects the challenges of applying this label during the manual annotation process. IAA scores for *Generalization* were the lowest of all the Taxonomy labels' IAA scores. The Classifier's performance with the *Generalization* label overall is higher than the IAA scores, though, by a range of 0.064 to 0.105. This indicates that the changes I made to the annotation instructions for labeling with the *Generalization* label during my manual review and aggregation of the annotated data (Chapter 5) did improve the consistency of annotations with this label in the training data. However, with an F_1 score less than 0.500, error analysis should be conducted to further understand how the training data could be improved for classification with the *Generalization* label.

6.6.1.2 Person Name and Occupation Classifier

The Person Name and Occupation Classifier is a multiclass sequence classification model trained to annotate tokens with the *Feminine*, *Masculine*, and *Unknown* labels from the Taxonomy's ***Person Name*** category, and the *Occupation* label from the Taxonomy's ***Contextual*** category. In addition to the preprocessing described in §6.4.2, I joined the *Gendered Pronoun*, *Gendered Role*, and *Generalization* annotations from the Linguistic Classifier to the data to pass them to the Person Name and Occupation Classifier as features. During manual annotation, annotators were instructed to apply ***Person Name*** labels based on the gendered terms referring to a person's name within the description in which that name appeared. Aligning with this practice, the ***Linguistic*** labels are associated with a description, so every sentence in a description that is input as training data to the Person Name and Occupation Classifier has a ***Linguistic*** label. When grouping the data by token, I ensured that any token with multiple ***Linguistic*** labels had a single ***Linguistic*** label string value (e.g. `"Gendered Role, Generalization"` rather than `"Gendered Role," "Generalization"`).

For training, each token's embedding, ***Linguistic*** label, token, `START` boolean, `END` boolean, and a bias value of 1.0 (as suggested in the sklearn-crfsuite documentation (Korobov, 2015)) were passed as features to the Person Name and Occupation Classifier, grouped by sentence in a list where each token's features appeared in the order the tokens appeared in the sentence. The Classifier uses the highest-performing model setup from earlier experiments (§6.5.2.2): a CRF model using the AROW algorithm with the parameters `variance` set to `1`, `max_iterations` set to `100`, and `all_possible_transitions` set to `True`. I used the same five-fold cross-validation approach as with the Linguistic Classifier, training and testing five instances of the Person Name and Occupation Classifier and saving the fifth instance of the model for reuse with Joblib.

For testing, I compared the Classifier's predictions to the manual annotations in the aggregated dataset at the token and annotation level. This chapter reports scores for the loose evaluation at the annotation level (Table 6.23); strict evaluation scores at the token level can be found in Appendix G (Table G.12).

label	FN	FP	TP	precision	recall	F_1
Feminine	553	1208	1146	0.487	0.675	0.566
Masculine	3402	3167	2247	0.415	0.398	0.406
Unknown	5982	3581	5170	0.591	0.464	0.520
Occupation	1278	901	1687	0.652	0.569	0.608
macro				0.536	0.526	0.525
micro				0.536	0.478	0.505

Table 6.23: **Cascade 1, Person Name and Occupation Classifier performance, loosely evaluated.** Performance scores for multilabel token classification of ***Person Name*** and *Occupation* labels using ***Linguistic*** labels as features. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across labels.

	between annotators			annotators vs. aggregated		
label	**precision**	**recall**	**F_1**	**precision**	**recall**	**F_1**
Feminine	0.604	0.601	0.597	0.992	0.682	0.807
Masculine	0.686	0.667	0.665	0.997	0.696	0.817
Unknown	0.677	0.725	0.678	0.996	0.800	0.879
Occupation	0.660	0.749	0.698	0.934	0.809	0.866

Table 6.24: **Inter-Annotator Agreement (IAA) for manual annotation with *Person Name* and *Occupation* labels, loosely evaluated.** In the "between annotators" columns, IAA scores between annotators 0 and 1, 0 and 2, and 1 and 2 were averaged to get the precision, recall, and F_1 scores displayed. In the "annotators vs. aggregated" columns, IAA scores between annotator 0 and the aggregated dataset, annotator 1 and the aggregated dataset, and annotator 2 and the aggregated dataset were averaged to get the precision, recall, and F_1 scores displayed.

The Person Name and Occupation Classifier yielded lower performance scores than the Linguistic Classifier, with a macro F_1 score of 0.525 and a micro F_1 score of 0.505 when evaluated loosely (Table 6.23). Per label, the Classifier annotated *Occupation* best with an F_1 score of 0.608, then *Feminine* (0.042 less), then *Unknown* (0.088 less), and then *Masculine* (0.202 less). Compared to the manual annotation process, the Classifier's F_1 scores per label are lower than the IAA scores per label (Table 6.24). The Classifier's F_1 score for the *Feminine* label is closest to that of the IAA between annotators, being only 0.031 lower. This similarity in scores could be due to the smaller number of

Feminine-labeled names relative to *Masculine*-labeled names in the manually annotated data, which likely results in less variation in *Feminine*-labeled names that makes it easier for a model to identify such names. There are far more *Unknown*-labeled names than *Feminine*-labeled names, yet the model performs worse with the *Unknown* label, further indicating that *Feminine*-labeled names have less variation than those labeled *Unknown*, because even with fewer examples the Classifier was able to pick up on data patterns indicating *Feminine* better than data patterns indicating *Unknown*.

The lower scores overall relative to the IAA scores, particularly for ***Person Name*** labels, are likely due to inconsistencies in the manual annotations with these labels. Error analysis indicated frequent confusion of *Masculine* and *Unknown* labels, with many names being labeled *Masculine* that should, in fact, have been *Unknown*. There were also instances where the Classifier correctly annotated a name as *Unknown* but the aggregated dataset had a *Masculine* label for that name, resulting in mistakenly higher FP and FN counts. Although this confusion occurs between *Feminine* and *Unknown* labels, there are fewer names annotated as *Feminine* than *Masculine*, and there are fewer total descriptions in which one *Feminine* name reoccurs, so the Classifier's ability to distinguish *Feminine* names is less impacted by this confusion. Future work should aim to improve the consistency of ***Person Name*** annotations in model training data.

The lower results of the Person Name and Occupation Classifier relative to the Linguistic Classifier was unsurprising given the nature of the language each classifier was intended to annotate. While ***Linguistic*** labels often annotate a single word, ***Person Name*** and *Occupation* labels often annotate longer word sequences, adding complexity to the patterns in the training data that a model must learn. Future work could experiment with including only *Gendered Pronoun*- and *Gendered Role*-labeled text as features to the Person Name and Occupation Classifier. Since *Generalization* was most difficult for the Linguistic Classifier to annotate with, removing it from the Person Name and Occupation Classifier's features could reduce inconsistencies in the feature data, potentially improving the Person Name and Occupation Classifier's performance.

6.6.1.3 Omission and Stereotype Classifier

The Omission and Stereotype Classifier is a multilabel document classification model trained to annotate descriptions with the *Omission* and *Stereotype*

labels from the Taxonomy's ***Contextual*** category. In addition to the preprocessing described in §6.4.3, I associated the ***Linguistic***, ***Person Name***, and *Occupation* labels from the previous two classifiers' predictions to the descriptions, generalizing the ***Person Name*** and *Occupation* labels' tags to labels (e.g. `B-Unknown, I-Unknown, I-Unknown` becomes `Unknown`) and removing duplicates (e.g. if a description had five tokens annotated as *Gendered Pronoun*, that description would have one *Gendered Pronoun* label). I then binarized these description-level labels and concatenated them to the TF-IDF matrix-representations of the descriptions.

For training, I passed the training data's concatenated TF-IDF and feature matrices to the model as features and the binarized *Omission* and *Stereotype* labels as targets. The Classifier is an SVM model, the highest-performing model from the Omission and Stereotype Classification Experiments (§6.5.3). I used the same five-fold cross-validation approach as with the Linguistic Classifier and Person Name and Occupation Classifier to train and test five instances of the Omission and Stereotype Classifier, again saving the fifth instance of the model for reuse with Joblib.

For testing, I compared the Classifier's predictions to the manual annotations in the aggregated dataset at the description level. I also ran a Baseline Omission and Stereotype Classifier (Table 6.27), using the same model setup as described above but without any ***Linguistic***, ***Person Name***, or *Occupation* labels input as features, to compare with the cascades' classifiers.

The Classifier performed well overall, with macro and micro F_1 scores of 0.747 and 0.715, respectively (Table 6.25). Comparing precision, recall, and F_1 scores per label, the Classifier yielded higher performance annotating with the *Stereotype* label than with the *Omission* label. Precision scores exceed recall scores by 0.307 for *Omission* and 0.199 for *Stereotype*. Relative to F_1 scores for agreement between annotators (Table 6.26), the Classifier outperformed the annotation of *Omission* and *Stereotype* labels by 0.227 and 0.389, respectively. However, relative to annotators' agreement with the aggregated dataset, the Classifier's performance is more similar, with an F_1 score of 0.101 less for the *Omission* label and an F_1 score of 0.006 more for the *Stereotype* label. The Classifier's results relative to the IAA scores are unsurprising considering that the motivation for creating a dataset that aggregated all annotators' labels was the subjectivity of gender bias: there were low levels of agreement between

label	FN	FP	TP	precision	recall	F_1
Omission	1829	381	2203	0.853	0.546	0.666
Stereotype	414	76	1187	0.940	0.741	0.829
macro				0.896	0.644	0.747
micro				0.881	0.602	0.715

Table 6.25: **Cascade 1, Omission and Stereotype Classifier performance.** Performance scores for document classification with *Omission* and *Stereotype* labels using ***Linguistic***, ***Person Name***, and *Occupation* labels as features to input into the document classifier. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across both labels.

	between annotators			annotators vs. aggregated		
label	precision	recall	F_1	precision	recall	F_1
Omission	0.531	0.409	0.439	0.997	0.641	0.767
Stereotype	0.464	0.447	0.440	0.995	0.705	0.823

Table 6.26: **Inter-Annotator Agreement (IAA) for manual annotation of *Omission* and *Stereotype* labels.** In the "between annotators" columns, IAA scores between annotators 0 and 3, 0 and 4, and 3 and 4 were averaged to get the or the precision, recall, and F_1 scores displayed. In the "annotators vs. aggregated" columns, IAA scores between annotator 0 and the aggregated dataset, annotator 3 and the aggregated dataset, and annotator 4 and the aggregated dataset were averaged to get the precision, recall, and F_1 scores displayed.

label	FN	FP	TP	precision	recall	F1
Omission	1857	356	2175	0.859	0.539	0.663
Stereotype	406	77	1195	0.939	0.746	0.832
macro				0.899	0.643	0.747
micro				0.886	0.598	0.714

Table 6.27: **Baseline Omission and Stereotype Classifier performance.** Performance scores for document classification with *Omission* and *Stereotype* labels without any additional labels input as features to the Classifier. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across both labels.

annotators yet I found all their annotations adhered to the annotation instructions during my manual review (Chapter 5). For the manual annotators and the Classifier, annotating with the *Stereotype* label proved easier than annotating with the *Omission label*. This, along with the higher precision and lower recall score for Omission, suggests that while most types of *Omissions* are easily recognizable, there is a subset of *Omission* types that is difficult to recognize.

This Classifier performs similar the Baseline Omission and Stereotype Classifier, with equal macro F_1 scores and a micro F_1 score 0.001 higher (Table 6.27). Per label, this Classifier outperforms the Baseline by 0.003 for the *Omission* label, while the Baseline outperforms this Classifier by 0.003 for the *Stereotype* label. This result indicates that including the ***Linguistic***, ***Person Name***, and *Occupation* labels as features for the Omission and Stereotype Classifier is helpful for annotating with *Omission* but not for annotating with *Stereotype*.

6.6.2 Cascade 2: LC > OSC

6.6.2.1 Linguistic Classifier

The Linguistic Classifier in this second cascade is the same as the Linguistic Classifier in Cascade 1. This ensures that any difference in the performance of this cascade's Omission and Stereotype Classifier relative to that of the Cascade 1 is attributable to feature engineering (i.e. the inclusion or exclusion of a Person Name and Occupation Classifier's predictions (model-made ***Person Name*** and *Occupation* annotations) as features. Please refer to tables 6.21 and G.10 for details of this Classifier's performance.

6.6.2.2 Omission and Stereotype Classifier

The Omission and Stereotype Classifier model setup here is the same as the model setup for Cascade 1, except that ***Person Name*** and *Occupation* labels are not input into the Omission and Stereotype Classifier as features. Instead, only the ***Linguistic*** labels, predicted by the Linguistic Classifier (§6.6.1.1), are binarized and input as description-level features to Omission and Stereotype Classifier. Again, I also binarized the *Omission* and *Stereotype* labels to pass

label	FN	FP	TP	precision	recall	F_1
Omission	1764	406	2268	0.848	0.563	0.676
Stereotype	367	93	1234	0.930	0.771	0.843
macro				0.889	0.667	0.760
micro				0.875	0.622	0.727

Table 6.28: **Cascade 2, Omission and Stereotype Classifier performance.** Performance scores for document classification with *Omission* and *Stereotype* labels using ***Linguistic*** labels as features to input into the document classifier. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across both labels.

them as targets for training the Omission and Stereotype Classifier. As with the previous cascade, I represented the descriptions as TF-IDF matrices and concatenated the features to these matrices. Maintaining this same data preprocessing across cascades ensures that any differences in Omission and Stereotype Classifier performances between the cascades is attributable to feature engineering (i.e. the inclusion or exclusion of previous classifiers' predictions as features), rather than from differences in model setups. For this reason, the training and testing processes are also the same as the previous cascade.

As with Cascade 1, the Omission and Stereotype Classifier in this second cascade performs better on the *Stereotype* label than the *Omission* label when comparing precision, recall, and F_1 scores (Table 6.28). Relative to the macro and micro precision, recall, and F_1 scores of Cascade 1's Omission and Stereotype Classifier, the Omission and Stereotype Classifier in this second cascade performs better. In both cascades, the macro, micro, and per label precision scores exceed the macro, micro, and per label recall scores. The macro F_1 score is 0.760, a 0.013 improvement over the Cascade 1, and the micro F_1 score is 0.727, a 0.012 improvement over Cascade 1. Per label, the F_1 score in this second cascade for the *Omission* label is 0.010 higher than Cascade 1 at 0.676; for the *Stereotype* label, 0.014 higher, at 0.843. These results indicate that the ***Linguistic*** labels alone provide better features than the ***Linguistic***, ***Person Name***, and *Occupation* labels combined.

This second cascade's inclusion of ***Linguistic*** labels as features also yields

an improvement over the Baseline Omission and Stereotype Classifier (which has no labels from the Taxonomy as features). The F_1 scores for the *Omission* and *Stereotype* labels in this cascade are higher than those of the Baseline by 0.013 and 0.011, respectively. This suggests that the annotation of gendered language, in the form of *Gendered Pronouns* and *Gendered Roles*, informs the annotation of gender biased language, in the form of *Stereotypes* and *Omissions*.

Relative to the agreement between manual annotators (Table 6.26) for the *Omission* and *Stereotype* labels, Cascade 2's Omission and Stereotype Classifier performs well. For *Omission*, the manual annotators' agreement with the aggregated dataset yielded an F_1 score 0.091 higher than the Classifier. For *Stereotype*, on the other hand, the Classifier's F_1 score is 0.020 higher than the manual annotators' agreement with the aggregated dataset. For the *Omission* and *Stereotype* labels, the Classifier's F_1 scores exceed the manual annotators' agreement amongst themselves by 0.237 and 0.403, respectively. For the same reasons described with Cascade 1 (§6.6.1.3), the Classifier's results relative to the IAA scores are unsurprising.

6.6.3 Cascade 3: PNOC > OSC

6.6.3.1 Person Name and Occupation Classifier

To continue enabling analysis of the impact of previous classifiers' predictions as model features, the Person Name and Occupation Classifier in this third cascade has the same preprocessing and model setup as the Person Name and Occupation Classifier in Cascade 1, except that no ***Linguistic*** labels are input as features to it. As such, this Classifier is a Baseline Person Name and Occupation Classifier.

This Baseline yielded the best performance on the *Feminine* label, then the *Occupation* label, then the *Unknown* label, and lastly the *Masculine* label (Table 6.29). F_1 scores per label range from 0.409 to 0.641. The precision scores are higher than the recall scores for *Masculine* and *Occupation*, but lower than the recall scores for *Feminine* and *Unknown*. This indicates that the Baseline is more sensitive to variation in data patterns for *Feminine* and *Unknown* than for *Masculine* and *Occupation*; and that the Baseline is more likely to miss a *Occupation* or *Masculine* label than make an incorrect annotation with one of those labels.

label	FN	FP	TP	precision	recall	F_1
Feminine	602	627	1097	0.636	0.646	0.641
Masculine	3708	1894	1941	0.506	0.344	0.409
Unknown	4062	4952	7090	0.589	0.636	0.611
Occupation	1188	926	1777	0.657	0.599	0.627
macro				0.597	0.556	0.572
micro				0.586	0.555	0.570

Table 6.29: **Cascade 3, Person Name and Occupation Classifier performance, loosely evaluated.** Performance scores for multiclass sequence classification of ***Person Name*** and *Occupation* labels (without ***Linguistic*** labels as features). This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across both labels.

Looking at the labels overall, the macro and micro precision scores are higher than the macro and micro recall scores. Comparing macro and micro F_1 scores of this Person Name and Occupation Classifier to that of Cascade 1, this Classifier performs better, with a macro F_1 score 0.047 higher at 0.572 and a micro F_1 0.065 higher at 0.570. Relative to the IAA between annotators (Table 6.24), this cascade's Baseline Person Name and Occupation Classifier performs better on *Feminine*, by 0.044, but worse on *Masculine*, *Unknown*, and *Occupation* by a range of 0.066 to 0.256. Relative to the IAA with the aggregated dataset (Table 6.24), this cascade's Baseline Classifier performs worse across all labels by a range of 0.166 to 0.408. For a strict evaluation of the Baseline Person Name and Occupation Classifier, see Table G.11.

Surprisingly, the performance of the Baseline Person Name and Occupation Classifier relative to Cascade 1's Person Name and Occupation Classifier indicates that the inclusion of a Linguistic Classifier's predictions (i.e. annotations with *Gendered Pronoun*, *Gendered Role*, and *Generalization*) as a model's features does not help the model classify with ***Person Name*** and *Occupation* labels. However, future work should investigate whether the following changes could improve the performance of Cascade 1's Person Name and Occupation Classifier over the Baseline: (1) including only gendered language predictions (i.e. *Gendered Pronoun* and *Gendered Role* annotations) as features or (2) improving the consistency of ***Person Name*** annotations in the training data.

label	FN	FP	TP	precision	recall	F_1
Omission	1728	424	2304	0.845	0.571	0.682
Stereotype	374	89	1227	0.932	0.766	0.841
macro				0.888	0.669	0.761
micro				0.873	0.627	0.730

Table 6.30: **Cascade 3, Omission and Stereotype Classifier performance.** Performance scores for document classification of *Omission* and *Stereotype* labels using ***Person Name*** and *Occupation* labels as features to input into the classifier. This table reports False Negative (FN), False Positive (FP) and True Positive (TP) counts and precision, recall, and F_1 scores per label; as well as macro and micro precision, recall, and F_1 scores across both labels.

6.6.3.2 Omission and Stereotype Classifier

The Omission and Stereotype Classifier model setup is the same as this model's setup in Cascade 1, except that ***Linguistic*** labels are not input into this classifier as features. Instead, only the ***Person Name*** and *Occupation* labels, predicted by the Baseline Person Name and Occupation Classifier, are binarized and input as description-level features to Omission and Stereotype Classifier. Again, I also binarized the *Omission* and *Stereotype* labels to pass them as targets for training the Omission and Stereotype Classifier. As with the previous cascade, I represented the descriptions as TF-IDF matrices and concatenated the features to these matrices. The training and testing process is the same as the previous two cascades' Omission and Stereotype Classifiers. Maintaining consistency in preprocessing and model setups across cascades, but varying model features, enables the comparison of Omission and Stereotype Classifiers across cascades, studying the influence of included features on the Classifiers' performances.

This third cascade's Omission and Stereotype Classifier performs well overall, exceeding the performance of all previous Omission and Stereotype Classifiers when measured with macro and micro F_1 scores (Table 6.30). This suggests that including ***Person Name*** and *Occupation* labels as features is more informative for the Omission and Stereotype Classifier than including ***Linguistic*** labels as features. However, per label, the results become more nuanced. This third cascade's Omission and Stereotype Classifier performs slightly worse on the *Stereotype* label, with an F_1 score of 0.841, 0.002 less than that of Cascade 2. This indicates that, for annotating descriptions with

the *Stereotype* label, including ***Linguistic*** labels as features may be more informative than including ***Person Name*** and *Occupation* labels as features. By contrast, the F_1 score for the *Omission* label Cascade 3's Classifier, at 0.682, is higher than the Baseline, Cascade 1, and Cascade 2 Omission and Stereotype Classifiers by a range of 0.005 to 0.019. This indicates that, for annotating descriptions with the *Omission* label, including ***Person Name*** and *Occupation* labels as features may be more informative than including ***Linguistic*** labels as features.

Relative to IAA between annotators (Table 6.26), per label precision, recall, and F_1 scores are higher with this third cascade's Omission and Stereotype Classifier by a range of 0.162 to 0.469. Relative to IAA with the aggregated dataset (Table 6.26), the Classifier's precision, recall, and F_1 scores for *Omission* are lower by a range of 0.070 to 0.152; for *Stereotype*, the Classifier's precision score is lower by 0.062 but the Classifier's recall and F_1 scores are higher by 0.061 and 0.018, respectively. These results again suggest that annotating with the *Stereotype* label is easier than annotating with the *Omission* label.

Figures 6.5 and 6.6 visualize a comparison of the performance of the baseline and cascades' classifiers for classifying gender biased language, in the form of *Omissions* and *Stereotypes*. Though the differences between cascades are small when visualized, in an NLP context, they are considered indicative of potential future directions for improving the classifier cascades. At the time of writing, reliable significance testing approaches for NLP have yet to be established (Søgaard et al., 2014). Standard statistical significance tests rely on the assumptions that data are independent and identically distributed,

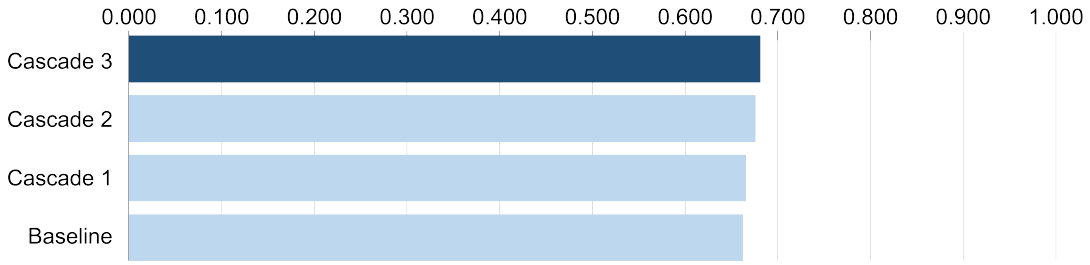


Figure 6.5: **F_1 Scores for Classifying Descriptions with *Omission*.** Cascade 3's Omission and Stereotype Classifier had the best F_1 score for classifying descriptions with the *Omission* label (dark blue bar) relative to the Baseline, Cascade 1, and Cascade 2.

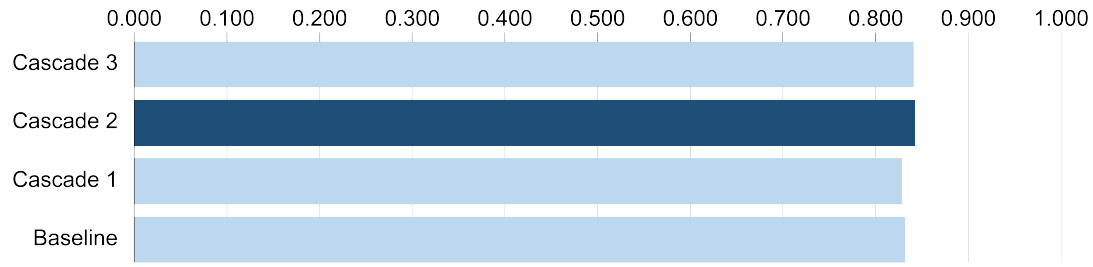


Figure 6.6: **F_1 Scores for Classifying Descriptions with *Stereotype*.** Cascade 2's Omission and Stereotype Classifier had the best F_1 score for classifying descriptions with the *Stereotype* label (dark blue bar) relative to the Baseline, Cascade 1, and Cascade 3.

neither of which are true of my data, nor most data that would be used in an NLP model (Dror et al., 2018). In the next section, I further discuss the implications of the quantitative evaluations of the cascades' performance on the identification of gender biases in text.

6.7 Discussion

Comparing the cascades' performance on classifying gender biased language, namely annotating HC Archives documentation with *Omission* and *Stereotype* labels, Cascade 2 and Cascade 3 yielded the best performance. Though Cascade 3 had the highest macro and micro recall and F_1 scores, as well as the highest F_1 score for the *Omission* label, Cascade 2 is a close second. Cascade 2's macro and micro F_1 scores were only 0.001 and 0.003 less than those of Cascade 3, and Cascade 2's macro and micro precision scores were higher than those of Cascade 3. Additionally, Cascade 2 yielded the highest F_1 score for the *Stereotype* label. Due to Cascade 3's use of a Person Name and Occupation Classier, which had an overall performance not much better than random chance (macro and micro F_1 scores were close to 0.5, or 50%, so similar to random chance), I have greater confidence in Cascade 2's ability to classify gender biased language in the form of *Stereotypes* and *Omissions*. Cascade 2 relied on a Linguistic Classifier, which had macro and micro F_1 scores of nearly 0.7 (70%) and 0.8 (80%), respectively; this performance was much higher than that of Cascade 3's Person Name and Occupation Classifier.

Error analysis on the Person Name and Occupation Classifiers in Cascade 1 and Cascade 3 (the Baseline) indicate that the aggregated dataset contains conflicting ***Person Name*** annotations. This may explain why the incorporation

of ***Linguistic*** labels as features for the Person Name and Occupation Classifier in Cascade 1 did not improve performance over the Baseline Person Name and Occupation Classifier. In future work, additional time could be dedicated to further cleaning of the *Feminine*, *Masculine*, and *Unknown* labels in the aggregated dataset based on the annotation instructions. Nonetheless, the lower performance of the Person Name and Occupation Classifiers relative to the Linguistic Classifiers and Omission and Stereotype Classifiers was unsurprising given the inconsistencies with which the ***Person Name*** labels were applied during manual annotation. The challenges these labels posed throughout the manual annotation (Chapter 5) and classification (Chapter 6) processes indicate the difficulty of the task of applying those labels, which required annotators to apply one of the labels to a name based on grammatically gendered terminology referring to that name within the description in which the name appears (Appendix D). Chapter 7 reports on stakeholder feedback on this approach to applying the labels, suggesting different approaches for future work.

Nonetheless, including gendered language as features, in the form of ***Person Name*** and ***Linguistic*** labels, as well the *Occupation* label, did improve the classification of gender biased language, in the form of *Omission* and *Stereotype* labels. However, providing a subset of these labels as features yielded better performance than providing all these labels as features. The inclusion of ***Linguistic***, ***Person Name***, and *Occupation* labels as features (Cascade 1 in §6.6.1) led to a performance improvement of only 0.003 in F_1 score over the Baseline Omission and Stereotype Classifier's performance for the *Omission* label, and led to a 0.003 decrease in F_1 score for the *Stereotype* label. By contrast, including only the ***Linguistic*** labels as features (Cascade 2 in §6.6.2) led to performance improvements for *Omission* and *Stereotype*, with F_1 scores increasing over the Baseline's by 0.013 and 0.11, respectively. Similarly, including only ***Person Name*** and *Occupation* labels as features (Cascade 3 in §6.6.3) led to performance improvements for *Omission* and *Stereotype* relative to the Baseline, with F_1 scores higher by 0.019 and 0.009, respectively. Looking at the scores for *Omission* and *Stereotype* individually, including ***Person Name*** and *Occupation* labels as features led to the best performance for classifying *Stereotypes*, and including only ***Linguistic*** labels as features led to the best performance for classifying *Omissions*. These results suggest that the

classification of *Omissions* and *Stereotypes* could be improved if these types of gender biased language were classified with two models, one for each label.

Across all cascades' Omission and Stereotype Classifiers, classifying descriptions as *Omission* proved more difficult than classifying descriptions as *Stereotype*. For *Omission*, precision ranged from 0.84 to 0.86, recall ranged from 0.53 to 0.58, and F_1 scores range 0.66 to 0.69. Comparing precision and recall scores, recall was lower by a range of 0.27 to 0.32. The higher precision and lower recall scores suggest that most instances of language that qualifies as *Omission* were easy to classify, but certain instances of language that occur less frequently were more difficult to classify. I posit that incomplete references to people (e.g. "Mrs. Baillie," "a woman") are types of easily-identifiable *Omissions* and completely missing names (e.g. a person's "Biographical / Historical" description naming their father but not their mother), because the words indicating the latter type of *Omission* vary more than the former type. Future work could investigate whether defining multiple labels to account for different types of gender-related *Omissions* could improve a classifier's performance.

Despite there being over twice as many instances of *Omission* (4,032 labels) than *Stereotype* (1,601 labels) in the aggregated data, all Omission and Stereotype Classifiers' performance for *Stereotype* exceeded that of *Omission*. For *Stereotype*, precision ranged from 0.93 to 0.94, recall ranged from 0.74 to 0.78, and F_1 ranged from 0.82 to 0.85. Comparing precision and recall scores, recall was lower by a range of 0.15 to 0.20, much less than the differences between *Omission*'s precision and recall scores. Nonetheless, as with *Omission*, the higher precision scores relative to recall scores suggest that most instances of language that qualifies as *Stereotype* are easy to classify, but certain instances of language that occur less frequently are more difficult to classify. Future work should include error analysis to determine which language patterns the classifiers easily classify and struggle to classify.

My approach to classifying gender biased language demonstrates a new approach to working with annotation inconsistencies, combining annotators' labels rather than selecting one annotator's label as the "correct" label among conflicting annotations. Data perspectivism inspired my approach. Data perspectivism encourages the publication of disaggregated datasets so researchers can analyze annotators' disagreements and utilize annotation

conflicts in model training (Basile, 2022; Basile et al., 2021). In my model creation approach, I built the perspectives of multiple annotators into my text classification models in an effort to work with, rather than eliminate, the subjective, contextual nature of bias and language. I recommend future work on bias in NLP and ML also experiment with data and model creation processes that permit multiple perspectives to be built into an NLP or ML system.

6.8 Conclusion

The results of the experiments with and cascades of classification models demonstrate the feasibility of an alternative to the typical top-down approach of ML model creation. Through a prioritization of quality over quantity, accuracy over efficiency, representativeness over convenience, and situated thinking over universal thinking, I created models that successfully classify types of gender biased language. Moreover, the performance of the classification models reported in this chapter indicate a correlation between gendered language, in the form of pronouns (*Gendered Pronoun*), nouns (*Gendered Role*), and proper names (*Feminine*, *Masculine*, *Unknown*); as well as job titles (*Occupation*); with gender biased language. I recommend future work on gender biased language classification break down the concept of gender bias into more specific types of biased language, so that the varieties of gender biases can be analyzed more closely and thus defined more clearly, and the correlations between the Taxonomy's labels can be better understood.

For the HC Archives, the performance of the Omission and Stereotype Classifiers shows promise for supporting collection reviews, currently a manual process of reviewing collection documentation for biases. Across cascades, the Omission and Stereotype Classifiers consistently achieved higher precision scores than recall scores, with precision scores ranging from 0.845 to 0.939. These scores are quite high, indicating that when the Classifiers annotate a description with *Omission* or *Stereotype*, the annotation is 84.5% to 93.9% likely to be correct. Using an Omission and Stereotype Classifier to identify potentially gender biased descriptions in the HC Archives' catalog could help the HC Archives team determine which collections' documentation to prioritize for manual review. The lower recall scores for the Omission and Stereotype Classifiers, which ranged from 0.539 to 0.771, indicate that only

53.9% to 77.1% of all manually-annotated gender biased descriptions were identified, reinforcing the need for the HC Archives employees' manual review of collections documentation. That being said, even if the Classifiers' recall scores were higher, manual review of collections documentation would still be recommended, because *Omission* and *Stereotype* annotations in the aggregated dataset do not account for all varieties of biased language, let alone gender biased language.

For the GLAM sector more broadly, the results of the Omission and Stereotype Classifiers indicate that text classification models can be created to support large-scale analysis of biased language in GLAM documentation, providing GLAM visitors with a sense of the skews in a GLAM catalog and providing guidance to GLAM employees conducting collection reviews. That being said, future work should run the classification cascades on catalogs from other GLAM institutions to evaluate the generalizability of the cascades and their conceptualization of gender biased language. Due to variations in description practices across the GLAM sector, the Omission and Stereotype Classifiers and the Person Name and Occupation Classifiers may not perform as well on Gallery, Museum, and Library documentation, which typically contain shorter descriptions than Archives' documentation. These classifiers are likely to work better on documentation that has lengthier descriptions, comparable to those of the HC Archives' catalog, because they consider sequences of text (either a sentence or description) when annotating with the Taxonomy's labels. By contrast, the Linguistic Classifiers are likely to perform well on a greater variety of GLAM documentation, because they consider one token at a time when annotating with the Taxonomy's labels. Future applications of this chapter's classifiers to other GLAM documentation should be accompanied by qualitative work with members of the GLAM institution who are experts on its collections' documentation, ensuring that differences in the documentation data and in conceptualizations of gender biases are taken into consideration when evaluating the models (e.g. the qualitative evaluation in Chapter 7).

Future work building on the classification model setups of this chapter could try several different approaches to preprocessing, word representation, feature engineering, and model evaluation. For preprocessing, punctuation could be removed, stop words could be removed, and tokens could be stemmed or lemmatized. Punctuation and stop words may be useful if patterns in sentence

structure help the models make classifications, but they also could have added noise to the training data. Including stems or lemmas as features for tokens could help the models identify similar words, improving classification results if commonly annotated words have the same root. Tokens could be represented using word embeddings of dimensions greater than 100, or word embeddings created using a different algorithm, such as skipgram.

Regarding feature engineering, future work could experiment with new or higher-quality features not included in this chapter's work. Features could be created from a list of non-unique labels, rather than unique labels, to emphasize the repeated presence of a label on a token or description. Tokens' part of speech tags or additional metadata fields from the HC Archives' catalog (e.g. date of material, language of material) could be included as features to see if they improve the performance of a model's ability to classify gender biased language. To estimate the improvements to the Omission and Stereotype Classifiers that could be achieved with higher quality features, the manual annotators' ***Linguistic***, ***Person Name***, and *Occupation* labels could be included as features instead of the models' annotations with these labels.

Regarding model evaluation, future work could build upon this chapter by varying model setups or the data the models classify. This chapter's model evaluation has focused on optimizing for F_1 score, considering precision and recall to be equally important when aiming to identify potential gender biases in language. That being said, should future work aim to optimize for precision, this chapter's experiments indicate that model setups different to those of the cascades would be best (§6.5). For example, the algorithm experiments for classifying *Omissions* and *Stereotypes* showed that Logistic Regression yielded the highest precision score, though SVM yielded the highest recall and F_1 scores. Additionally, the classifiers reported in this chapter could be further evaluated in their application to another archival catalog or to a catalog from another GLAM institution. Due to the contextual nature of gender biased language, I recommend the models first be evaluated on another catalog from a UK GLAM institution, maintaining the cultural context in which the models were trained. To complement the typical quantitative evaluation of models I reported in this chapter, I also recommend evaluating models qualitatively, with their intended audience. Chapter 7 demonstrates one approach to such qualitative model evaluation through a workshop with the HC team.

6.9 Recalibrations with the Classifiers

In this chapter, I continued the recalibration of ML demonstrated throughout this thesis with the creation of NLP models. Building on the work of those in the ML and NLP communities who take a critical approach to dataset and model creation (Aragon et al., 2022; Blodgett et al., 2020; Crawford and Paglen, 2019; D'Ignazio and Klein, 2020; Jo and Gebru, 2020; Rogers, 2021), I prioritized:

- **Quality over quantity** by using only data from the HC Archives' catalog that was manually annotated by members of the stakeholder groups of my research context (§4.1.6.1). Research from Kreutzer et al. (2022), Birhane and Prabhu (2021), and Crawford and Paglen (2019), among others, have found harmful biases in existing datasets created for ML models that result from a greater focus on creating large datasets than on creating high quality datasets. The presence of harmful biases in data is particularly problematic for my research context, training ML models to detect types of gender biased language.

- **Accuracy over efficiency** by conducting multiple types of evaluation on the classification models I report, recognizing that standard benchmarks and metrics to evaluate models for biases have yet to be well established (Welty et al., 2019). In this chapter I evaluate models through manual reviews of model predictions ("error analysis") in addition to the conventional quantitative metrics for evaluating model performance (§6.3.2). Chapter 7 provides a new, qualitative approach to evaluating models, introducing participatory classification evaluation and demonstrating its execution in a workshop with GLAM domain experts from the HC team.

- **Representativeness over convenience** by training custom word embeddings and by choosing not to fine-tune deep learning models. This ensures that any biases my models detect originate in the metadata descriptions of the HC Archives' catalog. Acknowledging the impossibility of disentangling biases in a pre-trained, deep learning model from biases in the HC Archives' documentation, I implemented a traditional ML approach, avoiding the risk of bias injections from such

models and their training data (Goldfarb-Tarrant et al., 2021; Steed et al., 2022).

- **Situated thinking over universal thinking** by positioning my text classification models in a case study (working with text data in British English from a Scottish archival catalog in the UK, focusing on gender bias from a 21st century perspective, and writing as an American), acknowledging that internal and external relations of any text corpus influence the meaning of text and thus the performance of models on that text (Beelen et al., 2021; Malik et al., 2022; van den Berg and Markert, 2020).

Chapter 7

Participatory Evaluation

> The true focus of revolutionary change is never merely the oppressive situations we seek to escape, but that piece of the oppressor which is planted deep within each of us.
>
> –Audre Lorde, *Sister Outsider* (1984, p. 123)

In this chapter I report on my final Participatory Action Research (PAR) activity: a workshop with the Heritage Collections (HC) team to evaluate manual and model-made applications of the Taxonomy of Gendered and Gender Biased Language (Chapter 5) to HC Archives documentation. With this workshop I complete the end-to-end integration of PAR with my Machine Learning (ML) model creation and evaluation processes. The PAR of this chapter complements the quantitative evaluation of models in Chapter 6, further investigating the research question: ***Can gender biased language be reliably annotated by domain experts to train a classification model to automatically annotate gender biased language?*** I created models to support stakeholders in the GLAM sector, so feedback from these stakeholders is necessary to evaluate how well the models live up to this intention. PAR activities are not yet common in ML model creation and evaluation (notable exceptions being Rodolfa et al., 2020 and Nobata et al., 2016), so in addition to contributing a qualitative evaluation of my annotated data and models, this chapter illustrates how stakeholder

collaboration can be incorporated into ML system creation, guiding researchers and practitioners in implementing similar evaluation approaches in the future.

7.1 Introduction

I ran a workshop with HC team members to seek their feedback on:

- **The Taxonomy's application (§7.4.1):** Although the HC team had provided me with feedback on the Taxonomy prior to the manual annotation process (Chapter 5), I was interested in their feedback on the Taxonomy as it was *applied* to HC Archives documentation during that process. Manual annotators' application of the Taxonomy represented the patterns the classifiers were trained to identify in Chapter 6.

- **The cross-collection measurements (§7.4.2):** Complementing the close-level reading of the Taxonomy as applied to individual descriptions, I was interested in the extent to which quantitative measures of gendered and gender biased language across the descriptions of a collection, and across multiple collections, would be useful to the HC team. The classification models scale up the annotation process, automatically annotating HC Archives documentation much faster than humans can manually review the documentation. This enables high level, cross-collection measurements, providing the HC team with a new view onto HC collections that covers hundreds of collections at one time.

To obtain this feedback, I defined two activities for the workshop: *Activity 1* focused on obtaining feedback on the Taxonomy's application and *Activity 2* focused on obtaining feedback on the cross-collection measurements.

In the following sections, I describe the workshop method (§7.2) and participants (§7.3); outline the workshop setup and procedure (§7.4); report observations (§7.5); discuss the observations relative to theories of feminism, critical discourse analysis, and heritage as a process (§7.6); and conclude with a summary of the workshop, and its implications for participatory approaches to ML and for research on bias in GLAM documentation (§7.7).

7.2 Method

The School of Informatics at the University of Edinburgh provided ethical approval (reference 2019/81479) for the workshop reported in this chapter. Aligning with the PAR approach of my Bias-Aware Methodology (Chapter 4), the workshop both served as a research method, to gather qualitative data, and aimed to facilitate change, specifically change in cultural heritage documentation practices and in ML model creation processes (Martin and Hanington, 2012b; Ørngreen and Levinson, 2017). Working with Rachel Hosker (University Archivist and Research Collections Manager) as a collaborator rather than a research subject, I collaborated with her to create an agenda, schedule a suitable time for members of the HC team to participate in the workshop, and invite members of the HC team to participate. Hosker stated her preference for taking responsibility for scheduling, reserving a room for, and inviting participants to the workshop. Recognizing that I was asking the HC team to dedicate a portion of their working day to participate in the workshop, I did not object. After the workshop, I circulated a questionnaire (Martin and Hanington, 2012c) to participants to gather information about their work and experience in GLAM, on the HC team, and with ML technologies. The questionnaire consisted of 18 questions, either multiple choice or short response, and took circa 10 minutes to complete. I used the questionnaire responses to characterize the workshop's participant group in §7.3.

The workshop was conducted as an open format workshop, similar to a semi-structured interview, with Hosker and I as joint facilitators influencing but not restricting the discussion (Ritchie and Lewis, 2003; Storvang et al., 2018). In the agenda, Hosker and I wrote pre-defined questions that were posed at the start of each workshop activity to prompt participant responses; then, we let participants' comments and questions guide the direction of the discussion for the remainder of the activity. To facilitate the discussion during each workshop activity, I created worksheets (detailed in §7.4) illustrating the potential capabilities of a digital interface to the text classification models I was developing at the time of the workshop (the complete versions of which I have reported in Chapter 6). The worksheets serve as technology probes, providing a tool with which I could research the HC team's needs in their work context, the quality of annotations of HC Archives documentation, and

potential applications of my text classification models to the HC team's existing workflows (H. Hutchinson et al., 2003). In addition to the participants' discussion throughout the workshop, their notes on the worksheets informed my analysis.

I analyzed the workshop discussion with a grounded theory approach, performing content analysis to inductively surface themes through open coding (Glaser and Strauss, 1980; Krippendorff, 2018; Robson and McCartan, 2016). I report these themes as "observations" (§7.5). My theoretical triangulation of feminist theories, critical discourse analysis, and heritage as a process (§3.3) guided my analysis (§7.5) and interpretation (§7.6).

7.3 Participants

Through conducting this PAR workshop, I engaged with the same stakeholder group, that of the HC Archives' employees, that I engaged with for my dataset curation and Taxonomy creation processes (Chapter 5). Five out of the 10 participants who attended this workshop had also attended the Taxonomy review workshop. Hosker emailed a workshop invitation to 14 members of the HC team; nine people accepted the invitation, leading to a total of 10 workshop participants including Hosker. Hosker was the sole participant holding a management position in HC.

The emailed workshop invitation included an agenda (Appendix H.3), participant information sheet (Appendix H.1), and participant consent form (Appendix H.2), enabling invitees to review expectations for workshop participation prior to deciding whether to attend the workshop. At the workshop, I brought printed participant information sheets and consent forms for the invitees who chose to attend. All 10 participants signed the consent form acknowledging their voluntary, anonymous participation and their right to withdraw from the workshop at any time. One participant only attended until the end of the workshop's first activity; the remainder of the participants attended for the entirety of the workshop.

Prior to the workshop, all but one participant had heard of ML, one participant had interacted with ML models, and one participant had created an ML model. Regarding encounters with biases in cultural heritage collections and their descriptions, participants reported encounters with biases related to

gender, sexuality, faith, race and ethnicity, nationality, language and accent, and disability. Participants described their gender as "female," "cisgender female," and "female-ish."[1] Their job titles can be broadly described as:

- Archivist (six participants),
- Curator (three participants),
- Librarian (one participant), and
- Manager (one participant).[2]

The higher education degrees participants held included degrees in subjects of Archival Science, History, Museum Studies or Museology, and Musicology. Participants described their job responsibilities as including acquiring, cataloging, curating, processing, developing, managing, preserving, interpreting, and facilitating discovery and access of collections; teaching with and giving training about collections; overseeing colleagues and volunteers; responding to enquiries about the collections; communicating with donors and stakeholders of collections; participating in international communities of practice; and research. Participants' experience working in the GLAM sector ranged from six years to over 10 years, and their experience working as at the University of Edinburgh ranged from less than one year to over 10 years.

7.4 Setup and Procedure

I structured the workshop around two activities, one activity for each type of feedback I sought from the HC team for a qualitative model evaluation. Prior to running the workshop, I met with Hosker to discuss an agenda (Table 7.1) and participants to invite to attend the workshop. I finalized the agenda and sent it to Hosker for feedback; Hosker approved the agenda and then emailed an invitation to attend the workshop to 14 members of her team. 10 participants attended the workshop in person (no participants attended online). With participants' consent, I audio-recorded the workshop using Zoom. Appendix H.4 provides a transcription of the workshop. All participants were sent the transcript and this chapter to review.

[1] Although the gender representation at the workshop was not particularly diverse, given the minoritized status of the represented genders, the workshop still contributes to the advancement of gender equity (also discussed in §5.1.2).

[2] Certain participants hold positions that fall into more than one of the listed job titles.

2:00-2:15	**Welcome:** Participant information and consent Introduction to research, workshop aims Questions from participants
2:15-3:00	**Activity 1: Taxonomy's Application** Review and discuss example outputs from a machine learning model that flag potentially gender biased language in metadata descriptions from the Archives' catalog.
3:00-3:10	**Break**
3:10-3:55	**Activity 2: Cross-Collection Measurements** Review and discuss example summary information about a machine learning model's findings across a subset of the Archives' catalog.
3:55-4:00	**Wrap up:** Questions from participants Final thoughts Thank you

Table 7.1: **Workshop Agenda, April 20, 2023.** The agenda for the workshop, which was held in the University of Edinburgh Library's Digital Scholarship Centre. The agenda displays times in the left column and tasks in the right column. During the workshop I gave participants a more detailed version of the agenda that with the activities' guiding questions (Appendix H.3).

To provide adequate context for the workshop participants to participate in these activities, and to follow the recommended ethical procedures of the University of Edinburgh's School of Informatics, at the start of the workshop I asked each participant to read and sign the participant information sheet and consent form, gave each participant an agenda, and then explained my research process and workshop aim. A pilot workshop I ran in preparation for this workshop, with three Ph.D. students and one post-doctoral researcher working in the Digital Humanities at the University of Edinburgh, had confirmed that the explanation I gave of my research process and workshop aim was understandable to people unfamiliar with my Ph.D. research and to people without ML experience. I described the supervised learning approach to creating text classification models, explaining how the models were trained on manually annotated data with the goal of automating the annotation process, providing a more efficient, scalable approach to finding potentially gender

biased language than hiring human annotators.

Next, I asked participants if they had questions. Participant 3 responded with a question about the data transformation process, asking whether a description could be traced back to its collection. I confirmed that yes, descriptions can be traced back to the collections in which they appear because I maintained the association of each description with its corresponding collection's identifier (the Encoded Archival Description Identifier (EADID)). Receiving no further questions, Hosker and I proceeded to *Activity 1*.

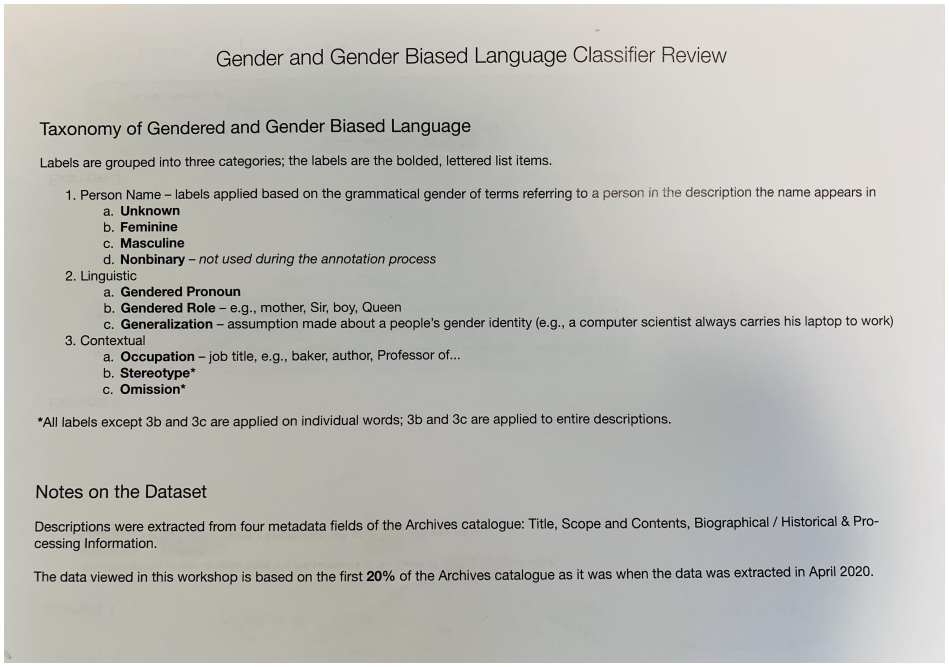
7.4.1 Activity 1: The Taxonomy's Application

Activity 1 focused on the Taxonomy's application, investigating participants' attitudes toward annotations of HC Archives documentation with labels from the Taxonomy of Gendered and Gender Biased Language. To guide participants' reflection and discussion, I created a worksheet for this activity with an outline of the Taxonomy (Figure 7.1) and three examples of annotated descriptions from the aggregated dataset created in Chapter 5 (Figure 7.2). The example descriptions together represent all of the Taxonomy's labels and illustrate the variety of language that could be annotated with a single label. Please refer to figures 6.2, 6.3, and 6.4 for the model cascades' labels on these descriptions.

I began *Activity 1* by talking through the Taxonomy, explaining the rationale behind each label. I then asked participants to take five minutes to review the example descriptions independently and note responses to two questions:

1. Do you agree or disagree with, or are you unsure about, the labels on the descriptions? Why?
2. How would you use this information?
3. What information is missing that you would need to support your tasks?

Though I set a timer for ten minutes, participants began to ask questions prior to the timer ending that led into a group discussion. In an effort to avoid enforcing a rigidity that would feel uncomfortable to participants, I chose not to interrupt the discussion to return participants to individual reflection. I report observations in response to the three questions of this workshop activity in §7.5. After one hour, I interrupted participants to thank them for their willingness to share their thoughts so openly and take a 10-minute break.



Gender and Gender Biased Language Classifier Review

Taxonomy of Gendered and Gender Biased Language

Labels are grouped into three categories; the labels are the bolded, lettered list items.

1. Person Name – labels applied based on the grammatical gender of terms referring to a person in the description the name appears in
   a. **Unknown**
   b. **Feminine**
   c. **Masculine**
   d. **Nonbinary** – *not used during the annotation process*
2. Linguistic
   a. **Gendered Pronoun**
   b. **Gendered Role** – e.g., mother, Sir, boy, Queen
   c. **Generalization** – assumption made about a people's gender identity (e.g., a computer scientist always carries his laptop to work)
3. Contextual
   a. **Occupation** – job title, e.g., baker, author, Professor of...
   b. **Stereotype***
   c. **Omission***

*All labels except 3b and 3c are applied on individual words; 3b and 3c are applied to entire descriptions.

Notes on the Dataset

Descriptions were extracted from four metadata fields of the Archives catalogue: Title, Scope and Contents, Biographical / Historical & Processing Information.

The data viewed in this workshop is based on the first **20%** of the Archives catalogue as it was when the data was extracted in April 2020.

Figure 7.1: **A workshop participant's *Activity 1* worksheet (front).** An overview of the Taxonomy of Gendered and Gender Biased Language.

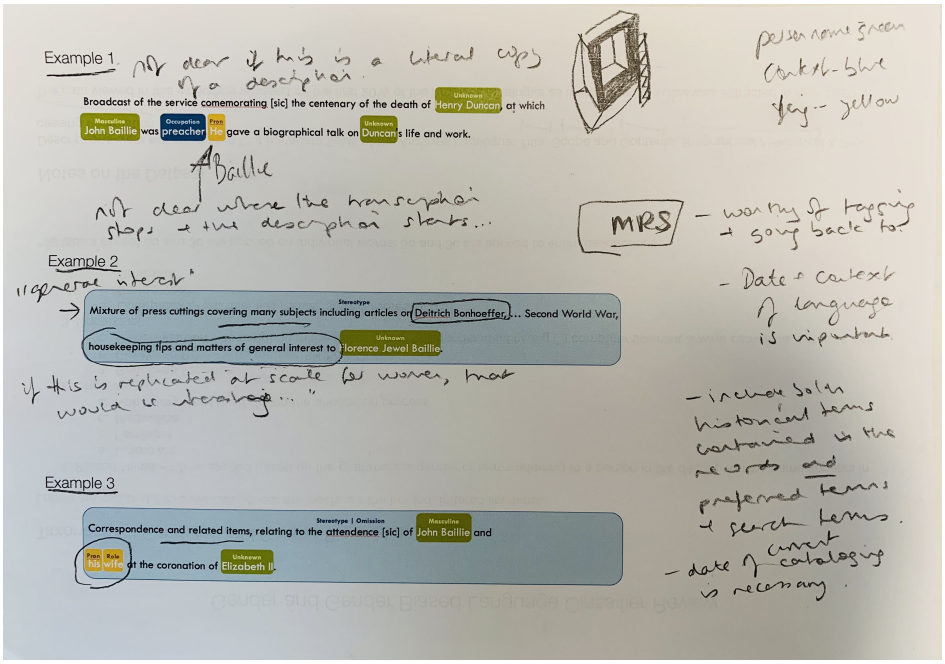


Figure 7.2: **A workshop participant's *Activity 1* worksheet (back).** Annotated description examples from from manually-annotated HC Archives documentation that served as model training, validation, and test data. The annotations are color-coded by Taxonomy category: yellow for ***Linguistic*** labels, green for ***Person Name*** labels, and blue for ***Contextual*** labels. Label names appear at the top of each annotation color block.

7.4.2 Activity 2: The Cross-Collection Measurements

After the 10 minute break following *Activity 1*, Hosker and I began the workshop's second activity. *Activity 2* focused on the cross-collection measurements, gathering participants' feedback on my approach to annotating HC Archives documentation at a higher level. Thus I created a worksheet for this activity that presented tables and charts summarizing measurements of gendered and gender biased language calculated using my text classification models' predictions. The front of the worksheet (Figure 7.3) displays tables that list the total counts of *Stereotype* and *Omission* labels for collections that received the highest number of annotations with these labels. The worksheet also displays tables that list the most common languages of the archival material in those collections. The back of the worksheet (Figure 7.4) displays three bar charts grouped into two sections. The top section visualizes the total number of descriptions that an early iteration of an Omission and Stereotype Classifier annotated with either the *Omission* label, the *Stereotype* label, or both labels. The bottom section visualizes calculations made with the predictions of an early iteration of a Linguistic Classifier. The top bar chart in this bottom section visualizes the total number of words classified with a ***Linguistic*** label (*Gendered Pronoun*, *Gendered Role*, *Generalization*). The bottom bar chart visualizes the Linguistic Classifier's performance, displaying the percentage of annotations that were true positives, false positives, true negatives, and false negatives.

I posed the following questions to participants to guide their review of this second worksheet:

1. What do you understand from the information on the worksheet? What questions do you have about the information you're seeing?
2. How would you use this information?
3. What information is missing that you would need to support your tasks?

As with *Activity 1*, participants were asked to spend five minutes reflecting individually before shifting to group discussion, but the group discussion recommenced before five minutes had passed. I again chose not to interrupt the discussion. Participants' discussion during *Activity 2* lasted 40 minutes.

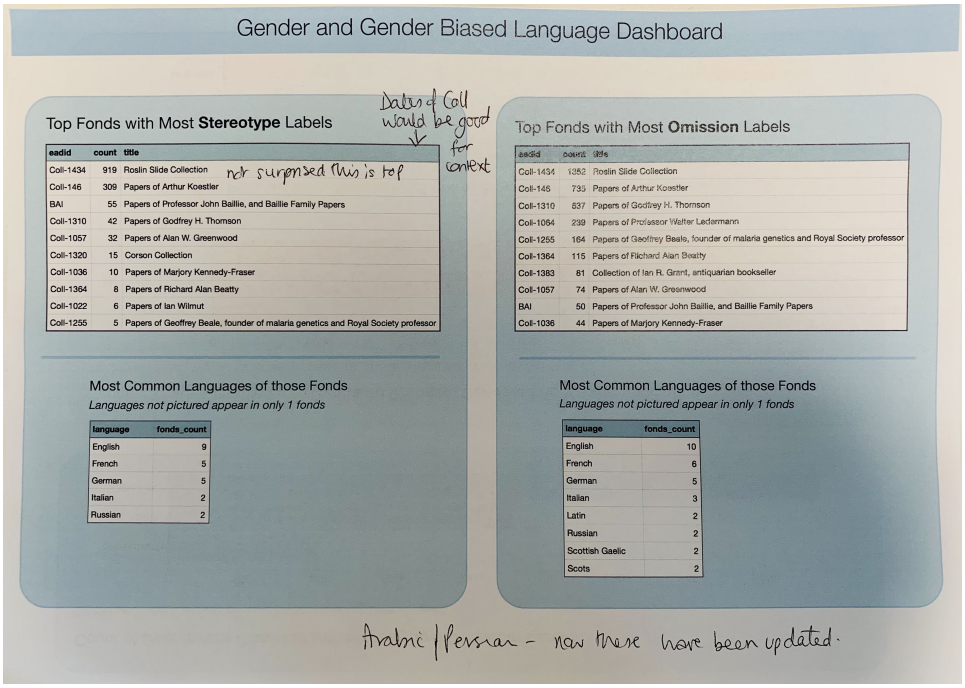


Figure 7.3: **A workshop participant's *Activity 2* worksheet (front).** Tables with summary information about classifier annotations with the *Omission* and *Stereotype* labels. Note: "fonds" is the archival term for collection.

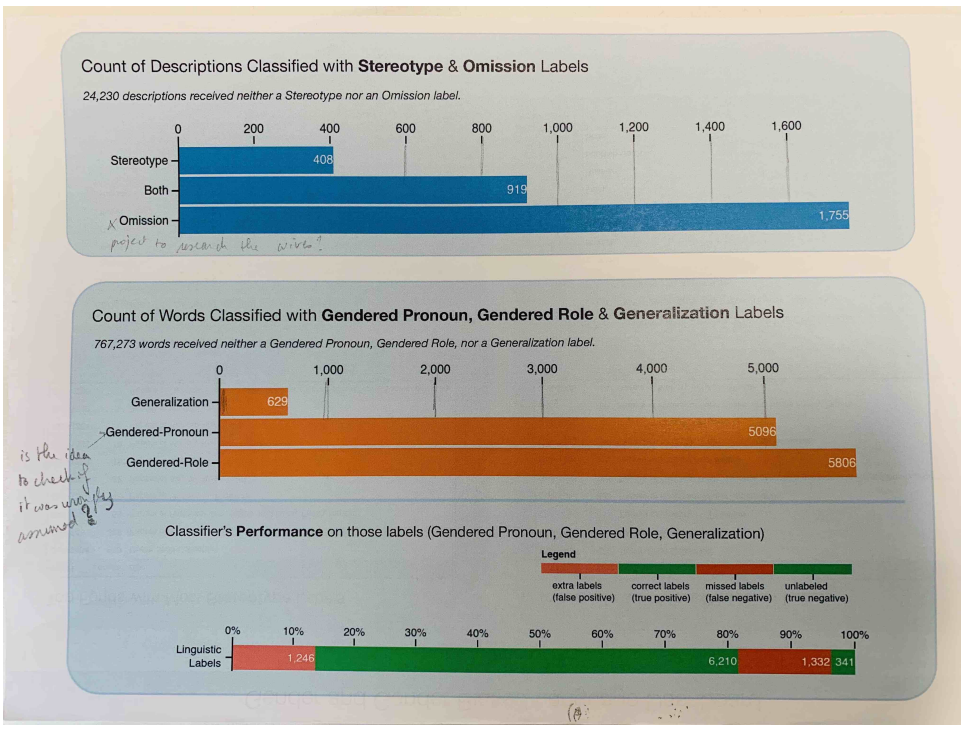


Figure 7.4: **A workshop participant's *Activity 2* worksheet (back).** Bar charts displaying summary statistics I calculated from classifier annotations.

7.4.3 Workshop Wrap Up

With five minutes remaining in the workshop, I interrupted participants' discussion to pose one more question, asking:

1. Is there any information on the worksheets that you would want to share with visitors to the collections, or if not the precise information on the worksheets, variations of it?

After hearing responses from two participants (described in §7.5.7), I ended the workshop, thanking participants for their attendance and contributions.

7.5 Results

In this section I detail my observations (**O1**-**O17**) of participants' discussion during the workshop, organizing the observations by activity (A1 or A2) and question (Q1, Q2, or Q3). To maintain participants' anonymity, I refer to them as P1, P2, P3, etc. As described in §7.2, I took a grounded theory approach to analyze the workshop results, performing content analysis to inductively surface my observations through an open coding of the discussion (Glaser and Strauss, 1980; Krippendorff, 2018; Robson and McCartan, 2016). For clarity, I made minor revisions to the quotes I include from participants.

7.5.1 A1, Q1: Agreement and Disagreement with Labels

This section summarizes observations relevant to the first question of *Activity 1* investigating attitudes toward the Taxonomy's application: *Do you agree or disagree with, or are you unsure about, the labels of the descriptions? Why?* Participants agreed and disagreed with the application of gendered language labels to varying degrees. Regarding the application of the gender biased language labels, however, participants expressed agreement and asked clarifying questions, but did not voice disagreement.

O1: Value of gendered language labels individually vs. in combination. Participants expressed agreement regarding the application of the *Gendered Role* label. P8 communicated an interest in seeing the quantity of different text spans that the *Gendered Role* labels annotated, saying that those quantities

could support the search for *"bias and gaps"* in collections' documentation. P8 explained that this label would be especially interesting for identifying *"where a 'Mrs.' is, and saying that's potentially a partner that needs further exploration,"* referring to the gaps in historical documentation about women's work and contributions (Beard, 2017; Graeber and Wengrow, 2021; Hessel, 2023b; Hessel and Beard, 2022).

Participants expressed disagreement regarding the application of the *Gendered Pronoun* and *Unknown* labels. Regarding the application of the *Gendered Pronoun* label, P2 and P5 stated that they would be interested in seeing gendered pronouns annotated if there was an assumption about a gender group being made (i.e. as with text annotated as *Generalization*), but not as it had been applied to all gendered pronouns in the archival documentation. Regarding the application of the *Unknown* label, P7 expressed disagreement with the way in which the label was applied based on gender information in a single description, rather than the gender information within an entire collection's descriptions. P10 communicated similar disagreement with the application of the *Unknown* label, pointing to an example where "Elizabeth II" was annotated as *Unknown*: *"in this case it's a woman in power, but then we're not recognizing that there were some women in power, if everybody's not identified by gender."* In this way participants communicated the value of gendered language for making historically minoritized gender groups more visible in GLAM documentation.

Conversely, P1 expressed agreement with the labels when considered in combination. P1 stated, *"It's not about just the* Gendered Pronoun*, it's actually when they come together, they produce something that's really useful."* P1 expressed interest in approaching annotations with the Taxonomy's labels in combination with one another, rather than considering each label individually.

O2: Concern with losing gender information. The conversation at the beginning of *Activity 1* communicated participants' concern with the Taxonomy's labels being an indication of language that should be removed or changed. P6 explained how, due to the minoritization of women, often a woman is only identified with a title and last name (e.g. "Mrs. MacDonald"), so *"if you lose 'Mrs.' then you lose all of their identity rather than enhancing their identity."* P8 gave a specific example of how the identity of a wife,

who was depicted in a painting with her husband, was able to be discovered thanks to the more extensive documentation of her husband. In this case, it was the wife who brought significant funds to the marriage that enabled her and her husband to develop their substantial art collection. The inclusion of the gendered title "Mrs." thus enables the discovery of further information about women, countering the lack of documentation about their roles and contributions throughout history.

Participants also discussed the value of identifying people by gender even if that gender is assumed. Regarding the application of the *Unknown* label, P7 expressed discomfort: *"I feel kind of weird, though, with it, in that now I can't say what their genders are [...] how can we say who they are if we can't assume that?"* Participants explained how describing people with gendered terminology can be useful even if an assumption was made about a person's gender, because that assumption reflects how the person was perceived and moved in society. The challenge is distinguishing between where a cataloger made an assumption about a person's gender and where a cataloger learned a person's gender from the way the person self-identified in collection material. P7 explained that for this reason, using quotes when describing collections is important to help future readers distinguish between information directly from collection material and information from a cataloger's interpretation of the material.

In response to participants' concern about losing information with the removal of gendered terminology, I clarified that the Taxonomy's labels were not intended to highlight language to be removed or changed, but rather were intended to make potential gender biases more easily identifiable.

O3: Bias in the Taxonomy's application. Regarding the application of the *Occupation* label, participants suggested that the label could have been applied more broadly, rather than only to job titles. Looking at examples 2 and 3 in Figure 7.2, P10 asked why "housekeeping" did not have an *Occupation* label, and said, *"And it's the same thing, you know, with 'coronation,' hinting at her [Elizabeth II's] occupation."* My instructions (Appendix D) for applying the Taxonomy's *Occupation* label reinforced occupational gender biases rather than pushing back against them.

O4: Subjectivity of gender bias. Differences in the interpretations of the

gender biased language labels, as well as an awareness of the subjectivity of bias and language, came through in participants' comments and questions. Regarding the *Occupation* label, P1 pointed out the subjectivity of the language with the question, *"But actually, is a preacher an occupation or a calling or what?"* That being said, the *Stereotype* label yielded the greatest debate.

Participants asked several questions about how I defined *Stereotype* and examples of text that had been annotated with this label. P5 spoke of the contextual nature of a stereotype, saying, *"it's changed by people, by culture, by things like this, so it seems really hard to define what it is."* P7 expressed concern about judgments implied with the annotation of a description as *Stereotype* (Figure 7.2, Example 2), saying, *"You could be doing her a huge injustice if her life's mission was housekeeping."*

This second example of an annotated description led to lengthy discussions about stereotypes. P8 asked for clarification about whether the *Stereotype* label had in fact been applied due to the association of a woman with "housekeeping." P3 responded by saying, *"I thought it wasn't just the housekeeping but that a woman's interests just get summarized into 'general interests.'"* Later in the workshop, P3 went on to say, *"It doesn't say, 'Florence Jewel Baillie was an avid researcher of Dietrich Bonhoeffer, the Second World War, housekeeping, etc.' It just says here are some news cuttings on some things that she read about. Which kind of makes a judgment. That's what I thought the* Stereotype *was in this instance."* Though I saw both interpretations as motivation for the *Stereotype* annotation, I chose not to respond to P8's question myself. I was interested in participants' interpretation and did not want to put mine forward as the "correct" or "intended" interpretation.

O5: Discomfort with bias evaluations. At the beginning of the workshop I stated that my goal with the annotations I presented was to highlight how historical minoritization and systemic biases come through in language, and that my goal was not to blame catalogers for their language choices. Nonetheless, participants voiced discomfort with the process of reflecting upon and discussing gender biases in the archival documentation. In the post-workshop survey, one participant commented, *"I found [the workshop] intimidating because I felt like my work, and thus myself, was being judged (even though the examples chosen were not my work, but it felt inappropriate to display*

this fact)." During the discussion of Example 2's (Figure 7.2) *Stereotype* label, P3 observed, *"It's very self-revealing to say what you think."* Discomfort thus seemed to stem from a sense of responsibility and fear of judgment.

As a workshop facilitator, Hosker responded to these comments with reassurance, encouraging people to voice conflicting opinions and interpretations. For example, during *Activity 1*, Hosker stated, *"I love being in this space where we can sort of unpick things even if we disagree or we have different ideas, it's really, really healthy."* As a workshop facilitator, I also aimed to reassure participants, explaining that I was focused on understanding the complexities surrounding biased language; I was not interested in unrealistically simple interpretations of biased language that identified a single source of bias on which to assign blame.

7.5.2 A1, Q2: Anticipated Next Steps

This section summarizes observations relevant to *Activity 1*'s second question investigating attitudes toward the Taxonomy's application: *How would you use this information?*, in reference to the annotated descriptions on the back of worksheet 1 (Figure 7.2). Participants saw model annotations of archival documentation as flags that could support and inform their next steps of collecting, describing, and managing collections through manual processes. That being said, the models could not address all the challenges related to bias in GLAM documentation due to the complicated networks of stakeholders that influence documentation practices.

O6: Models as a tool to support existing workflows. Participants explained that upon being presented with the models' annotations, the next step would be to have a member of the HC team revisit the annotated descriptions. P5 saw the models as, *"a way to flag something that might be problematic. But then you've got to have an archivist or cataloger go back,"* rather than leaving a model to determine whether or not changes should be made to the documentation. Participants thus saw the models as tools to support existing practices.

Participants also saw the models' annotations as useful for informing guidelines to documenting collections. P2 proposed that the models' annotations could help the HC team write a *"how to"* guide for describing

collection material. P8 responded, *"I think that would be quite useful to have,"* in reference to the participant's current work with volunteer catalogers, to raise volunteers' awareness about gender biases in GLAM documentation.

O7: Models as a tool for self-reflection. Participants discussed the value of the models' for self-reflection on descriptive practices. Their comments and questions communicated a recognition of the inevitability of bias. Participants spoke of how concepts of bias evolve, and how different stakeholders of a collection may have different opinions about which terminology is correct and biased. Rather than trying to fix or remove bias, P1 proposed that the aim could be awareness, and that the models could help the HC team in *"being aware of our practice, or aware of where we've taken old catalogs and perpetuated things."* Participants saw the models as tools to help them view collections and their documentation from a different perspective.

O8: Model outputs as evidence to support resource requests. P1 spoke of how the models could provide evidence for why project funding and new hires were needed to work on particular documentation efforts. P1 discussed the models' annotations as *"an evidence base"* to support requests for interns and trained professionals. This participant saw the models' annotations as a way to demonstrate that *"there is a skill to this [GLAM collection description] and these skills of archivists, librarians, and curators are needed in this space."* Student interns are not suitable for all documentation work, so participants value evidence that will demonstrate the need for their professional skills.

O9: Barriers posed by institutional and professional structures. While acknowledging the power held as the people describing collections, participants also discussed limitations to their power with regard to institutional priorities, as well as professional standards and expectations. Regarding the limited power one can exert within an institution to request necessary resources, P1 stated, *"We will always get back to having an army of catalogers that we need."* Regarding the limited power one can exert over GLAM cataloging standards and expectations for interoperable GLAM systems, participants discussed the tension between standards and the reality of heritage material. P3 pointed out that the biases sit not only within cataloging standards, but also *"the whole*

profession." Decisions about cataloging standards are made by professional boards that often lack adequate representation of minoritized communities. As noted in Chapter 4, the HC team described themselves as activists in the area of cataloging. P1 pushed back against the GLAM profession's emphasis on knowing large quantities of cataloging standards and adhering to them unquestioningly. In reference to GLAM practices, P1 stated, *"We were taught to accept them and that's, well, I don't want to. I want to see them as a framework."* Similar to my view of ML systems, participants view GLAM documentation as having the potential to reinforce or subvert societal power relationships.

7.5.3 A1, Q3: Critiques of the Information Presentation

This section summarizes observations relevant to *Activity 1*'s third question investigating attitudes toward the Taxonomy's application: *What information is missing that you would need to support your tasks?* To present example annotated descriptions, I used the color encodings from the manual annotation process: yellow for ***Linguistic*** labels, green for ***Person Name*** labels, and blue for ***Contextual*** labels.

O10: Purpose of color encodings. When reviewing a catalog's annotated descriptions individually, distinguishing specific categories of labels through color coding may be unnecessary. After explaining how the descriptions' labels were color-coded according to the Taxonomy, and asking participants to write down individually what they agreed and disagreed with on worksheet 1, P7 asked, *"Do the colors mean something?"* Color-coding the labels to distinguish annotations for gendered and gender biased language may provide a more informative distinction for a close reading activity such as *Activity 1*.

7.5.4 A2, Q1: Interpretation of Summary Information

This section summarizes observations relevant to the first question of *Activity 2* investigating attitudes toward the cross-collection measurements: *What do you understand from the information on the worksheet? What questions do you have about the information you're seeing?* In response to worksheet 2's tables (Figure 7.3), participants communicated that they were not surprised to see certain collections appear in the list of those with the highest gender biased

label counts, but wondered whether the models annotated those collections for the reasons they expected. Participants showed little interest in worksheet 2's quantitative measures of model performance (Figure 7.4).

O11: Uncertainty about models' capabilities. Upon reviewing the tables of collections with the highest counts of model-annotated descriptions with the *Omission* and *Stereotype* labels, P1 expressed interest in three of the collections that worksheet 2 listed. Those collections' presence in the tables aligned with P1's knowledge of the biases associated with the collections, which were related to eugenics, misogyny, imperialism, and gender stereotyping. Regarding the *Papers of Godfrey H. Thomson*,[3] P1 exclaimed, *"I'm absolutely fascinated that Godfrey Thomson has come up,"* explaining that high *Stereotype* counts could be due to *"the time period he worked in but also the environment. I think some of his work early on looked at eugenics and that kind of classification of people."* Regarding the *Papers of Arthur Koestler*,[4] P1 noted that the documentation of this collection came from a historical catalog, rather than being written by a member of the HC team. That being said, P1 expressed uncertainty about whether the models had labeled these collections for the reasons she expected, stating, *"we'd have to investigate further."*

P7's comments expressed the strongest doubt in the models' capabilities relative to those of other participants. In regards to the presence of the *Roslin Slide Collection*[5] in the tables recording collections with high *Omission* and *Stereotype* counts, P7 stated, *"I suspect there is some complicated problem with the data and the collection, and the way the data is being compiled, which is why it's got such high numbers there."* P1 responded that the presence of the *Roslin* collection in the tables did not surprise her, explaining, *"The images [in the collection] were collected during the height of empire and colonialism, and there is language within it that is gendered. There's also how women are represented and that can be in terms of different parts of the world and communities."* I seconded this explanation based on my own experience annotating the descriptions of the *Roslin* collection, which I had found offered many examples of the

[3]Documentation of the *Papers of Godfrey H. Thomson* is available at: archives.collections.ed.ac.uk/repositories/2/resources/85818.

[4]Documentation of the *Papers of Arthur Koeslter* is available at: archives.collections.ed.ac.uk/repositories/2/resources/85878.

[5]Documentation of the *Roslin Slide Collection* is available at: archives.collections.ed.ac.uk/repositories/2/resources/85706.

intersectional nature of bias along the axes of racialized ethnicities and gender.

O12: Little interest in quantitative model evaluation measures. Participants did not voice questions or comments about the *Activity 2* worksheet's bar charts. Handwritten notes on participants' worksheets (e.g. Figure 7.4) suggest that participants were more interested in gaining an understanding of the type of language that the models annotated with the Taxonomy's labels.

7.5.5 A2, Q2: Anticipated Next Steps

This section summarizes observations relevant to the second question of *Activity 2* investigating attitudes toward the cross-collection measurements: *How would you use this information?*, referring to the tables and bar charts (figures 7.3 and 7.4). Upon seeing the results from the models, participants would next like to gauge the extent to which the models' annotations align with their own interpretations of HC Archives documentation.

O13: Manual review of language that models flag. Participants expressed both interest and skepticism in the models' capabilities. P1 proposed, *"I wonder whether actually, if there were collections we would suggest putting through the models to test them, to see if they pick up on things that we know are in those collections, whether that would be something useful."* Other participants confirmed that they would be interested in reviewing model annotations on documentation of collections with which they were familiar to evaluate the models' performance.

Additionally, similar to **O6**, participants saw the models' ability to flag potentially problematic language useful as an added step for collection reviews that could increase the efficiency of that process overall. P9 said, *"it would maybe enable this kind of revisiting and correcting to be done more quickly and with less labor."* The models could provide a first pass over collections' documentation that HC team members could use as a guide for prioritizing specific collections' documentation to manually review, rather than manually reading through all existing documentation.

O14: Desire for more time to document collections. Similar to **O9**,

participants spoke of time constraints that limit the research they can conduct on collections to inform their documentation. P3 spoke about the implications of the increased scale of GLAM collections since the advent of digital technologies: *"For digital archives, at the scale some of them are, the human race will die out before we'd be able to catalog every website."* P8 spoke of the mismatch between the time needed and time available for describing a collection currently being cataloged: *"[...] then you have to research her relationship, his relationship to get a hint as to what Mary's role was as well, and that takes you into a three year Ph.D.! And you've got six months to catalog the collection!"* Participants communicated an unresolvable tension between making as much material as possible discoverable (i.e. documenting as much material as possible so visitors can search for it in online catalogs) and conducting lengthier research into collections that could challenge social biases.

O15: Sense of responsibility to past, present, and future stakeholders. Participants communicated a sense of responsibility to historical people and communities represented in collections, to present-day members of those communities, and to future visitors who will look for material about those people and communities. Reflecting on the role of GLAM catalogs and the process of documenting collections, P9 commented, *"I think it's sort of socially really important [...] because so many people are going to look it [the catalog] up, students or members of the public, and refer to that as a kind of non-biased sort of factual resource. So it's different from writing something where you assess your positionality and make that clear to the reader [...] it's a big responsibility to get right as well."* Participants expressed concern about navigating this responsibility when stakeholders may have conflicting points of view.

When attempting to meet their responsibility to stakeholders in the past, present and future, participants experienced time not only as a limitation (**O14**), but also as a complication. P6 gave an example in which a historical person in collection material referred to themselves as a *"Highland traveler,"* but a person in that present-day community had contacted the HC team to request that term, today considered derogatory, be changed to today's accepted term, *"Gypsy Roma Traveler."* P6 explained the tension between catalogers' responsibilities to different stakeholders by saying, *"So you then get into layers of, well who's say is it? Is it the current community, or is it the person who's own*

voice it is?" P4 gave an example related to gender, explaining that historical members of the trans community used terms to describe themselves that are considered offensive today. Nonetheless, P4 viewed this language as *"something that's an example of trans history or gender non-conformity"* and thus important to include in catalogs.

P7 posed a question that communicated the temporary nature of any solution to a diachronic change in language, asking, *"What do you do when [...] in 20 years time, something you created is deemed offensive?"* P2 added that diachronic changes in language are unpredictable, with certain terminology evolving from having positive to offensive connotations and other terminology evolving from having offensive to positive connotations (also noted in Schulz, 2000; Shopland, 2020). Participants came to a consensus that including as much terminology as possible, with contextualization, is the best approach, despite their dissatisfaction with the description becoming *"clunky-looking"* or *"difficult to read."* Their discussion indicated an acknowledgment of time as an ever-present, unpredictable challenge that means the work of addressing biased language is always ongoing.

7.5.6 A2, Q3: Critiques of the Information Presentation

This section summarizes observations relevant to the third question of *Activity 2* investigating attitudes toward the cross-collection measurements: *What information is missing that you would need to support your tasks?* Critiques of the information presentation during *Activity 2* focused on the tables on the front of worksheet 2 (Figure 7.3). As noted in **O12**, participants did not comment on or ask about the bar charts on the back of worksheet 2 (Figure 7.4), suggesting that the data being visualized was not of interest to participants.

O16: Value of dates to understand context. Participants expressed the importance of including dates wherever available, because dates situate a description's language in the historical context in which the language was produced. Regarding potential biases in the language of the "Title" metadata field, P3 said, *"we probably inherit that [...] so it'd be good to be able to isolate*

language that's been created recently by an archivist." P1 and P3 spoke of the value of dates alongside records of different versions of a catalog description. P1 explained, *"it's about dating descriptions and not overwriting them."*

7.5.7 Wrap Up, Q1: Sharing Information with Visitors

This section summarizes observations relevant to the question I posed to wrap up the workshop: *Is there any information on the worksheets that you would want to share with visitors to the collections, or if not the precise information on the worksheets, variations of it?* Most participants did not respond to the question of whether they would like to share the information of the worksheets, or variations of that information, with visitors.

O17: Distinct needs of GLAM employees and visitors. Regarding the possibility of sharing information on either worksheet with visitors to HC's catalog and collections, P1 suggested, *"I think it'd be really good to ask some of our User Services people who get inquiries in."* Another participant stated that *"It'd be useful to speak to users [...] because I've got a feeling they come with knowledge that we don't have,"* recalling earlier discussions of documentation as a work-in-progress intended to help researchers discover potential historical gaps they can investigate further. The results of such an investigation could be brought back to HC for *"enhancing or building on"* existing documentation.

During the pilot workshop, in response to worksheet 2's visualizations of cross-collection measurements, a participant stated, *"I just don't really think about the fact that they [GLAM catalog's descriptions] are created in a certain time period by certain people. So I think it adds a layer of something to think about when you're looking for materials [...] it's a reminder as a researcher that this is not some sort of neutral, objective tag of what you actually can find."* Speaking with the User Services team and researchers who use HC's catalogs and collections is an area for future work.

7.6 Discussion

The workshop with members of the HC team provided a qualitative evaluation of the the text classification models to complement the conventional,

quantitative evaluation of models reported in Chapter 6. The qualitative evaluation consisted of GLAM domain experts' feedback on manual annotators' application of the Taxonomy, which determined the language patterns guiding models' classification of text (*Activity 1*), and feedback on the models' predictions, which I presented in aggregate as cross-collection measurements (*Activity 2*). Feedback on the Taxonomy's application and utility of cross-collection measurements varied from one participant to the next. The evolving nature of language (**O14**, **O15**) and subjectivity of bias (**O3**, **O4**) seemed to make participants hesitate to express confidence in the annotations. Still, they identified four areas of value that the models held:

1. Supporting collection reviews with estimates of the quantities of potentially biased descriptions per collection, informing the prioritization of collections for review (**O1**, **O6**, **O13**)
2. Informing description-writing guidelines for students and volunteers who work with the HC team to document collections (**O6**)
3. Facilitating self-reflection in descriptive practices (**O5**, **O7**)
4. Providing evidence of resource needs for cataloging projects (**O8**, **O14**)

All four value areas involve human-ML collaboration, with the models serving as tools for the HC team to use to support and inform existing practices. Workshop participants spoke of the importance of context for deciding when to use the models. P1 likened the text classification models to Transkribus,[6] a handwriting recognition technology, stating that the question to consider is *"When you deploy this and when you don't, depending on the collection knowledge and cataloging required."* The HC team's attitude toward the models reject the assumption that computational technologies alone always offer the best solution or approach, a form of bias that Broussard terms "technochauvinism" (2019). The HC team members view my models as valuable for a "human-in-the-loop" system, where "a machine does a lot of the work but meaningful human work and meaningful human interaction are prioritized" (ibid.). The HC team's critical approach to my models is consistent with explorations of ML applications in the GLAM sector (Averkamp et al., 2021; Baker, 2020; Cordell, 2020; Jaillant, 2022; Padilla, 2019).

[6] readcoop.eu/transkribus

The lack of consensus among participants around agreement and disagreement with the Taxonomy's application reflects the subjective, contextual nature of language (**O3**, **O4**). Factors of time, culture, and politics, among others, influence what terminology will be deemed appropriate or inappropriate. GLAM collections do not easily fall into categories, despite categorization being the aim of GLAM catalogs' metadata standards and classification schemes. Reflecting on the process of cataloging certain collections according to certain metadata standards, P1 stated, *"They just didn't fit. It was, you know, trying to shoehorn things into a standard way of doing things [...] the messy archival world is not like that. Life does not conform at all."* P1's reflections on the mismatch between archival material and standards for describing them reflect Bowker and Star's (1999) characterization of classification as an attempt to give structure to documentation that inevitably emphasizes one narrative and silences others.

Just as classification is simultaneously "necessary" and "dangerous" (Bowker and Star, 1999), so too is description. Choosing what to describe in detail and what to summarize more vaguely privileges certain points of view over others (Duff and Harris, 2002). Data, whether in catalogs or otherwise, are abstract representations that reflect particular assumptions about the world (Gitelman and Jackson, 2013). Through first-hand experience with classification and description, workshop participants were well aware that the effort to avoid harmful biases in data will always be a work-in-progress effort. Similarly, ML models, which rely on data, also are always works-in-progress that need updating over time and across use cases (Ciora et al., 2021; Goree and Crandall, 2023; Havens et al., 2020, 2022; Rodolfa et al., 2020). As Eubanks (2017) notes, models are always outdated because they rely on data produced in the past. Drucker (2021) writes, "The gap between representation and phenomena can never be closed...the point is to push the recognition of the interpretative machinations of the representational process into view" (p. 565). Despite participants' disagreements on the accuracy of the annotations, the annotations do bring a particular interpretation of language into view, prompting critical reflection on potential biases in descriptions (**O5**, **O7**).

Tai's (2021) framework of cultural humility recognizes the inability of members of the GLAM sector to be experts in every community represented in GLAM collections and their documentation. Such expertise is unrealistic

to ask those who classify and describe collections to achieve (**O13**). Rather, those who classify and describe GLAM collections should aim to be stewards of the collections (Odumosu, 2020; Tai, 2021; Wurl, 2005). Their responsibility is not a responsibility to rectify all historical wrongs, but rather to approach collections and their documentation through a lens of "feminist ethics of care" (Caswell and Cifor, 2016). Workshop participants' discussion of the importance of including as many terms as possible in documentation (**O15**), including historical terms considered derogatory today, demonstrates how they work to live up to that responsibility, giving people both past and present the power to describe themselves with their chosen terminology.

The ways in which participants communicated their responsibility to consider past, present, and future stakeholders in GLAM collections (**O15**) demonstrates the suitability of the theory of heritage as a process. In Smith's (2006) conceptualization, heritage is a process of "passing on and receiving memories and knowledge" (p. 2), drawing on the field of critical discourse analysis that characterizes language, social practices, and societal power relations as intertwined (see §3.3 for a more detailed discussion of these theories). Smith (2006) writes about the political power that comes from choosing how one's own community will be described and identified. Participants' interest in reflecting on description practices (**O7**), dedicating time to consider where classification schemes and metadata standards simply do not fit, shows their recognition of the need for alternative heritage discourses that privilege, rather than silence, the perspectives of minoritized communities. That being said, I emphasize that those who write GLAM catalogs' metadata descriptions are not solely responsible, nor should they feel guilt or fault, for the injustices that minoritized communities of people continue to experience.

Drawing on feminist political philosophy, Young's (2011) conceptualization of "structural injustice" provides a "social connection model" of responsibility relevant to the challenge of catalog biases, as well as data biases more broadly. Structural injustice refers to societal structures with highly complex relationships between actors and their actions. Young's (2011) social connection model of responsibility is primarily forward-looking, focused on taking constructive action rather than assigning blame. At times during the workshop, participants expressed concerns about distinguishing between different sources of bias (e.g. bias from heritage material, bias inherited from

a previous catalog, or bias from a cataloger's word choice for a description). I argue, however, that the source of bias should not be a focus in ML research on social biases and related topics of fairness, ethics, and responsibility. Recalling the complications from professional and institutional barriers (**O9**), I argue that while members of the HC team, and other GLAM institutions, have a responsibility to try to counteract structural injustices reflected and perpetuated in their catalogs' documentation, they are not solely responsible for, nor are they to blame for, biases in that documentation. Young (2011) suggests collaborative action involving people from minoritized communities as particularly important for counteracting structural injustice, because members of those communities will better understand how those injustices affect them than people who do not experience harms from those injustices.

Along with the feminist theorizing of Young's (2011) structural injustice and Caswell and Cifor's (2016) ethics of care, I looked to D'Ignazio and Klein's (2020) feminist approach to Data Science, data feminism. Data feminism's seven principles include, "elevate emotion and embodiment." Participants' comments suggested feelings of discomfort, for example, P3 saying, *"It's very self-revealing to say what you think,"* and multiple participants provided specific examples to demonstrate the difficulties of tracing the origins of biases in a description. I am grateful for participants' willingness to put themselves in a place of discomfort and participate in the workshop, because their participation enabled me to bring the discussion of biases out of the technical and into the social realm, where biases have impact. That being said, participants' discomfort also points to a direction of future work on qualitative evaluations of data biases: how can one highlight potential data biases without implying blame on any particular person or organization for the presence of bias? Participants' expressions of discomfort and of doubt in the models' capabilities could indicate an aversion to being told how to think, a well-known response described in Psychology literature with the theory of reactance (Brehm and Brehm, 1981). Looking to the field of Design for guidance, "Embedded Design" (Flanagan and Kaufman, 2017; Kaufman et al., 2021; Kaufman and Flanagan, 2015) offers an approach for future work. Embedded Design uses subtle design mechanisms, rather than the overt text visualizations of the workshop worksheets, that has been shown in game research to avoid prompting reactance and reduce game players' social

biases (Kaufman and Flanagan, 2015).

In addition to the self-reflection that participants spoke of the models facilitating for them and their descriptive practices, participants' feedback facilitated my own self-reflection. I have attempted to document my positionality and biases throughout my research, recognizing that I am susceptible to the biases of the societal structures in which I live. That being said, this recognition does not ensure I can avoid its pitfalls. When P10 said that "housekeeping" or "coronation" could have had an *Occupation* label, I realized that the instructions I had written for applying the Taxonomy upheld the authorized heritage discourse (Smith, 2006) around occupations. With the aim of maximizing annotator agreement, I had instructed annotators to label job titles with the *Occupation* label, and not to label text that referenced work or employment without a job title (Appendix D). My instruction perpetuated the silencing of a minoritized perspective that I was trying to bring to the forefront. Had I not focused on the formality of a job title, *Occupation* annotations could have called attention to the under-documented, under-recognized histories of women working. As P9 stated, women often were not allowed to have professions historically. Annotating text that referred to work women performed (whether compensated or not), even if they were not described with a job title, could have highlighted areas for further investigation during collection reviews, as well as for historians researching contributions of historical women. This will have to be an area for future work.

7.7 Conclusion

The participatory evaluation this chapter reports has implications for bias mitigation in ML (§7.7.1), and for GLAM and ML collaborations (§7.7.2).

7.7.1 Bias Mitigation in ML

Collaboration with stakeholders offers insights on the complexities of bias and the uncertainties in data. In my workshop, participants explained how the presence of gendered language, such as the title "Mrs.," enables them to identify opportunities to counter the authorized heritage discourse that under-documents women's contributions (**O1**). Participants also spoke of

the difficulty, and often impossibility, of distinguishing between information a previous cataloger assumed and information that cataloger knew (**O2**). The challenge of tracing the origins of language, and thus the biases in that language, give further impetus to adopting Young's social connection model of responsibility, focusing on taking action to make improvements for the future, rather than attempting to assign blame to past actors. Participants' insights on model data guides ML researchers and practitioners in focusing their models on tasks that stakeholders consider both valuable and reliable.

Participatory approaches to ML are particularly valuable for bias mitigation due to the distinct perspectives of collaborators relative to an ML researcher or practitioner. Workshop participants' critiques of *Occupation*'s application (**O3**) brought my attention to a bias against women I wrote into my annotation instructions (Appendix D). Participants' critique of the *Unknown* label's application (**O2**) communicated similar concerns, with certain participants viewing its application as perpetuating the invisibility of women in historical records. Participants' willingness to share their reactions to the information I presented to them facilitated my own self-reflection, heightening my awareness of biases I unintentionally perpetuated in my work.

7.7.2 GLAM and ML Collaborations

For interdisciplinary work between GLAM and ML, GLAM collaborators and stakeholders do not need deep computational or statistical knowledge to critique the training data and predictions of ML models. Workshop participants provided valuable feedback on my classification models' training data (*Activity 1*, the Taxonomy's application) and predictions (*Activity 2*, the cross-collection measurements) after my conceptual overview of the supervised learning approach I employed to create the models. For example, my workshop participants identified four areas of value that my models could provide for their work (§7.6). Nonetheless, members of the GLAM sector may not be familiar with considering their collections' biases (or other characteristics) at a high level, due to their typical focus on individual cultural heritage records or individual collections. Offering a greater variety of cross-collection measurements to members of the GLAM sector than I provided in *Activity 2* may yield more discussion on the potential value of such high-level views of

GLAM documentation.

ML researchers and practitioners collaborating with members of the GLAM sector, or working with GLAM data, should be aware of the distinct characteristics of GLAM data. For example, in addition to documenting cultural heritage records, the metadata descriptions of GLAM catalogs are themselves cultural heritage records. As such, approaches to hate speech detection and and gender bias that remove or change language in the data are unsuitable. GLAM documentation provides GLAM visitors with evidence of historical biases as well as shifts in attitudes and perspectives that have moved society towards a more just future. Workshop participants' approach to adding terminology to existing descriptions, rather than replacing outdated terminology with contemporary terminology (**O15**), provides an example of this practice. The conscientiousness of members of the GLAM sector to stakeholders past, present, and future demonstrates an acknowledgement of responsibility much needed in ML research and practice. Furthermore, the understanding of the situated nature of collections and their documentation in the GLAM sector offers a valuable approach to data for ML that could minimize the harms from biased ML systems resulting from oversimplified conceptualizations of bias.

Areas of research for future work on GLAM and ML collaborations for bias mitigation include information design, annotation approaches, and approaches to qualitative model evaluation. Future work is needed to experiment with text visualizations that draw attention to potentially biased language without implying an intention of removing or changing that language (**O2**). Annotation approaches that encode assumed gender information (**O1**) without risking representational harm by perpetuating gender biases is also an area for future research; literature on visualizing data uncertainties or multiple perspectives on a dataset could offer a starting point (Havens, Bach, et al., 2022; Kay et al., 2016). Lastly, future research on qualitative model evaluation approaches should investigate how to incorporate ML models into existing GLAM workflows, such as those for collection reviews or descriptive practices, that enable GLAM collaborators to evaluate the models within the context of their work, comparing their own knowledge to the performance of the models.

7.8 Recalibrations with the Participatory Evaluation

In this chapter, I continued the recalibration of ML demonstrated throughout this thesis with a human-centered approach to evaluating NLP models. With the workshop for participatory model evaluation, I prioritized:

- **Quality over quantity** by asking the HC team, represented by the 10 workshop participants, to provide feedback on the application of the Taxonomy of Gendered and Gender Biased Language (Chapter 5). I and the hired annotators applied the Taxonomy to the HC Archives documentation in order to create an annotated dataset for training text classification models (Chapter 6). The application of the Taxonomy thus shapes the quality of the annotated dataset and the models trained on that dataset.

- **Accuracy over efficiency** by conducting multiple types of evaluation on the classification models I report, recognizing that standard benchmarks and metrics to evaluate models for biases have yet to be well established (Welty et al., 2019). Chapter 6 provided the quantitative evaluations with standard NLP metrics (i.e. precision, recall, and F_1 score) and this chapter provides a new, qualitative approach to evaluating models in collaboration with model stakeholders (i.e. a workshop with the HC team).

- **Representativeness over convenience** by engaging members of the HC team, who are domain experts on my dataset of HC Archives documentation, to provide a qualitative evaluation of my annotated data and classification models.

- **Situated thinking over universal thinking** by continuing to position my work in a case study, facilitating this chapter's qualitative evaluation with representatives of the HC team for my work with HC Archives documentation.

Chapter 8

Discussion

> "There is no use trying," said Alice; "one can't believe impossible things." "I dare say you haven't had much practice," said the Queen. "When I was your age, I always did it for half an hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."
>
> –Lewis Carroll, *Through the Looking-Glass* (1871, 2019 ed., p. 227)

I investigated three research questions for this thesis:

1. Can existing methods of identifying and categorizing gender biased language in Natural Language Processing (NLP) research be applied to archival metadata descriptions ("archival documentation")? Why or why not?
2. What types of gender bias are present in the language of archival documentation?
3. Can gender biased language in archival documentation be reliably annotated by domain experts to create data on which to train NLP classification models?

This chapter describes challenges I faced while investigating these questions

(§8.1); discusses limitations of my research (§8.2); summarizes implications of my main argument, looking to the Gallery, Library, Archives, and Museum (GLAM) sector to guide a recalibration of Machine Learning (ML) for social biases (§8.3); and proposes directions for future work (§8.4).

8.1 Challenges

While carrying out the research reported in this thesis, I faced challenges regarding the nature of Participatory Action Research (PAR), and the uncertainty and subjectivity of gender, bias, and language. During PAR activities I conducted to collaborate with the Heritage Collections (HC) team, I faced the challenge of attempting to meaningfully engage with the team while also respecting their responsibilities as HC employees. Rachel Hosker, University Archivist and Research Collections Manager in HC, expressed enthusiasm for engaging with my research from the time I submitted my Ph.D. research proposal, and encouraged her colleagues to engage with my research as well. That being said, as with many GLAM institutions, the HC Archives is always in need of additional resources to acquire, appraise, describe, preserve, manage, and facilitate access to its collections. I needed to ensure I did not request too much of Hosker and the HC team, which would push the collaboration towards being beneficial only for myself rather than being mutually beneficial. Furthermore, as a Ph.D. project, the expectations for me as the Ph.D. researcher were to primarily execute the work myself. The challenges of any collaborative or otherwise participatory research effort will be unique to the funding situation and employment obligations of those involved.

The uncertainty and subjectivity associated with gender, bias, and language also posed challenges for my research. Grammatical gender does not correlate precisely with gender identity (Scheuerman, Spiel, et al., 2020; Shopland, 2020; Spiel et al., 2019). The use of a feminine term such as “Ms.” does not tell you a person’s gender identity with certainty. A historical person may have used different terminology to describe their gender identity than we use today (Shopland, 2020; Chapter 7). A person’s gender identity may change over time, and if GLAM documentation provides evidence of a person’s gender transition, the person’s safety or job security could be impacted (Dunsire, 2018). Moreover, calling out a person’s gender, if among minoritized genders,

contributes to the othering of that gender (Hessel, 2023a, 2023c) and reinforces the authorized heritage discourse (Smith, 2006).

That being said, recording and considering indicators of historical figures' likely or possible gender identities is necessary for studying how privilege and oppression have operated and continue to operate along the axis of gender. A Whitechapel Gallery curator explained that specifying the gender of artist Lillian Holt could "show how isolated in her gender she was, and how the Borough Group, of which she was a member, is always written about in regard to the men included in it, rather than her" (Hessel, 2023c). Keskustalo et al. (2023) write of the value of studying letters exchanged between people of the same and different genders for understanding human emotions and relationships. During the workshop discussion (Chapter 7), participants discussed the possible value in assuming a person's gender identity based on how their appearance fits gender stereotypes. If assumptions that reinforce gender biases have been made based on images of a person or other people's descriptions of that person, these assumptions reflect how that person would have moved through society, providing insight on how and where the person would have experienced privilege and oppression.

Nonetheless, adhering unquestioningly to stereotypical gender representations when describing cultural heritage material would fail to address the changes of gender as a concept over time, and the reality of the diversity within each gender identity. While gender stereotypes may have their value for historical research, they also are a source of oppression for minoritized gender groups, namely people who do not identify as a cisgender man. With my thesis I urge people to take a more critical approach to gender, questioning the most prevalent representations and interpretations. In an interview with art historian Hessel, classicist Beard speaks of how men have relied on loud and extreme depictions of women in an effort to make misogyny seem "natural" (2022). A careful revisiting of historical and archaeological evidence, however, uncovers numerous examples of societies in which men were not the dominant gender group (Graeber and Wengrow, 2021). Crumley (1995), for example, introduced the term "heterarchy" to describe societies in which power shifted between gender groups based on contextual factors, such as the seasons or whether a society was in a time of peace or conflict. This depiction of power challenges the hegemonic idea that shifting power relations in society must be accompanied

by instability and upheaval (Graeber and Wengrow, 2021).

As Beard (2017) suggests, perhaps in order to address the uneven distribution of power between genders, we should change how we think about power, rather than urging women or other minoritized gender groups to change. While advice from books such as Sandberg's *Lean In* (2015), and Babcock and Laschever's *Why Women Don't Ask* (2008) and *Ask For It* (2009) certainly were written to support and encourage working women, changing one's behavior as a woman in an attempt to mirror the men around you can backfire. Examples of attempts to silence women who participate in conversations and decision-making in both public and private spheres stretch back to antiquity (Beard, 2017). Historical and contemporary descriptions of women's voices have attempted and continue to attempt to undermine women's authority and right to speak, from complaints about the higher pitched sound of a woman's voice or the use of derogatory verbs such as "bark" and "whine" rather than "speak" or "say" (ibid.). These sexist depictions have become a part of Western cultural infrastructure, making them appear natural and inevitable when in fact they are entirely constructed (Beard, 2017; Graeber and Wengrow, 2021). As human-made, socio-technical systems, ML systems replicate these cultural constructions, which is why presenting these systems as objective or capable of predicting the future is dangerous.

So where to begin? If anything is to be changed, it must first be noticed. As set out in Chapter 1, my motivation for creating models to *identify* gender biased language came from the gap in understandings of bias in ML. Publications on minimizing or removing bias rarely defined bias, and how can something undefined be minimized, removed, or even measured? Due to the value of studying the past for understanding and improving the present (Graeber and Wengrow, 2021; Hicks, 2018, 2021; Smith, 2006), along with the GLAM literature discussing how describing collections can either uphold or push back against societal power relations (Caswell and Cifor, 2016; Ortolja-Baird and Nyhan, 2022; Tai, 2021), I positioned my work at the intersection of GLAM and ML. By creating gender biased text classifiers that annotate potential gender biases in archival documentation, I aim to make the gender biased structures of society visible, encouraging people to notice, question, and discuss them. I present the work of this thesis as a response to Drabinski's (2013) call to make the underlying structures of a

GLAM catalog *visible*, and to Beard and Hessel's urging not to remove, but to *take notice* of how images and language uphold biased and unjust structures in society (Beard, 2017; Hessel and Beard, 2022). Once we notice unjust structures, we can study how they operate and begin to deconstruct them, replacing them with more equitable societal structures that do not replicate the mistakes of the past (Friedman and Nissenbaum, 1996; Hicks, 2018; Morrison, 2021). This thesis contributes to the first step, noticing unjust structures, by creating models that encourage people to take notice of the gender biases written into the descriptive language of an archival catalog.

8.2 Limitations

There are several limitations to the work I report in this thesis regarding the use of text classification models, the annotated data on which those models were built, the PAR activities, and the case study approach. As with all ML models, the text classification models I created are limited by the data on which they are trained and tested (chapters 5-6). Though the quantitative measures of inter-annotator agreement for the aggregated dataset that I used to train, validate, and test the models are reasonably high, the annotation of language, particularly biased language, is a subjective task. HC team members' conflicting attitudes towards annotators' application of the Taxonomy's labels (Chapter 7) demonstrated the difficulty of evaluating annotations of biased language, whether model-made or human-made. Iterative collaboration with model stakeholders is valuable for creating ML models that meet those stakeholders' expectations, but the process may not lead to consensus. Interpretations of data and opinions about what is a "correct" annotation will inevitably vary from one person to another.

Moreover, regarding the annotated datasets of archival documentation (Chapter 5), the contents of the datasets should be expanded to more comprehensively account for different types of gender biased language. Archival documentation that includes greater representation of trans, non-binary, and gender diverse identities is needed to more fully account for biases against these gender groups. Collaborations with community-run Archives that focus on managing the cultural heritage records of minoritized gender groups could help to broaden the perspectives currently represented in

the datasets. That being said, this additional data should be obtained with the consent of the Archives, and should be interpreted with domain experts, ideally those who work or volunteer at the Archives. Funding would likely be needed from external grants to enable this collaboration, as manually annotating language is a labor-intensive process.

The annotated data of this thesis are also limited by the interpretation required to create data, and thus the inevitable subjectivity of any dataset (Drucker, 2021; Gitelman and Jackson, 2013). In an ideal world, evaluations of biased language in GLAM documentation would be based not only on reading the descriptive language of the documentation, but also on reviews of the cultural heritage records being described. Due to the resource constraints GLAM institutions typically face (e.g. the need for an "army of catalogers" that P1 spoke of in Chapter 7), reviewing all, or even a majority of, cultural heritage records alongside their descriptions was not feasible.

As noted in Chapter 4, collaborating with all stakeholders of an ML system may not be possible, and was not possible for the research I report in this thesis. Stakeholders who were unavailable for me to collaborate with include former employees of the HC Archives and the people represented in the HC Archives collections. Furthermore, though I collaborated with members of the HC team throughout the entirety of my research, perspectives from the HC Archives' User Services team are missing, as noted in Chapter 7. Additionally, though I collaborated with two HC Archives' visitors as annotators during the manual annotation process, the perspectives of visitors are missing from the data and model evaluation process.

Lastly, the case study approach has limitations due to its narrow focus on a particular context. As discussed in Chapter 4, case studies are well-suited to research bias because bias is highly situated, varying with cultural, political, temporal, and linguistic contexts. Nonetheless, a series of case studies is needed to better distinguish which aspects of a research project are generalizable across contexts and which aspects of a research project are unique to a specific context. Drawing on my knowledge of documentation practices in GLAM, I suspect that the text classification models would translate to other Archives more than Galleries, Libraries, or Museums. That being said, experiments running the models on other GLAM catalogs and evaluating the models with those catalogs' experts are needed to more precisely evaluate

the generalizability of the models. Similarly, drawing on my knowledge of cultural and linguistic similarities among English-speaking countries, the types of gender biased language in an archival catalog in the UK are likely to be similar to the types of gender biased language in an archival catalog from the US, Canada, or Australia. Still, experiments running those models on other countries' Archives catalogs are needed to measure the similarities and differences in the types of gender biased language across those catalogs.

8.3 Implications

In this section I summarize the recalibration of ML that I have proposed and demonstrated in this thesis (§8.3.1), and the guidance the GLAM sector offers in enacting this recalibration (§8.3.2).

8.3.1 ML Research and Practice

I have argued for the importance of recalibrating ML research to empower minoritized communities and more effectively mitigate harms from social biases. My contributions in this thesis demonstrate how the recalibration may be executed throughout the entire process of creating an ML system. In my research, I drew on GLAM literature regarding the power and subjectivity in GLAM cataloging practices (Adler, 2017; Adler and Harper, 2018; Agostinho et al., 2019; Duff and Harris, 2002; Olson, 2001; Smith, 2006) to guide my definition of four new prioritizations for ML. Here I discuss the implications of the four priorities I proposed to make this recalibration.

8.3.1.1 Quality over Quantity

Prioritizing quality over quantity involves more critical approaches to data curation, including the collection and interpretation of data. I carried out the prioritization of quality in my data curation by:

- Defining a specific bias of focus, namely gender bias in British English text (Chapter 4),
- Identifying stakeholders to collaborate with on evaluating my annotated dataset and classification models (Chapter 4),

- Curating a bespoke dataset containing only text relevant to my research context (Chapter 5),
- Hiring a small number of annotators to label that dataset (Chapter 5),
- Creating text classification models through supervised learning with that annotated data only (Chapter 6), and
- Evaluating the annotated data and models qualitatively in a workshop with HC team members (Chapter 7).

For the ML community, improving models' data quality will facilitate greater reproducibility. More careful collection, curation, and interpretation of data will require more detailed documentation of data-related decisions. More detailed documentation would facilitate reflection among an ML project's researchers or practitioners, while also enabling future researchers or practitioners to more easily replicate, anticipate harms from, and build upon the work (Bender and Friedman, 2018; Gebru et al., 2018; Yang et al., 2018).

Improving models' data quality will require greater collaboration with diverse groups of ML model stakeholders (Blodgett et al., 2020; Crawford, 2017; Devinney et al., 2022; Stańczak and Augenstein, 2021). Domain experts on different subsets of a model's training data can provide guidance on collecting, curating, and interpreting that subset's data. Literature on creating quality GLAM documentation holds relevance for ML. Duff and Harris (2002) remind their reader that archival records are "understandable only in the ever-changing broader context of society" (p. 263) which is applicable to all data (Gitelman and Jackson, 2013), including ML datasets (Eubanks, 2017). Tai's (2021) framework of cultural humility for members of the GLAM sector is also applicable to ML researchers and practitioners: no one person can be an expert in all the communities of people represented in an ML dataset. Greater collaboration with the public would not only lead to better quality data, but would also lead to greater ML literacy.

Principles from social justice oriented approaches to research can provide guidance to researchers and practitioners interested in critical approaches to dataset creation that emphasize data quality. With data feminism, D'Ignazio and Klein (2020) propose one such approach, drawing on intersectional feminism to encourage data-driven work that empowers minoritized communities. The authors distinguish between data work that secures power and data work that

challenges power. Challenges to societal power structures consider context, reflexivity, and co-liberation to develop lasting, holistic datasets and data-driven systems (ibid.). Similarly, Aragon et al. (2022) emphasize the importance of accompanying analysis with reflection for ethical, human-centered Data Science. Costanza-Chock's (2018) *design justice* also offers a context-oriented approach, focusing on collaboration with communities to guide design processes towards solutions that empower them.

8.3.1.2 Accuracy over Efficiency

Prioritizing accuracy over efficiency in ML system creation involves collaboration with stakeholders end to end, from problem definition to dataset curation to model evaluation. In my research, this included:

- Collaboration with the HC team to set the focus of my research as gender biased language (Chapter 4),
- Collaboration with the HC team to determine which metadata fields' descriptions to extract from the HC Archives' catalog (chapters 4-5),
- Collaboration with the HC team to finalize the Taxonomy of Gendered and Gender Biased Language (Chapter 5),
- Collaboration with visitors to Archives as hired annotators to label HC Archives documentation according to the Taxonomy of Gendered and Gender Biased Language (Chapter 5),
- Collaboration with Rachel Hosker, University Archivist and Research Collections Manager in HC, to write a publication on the gender biased classification of HC Archives (Havens et al., 2023; Appendix K), and
- Collaboration with the HC team to evaluate the manual and automated applications of the Taxonomy (Chapter 7).

For the ML community, prioritizing accuracy over efficiency in the research process requires more critical thinking about the way in which an ML system's performance is evaluated. Choosing or creating suitable metrics, benchmarks, and measurement instruments for a model requires reflection on the model's underlying data and the model's use case (Jacobs and Wallach, 2021; Raji et al., 2021; Welty et al., 2019). Though this reflection takes additional time, it will improve ML researchers and practitioners' ability to

anticipate and minimize harms from inaccuracies, mistakes, and biases in models, benefiting all model stakeholders. Qualitative research practices from Human-Computer Interaction, Design, and the Social Sciences offer examples of how to incorporate stakeholder feedback in data and model evaluation processes. Regarding qualitative dataset evaluation, Aragon et al. (2022) summarize qualitative methods relevant for human-centered approaches to Data Science. Among others, the authors encourage people working with data to use thematic analysis, action research, and participatory design. Regarding qualitative model evaluation, Markl and Lai (2021) propose evaluating the performance of Automatic Speech Recognition systems using a combination of quantitative and intersectional benchmarks, qualitative error analysis, and user experience methods (e.g. surveys, interviews, ethnography).

Approaches from the GLAM sector offer guidance on how to evaluate ML systems' accuracy with less efficiency and greater criticality. In Archival Science, Duff and Harris (2002) suggest that the concept of provenance, referring to the origins of archival records, be reconceptualized in the plural: provenances. They write of the multitude of perspectives, contexts, and relationships that should be considered to more accurately represent "the complex, messy present and the pasts it invokes" (p. 280). The authors' acknowledgement of an archivist inevitably making interpretations when describing archival material recalls Smith's (2006) theory of heritage as a process of continuous usage, recreation, and adaptation. Thus just as prioritizing quality over quantity requires additional time for reflective practices, so too does prioritizing of accuracy over efficiency. The implications of data's multiplicity and incompleteness deserves greater reflection in ML research and practice (D'Ignazio and Klein, 2020; Drucker, 2021; Duff and Harris, 2002; Haraway, 1988; Harding, 1995).

8.3.1.3 Representativeness over Convenience

Prioritizing representativeness over convenience in the ML system creation process requires reflection upon how well data can reflect the real-world context within which an ML model is intended to function. In my research I let this reflection guide my decisions regarding:

- The data source for my models' training, validation, and test data, i.e. the

HC Archives catalog only (Chapter 5),

- The word embeddings I used to represent word meanings for my text classification models, i.e. custom embeddings trained on HC Archives documentation only (Chapter 6),
- The models I chose for classifying gender biased language, i.e. traditional ML rather than pre-trained, deep learning models (Chapter 6), and
- The people I collaborated with to annotate data, i.e. Gender Studies and Archives experts (Chapter 5), and to create and evaluate data and models, i.e. representatives of the HC team (chapters 4, 5, and 7).

Due to ethical concerns with scraping web data without data owners and producers' consent (Buolamwini and Gebru, 2018; Crawford, 2021), environmental concerns with the cost of training large models (Bender et al., 2021; Strubell et al., 2019), social justice concerns regarding the perpetuation and amplification of biases in models created with convenient-to-collect data (de Vassimon Manela et al., 2021; Jentzsch and Turan, 2022; Jiang and Fellbaum, 2020; Jin et al., 2021; Lu et al., 2020; Martinková et al., 2023; Tan and Celis, 2019), and ethical and validity concerns with approaches to creating large-scale annotated data (Blodgett, Lopez, et al., 2021; Irani, 2015, 2016; Kreutzer et al., 2022), research on approaches to creating more representative datasets and models would benefit the ML community and broader public. History illustrates that research built upon what is conveniently available leads to unethical practices across disciplines that reinforce unjust societal power structures (Perez, 2019). Consider the use of Henrietta Lacks' cancerous cells for medical research without her consent (Skloot, 2011) and the undervaluing of crowdworkers' labor (Gray and Suri, 2019; Irani, 2015). To create ML systems that represent the systems' stakeholders, and do so in an accessible and socially beneficial manner (Verdegem, 2021), existing ML practices must be rethought.

To create representative systems, ML researchers and practitioners should not speak for communities to which they do not belong. Rather, ML researchers and practitioners must engage in collaborative, participatory processes with communities outside their own, creating space for people to communicate their perspectives and experiences for themselves (Broussard, 2019; Irani, 2016). When writing of librarians and their cataloging work, Olson (2001)

writes, “Instead of possessing power exclusively, we who are on the inside of the information structures must create holes in our structures through which power can leak out” (p. 659). I argue that not only does the ML community need to “subdivide and reengineer failing systems at a basic, structural level” (Hicks, 2021, p. 155), it also needs to subdivide and reengineer the processes and practices that create those systems in the first place, making space for communities to exert power in their reengineering.

Digital technology’s rapid development has far outpaced legislation, however historical examples of legislation passed in reaction to harmful research and working practices illustrate the possibility of preventing the same harms in the future. Such legislation can empower minoritized communities of people to push back against and prevent oppression. For example, *The Belmont Report* of 1979 was made in response to the Tuskegee Syphilis Study in the US, and led to the creation of Institutional Review Boards to more rigorously uphold ethical standards in American research practices. (Department of Health, Education, and Welfare, 1979). Approaching ML systems through the lens of critical discourse analysis (Bucholtz, 2003; Fairclough, 2003; van Leeuwen, 2009), viewing data as discourse, provides a useful framework for understanding the power exercised in ML dataset creation and use, and thus identifying where legislation could be developed to adjust imbalanced power relationships between ML systems’ creators and stakeholders.

8.3.1.4 Situated Thinking over Universal Thinking

Prioritizing situated thinking over universal thinking encourages ML researchers and practitioners to approach ML systems as socio-technical, rather than purely technical, systems. In this thesis, I present my research as a case study to emphasize the situated nature of my research, describing:

- My cultural, linguistic, and temporal context, i.e. researching as an American ciswoman in the 21st century using a corpus of British English text from the University of Edinburgh’s HC Archives’ catalog; and
- My data, models, and quantitative and qualitative evaluations relative to that context.

Providing these case study details enables future researchers to compare and contrast their research context with mine, and thus make informed decisions

about which aspects of my research they should replicate and which aspects they should tailor to their own research. As discussed previously, considering context improves the quality, accuracy, and representativeness of ML systems.

More specifically, a consideration of contextual factors informs the way in which a model is designed to interpret data and to inform the formulation of the task a model is meant to complete. For example, considerations of the language and accents of a region in which an NLP or Automatic Speech Recognition (ASR) model will be deployed are necessary to ensure the model can serve the people living in that region. Supposedly universal, or generalizable, approaches to creating NLP and ASR models has resulted in their poor performance with non-dominant variations of English, despite English itself being the dominant language in speech and language technology research (Blodgett et al., 2016; Markl, 2022a). Samorani et al. (2022) provide an example of how considering contextual factors informs task formulation.

Working on medical appointment scheduling in the US, Samorani et al.'s (2022) consideration of racist social structures that continue to oppress non-white, and especially Black, communities prevented the replication of those unjust social structures in their proposed scheduling model. The authors' ML model experiments showed that so-called state-of-the-art approaches to creating fair algorithms exhibited more racial biases than the model they created. The authors' model used an algorithm that aimed to minimize the longest wait time any patient could be assigned, while state-of-the-art approaches relied on algorithms that applied weights based on demographic data. Thus Samorani et al.'s (2022) ML model was both less biased and less invasive, as it didn't require data on patients' personal information such as racialized ethnicity. Again, critical discourse analysis provides a useful theoretical lens for thinking about ML models and their data, particularly for text-based data, because this theoretical framework considers history to analyze the power relationships at play in discourse, i.e. written or spoken language (Bucholtz, 2003; Fairclough, 2003; van Leeuwen, 2009).

8.3.2 A Leading Role for GLAM

As the size of ML training data and and model parameters has grown, so too has the cost of the computing resources needed to analyze the data and

train the models. As a result, corporations such as Google and OpenAI, with an investment from Microsoft, have been major players in ML research and model creation (Devlin et al., 2019; OpenAI, 2023); academic research labs and public institutions cannot afford the same computing resources of those corporations (Bommasani et al., 2021). Though corporations often do not publish the training data and architecture of their models, they do publish the models themselves as pre-trained models, meaning trained on a large and, problematically, unknown dataset. Researchers and practitioners can then fine-tune, or customize, the model to a different dataset without having to train the model from scratch. Bommasani et al. (2021) term such pre-trained models "foundation models," indicating their widespread use as well as their unfinished nature, being intended for customization to particular tasks.

Foundation models' widespread use crosses into research and industry. For example, at the time of writing, on the ML model and dataset platform HuggingFace,[1] 13 out of the top 30 most downloaded models were a derivative of Google's generative language model BERT (Devlin et al., 2019), and Google, GitHub and Overleaf have deployed customized foundation models for online search,[2] programming,[3] and writing in LaTeX,[4] respectively. The risk with the widespread use of foundation models is that they are created and deployed by corporations with commercial motivations, prioritizing quantity, efficiency, convenience, and universal thinking. Corporations are not incentivized to upend society's existing power structures, because those power structures serve their commercial interests (Noble, 2018; Thornton, 2017).

There is an opportunity for the GLAM sector to lead ML system creation in a new direction, demonstrating an alternative approach that prioritizes quality, accuracy, representativeness, and situated thinking. GLAM institutions exist to provide access to knowledge for the public. Noble (2018), Cordell (2020), and Jaillant (2022) have written of the role Libraries can play by emphasizing ethics and diversity instead of commercial interests. Moreover, as described across this thesis, researchers and practitioners across Galleries, Libraries, Archives, and Museums have already been grappling with the inevitable challenges of bias in collections and collections' documentation (Adler, 2016, 2017; Adler

[1] huggingface.co
[2] blog.google/products/search/search-language-understanding-bert
[3] github.com/features/copilot
[4] www.writefull.com/writefull-for-overleaf

and Harper, 2018; Berry, 2020; Caswell, 2022; Caswell and Cifor, 2016, 2019; Duff and Harris, 2002; Noble, 2018; Odumosu, 2020; Padilla, 2019; Schwartz and Cook, 2002; Stoler, 2002; Tai, 2021). The recognition of the partiality of data in GLAM positions the sector well to lead the ML community away from technochauvinist approaches and towards human-in-the-loop approaches, where human-machine collaboration is recognized as a more promising path for improving technology than fully autonomous systems (Averkamp et al., 2021; Broussard, 2019, 2023; Cordell, 2020; Irani, 2016; Noble, 2018).

While ML models' ability to automate processes at large scale certainly offers efficiency gains for GLAM, such as automating and scaling up metadata creation (Padilla, 2019; Yilmazel et al., 2004), researchers and practitioners in the sector have taken a cautious and critical approach to ML deployment. Both independently and collaboratively, members of the GLAM sector have executed and commissioned case studies (e.g. the case studies of AI applications for Archives in Jaillant, 2022), experiments (e.g. training a language model to write like a historical cataloger in Baker, 2020), and reports (e.g. those commissioned by OCLC (Padilla, 2019) and the Library of Congress (Averkamp et al., 2021; Cordell, 2020)) to investigate both the capabilities and limitations of ML models. Just as the ML community is interested in reusable data and models (Bommasani et al., 2021; Raji et al., 2021), so too is the GLAM sector. That being said, the limits of supposedly universal data and models that have more recently become evident for the ML community (Birhane and Prabhu, 2021; Buolamwini and Gebru, 2018; Crawford and Paglen, 2019) have parallels with the GLAM sector that have already prompted large-scale changes among GLAM, such as efforts to incorporate localization in the global cataloging standard Resource Description Access (Dunsire and Willer, 2014). Collaboration between GLAM and ML researchers and practitioners offers the opportunity to address resource constraints in GLAM and recalibrate ML in support of social justice.

8.4 Future Work

There are many directions for future work on the data annotated for gender biases (Chapter 5) and training the gender biased text classification models (Chapter 6). As previously discussed (§8.2), the dataset could be expanded to

include GLAM documentation from catalogs beyond that of the HC Archives, and the application of the Taxonomy to annotate the documentation could be modified based on feedback from members of the HC team or other stakeholder groups. Regarding the text classification models, additional experiments could be run to further investigate model setups that yield the best performance, and alternative data perspectivist approaches could be applied to train models on the individual annotators' datasets (Basile et al., 2021; Davani et al., 2022), also previously discussed (§6.8). Experiments could be run with the models on English-language catalog documentation of GLAM institutions from other countries to study the similarities and differences in the gender biased language of those catalogs. The most exciting and promising directions for future work on gender biases in data and models that I suggest, however, revolve around interdisciplinary collaborations.

8.4.1 Collaboration with Artists and Designers

Thinking beyond the future research directions discussed in Chapter 6, on new model setups and experiments for gender biased text classifiers, I see exciting possibilities for future work on creative explorations of ML. ML researchers and practitioners could collaborate with artists and designers to facilitate critical reflection on ML system creation practices. As Flanagan (2009) writes, "artists can challenge ideas, beliefs, and social expectations and subsequently transform them in their work" (p. 12). For ML researchers and practitioners, creative engagements with ML can make invisible societal structures that are unintentionally engineered into ML systems visible. An awareness of how social biases uphold unjust societal structures enables ML researchers and practitioners to consider alternative, subversive practices to ML dataset and model creation. For the public, creative engagements with ML have exposed the inner-workings of models and the contents of large-scale data, improving ML literacy and informing legislative processes that can define the boundaries of permissible ML practices.

Artists already using ML systems with whom ML researchers and practitioners could seek collaborations include Pip Thornton, Mary Flanagan, and Mimi Ọnụọha. Thornton's *{poem}.py* (2016) interrogates the ML model behind Google AdWords, printing poems on receipts where each word in the

poem has a particular cost that sums to the total price of the poem. The artwork prompts reflection upon the commercialization of language and its influence on our word choice (Thornton, 2017). Flanagan's *[help me know the truth]* (Flanagan, 2017) prompts reflection upon the categorizations of computational neuroscience algorithms. The artwork actively engages viewers, asking them to take a selfie that is subsequently processed by algorithms to produce two images; viewers are then asked to categorize the images of their fellow viewers. Flanagan describes the piece as an investigation into "how computational neuroscience techniques can uncover the categorizing systems of the mind, and how they are therefore subject to socially constructed fears and values" (2017). Reflecting upon the role of, and perhaps over-reliance on, crowdwork in ML practices (Aroyo and Welty, 2015; Welty et al., 2019), the piece also calls to mind the judgments involved in crowdwork and thus the social biases encoded in ML systems reliant upon crowdwork. Ọnụọha's *The Library of Missing Datasets* (2016, 2017) prompts reflection upon the power exercised through data collection. The piece consists of a file cabinet of folders labeled with dataset names. Every folder, however, is empty. The piece calls attention to "cultural and colloquial hints of what is deemed important" (Ọnụọha, 2016). These three pieces demonstrate how artists challenge assumptions and beliefs about ML systems, and transform the way ML researchers and practitioners, as well as the public more broadly, view and interact with these systems.

In my own work, I've engaged in a Design method, Speculative Design, to envision a human-in-the-loop ML system for GLAM (Havens, 2021). Drawing on Dunne and Raby's *Speculative Everything* (2013), I designed a User Interface (UI) to a GLAM catalog that would allow the institution's community partners to contribute to the catalog's descriptions of heritage collections (Figure 8.1). With each revision to a catalog record, a new version of that record would be saved (Figure 8.2). GLAM employees and the public could view the different versions of the record (Figure 8.3), and could view when and which group made every revision to the record (Figure 8.4). The HC team's discussion of the importance of dating documentation and challenges with updating its language over time motivated the design of my speculative

UI. The Six Degrees of Francis Bacon[5] and The Pelagios Network[6] projects further inspired the partnerships I envisioned between a GLAM institution and community organizations. These projects have developed approaches to large-scale collaboration on building and expanding upon network data that consider domain expertise when determining who can contribute to their networks; GLAM could develop a similar approach for inviting community groups represented in their collections into the documentation process. I envisioned these community partnerships as one way in which the GLAM sector could "create holes in [its] structures through which power can leak out" (Olson, 2001, p. 659). Furthermore, recognizing the need for alternative information discovery processes, I designed an exploratory data visualization to guide visitors through the GLAM catalog. By avoiding a ranked list of search results, this exploratory interface emphasizes uncertainty and relationships, putting the visitor in charge of navigating and filtering more than an ML model behind the scenes.

This speculative UI then led me to consider the possibilities for NLP models that could make use of the dated versions of GLAM documentation. Envisioning a future in which NLP researchers and practitioners executed more work in case studies (Havens et al., 2020), and less work aiming for universally-applicable systems, I asked, *What could we learn about how language evolves if we had 100 years of revisions to our cultural heritage documentation to look back on?* This led me to design a textbook trained on the dated versions of cultural heritage records that were written collaboratively by GLAM experts, community organizations, and subject matter experts (e.g. a collection of piano music described by a pianist). Inspired by the work of Graeber and Wengrow (2021) and Beard (2017), which problematize the idea of a single past, or single historical narrative, I designed a textbook that tells history from multiple perspectives side by side (figures 8.5 and 8.6). The process of creating such speculative works such as the GLAM catalog UI and history textbook ensure that as we critique and deconstruct biased (or otherwise problematic) technologies, we have a blueprint for how we want to rebuild those technologies.

[5] www.sixdegreesoffrancisbacon.com
[6] pelagios.org

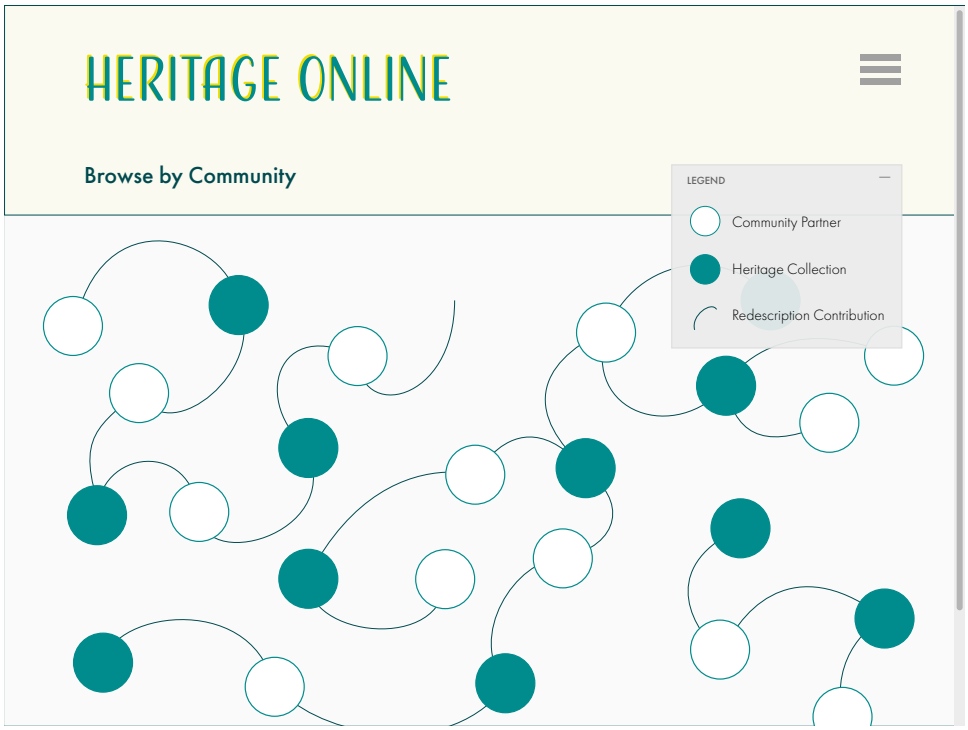


Figure 8.1: **Speculative cultural heritage catalog UI, network view.** Employees and visitors to the online catalog could browse the collections it documents in a network visualization that uses visual encodings to indicate which collections were described by a community partner.

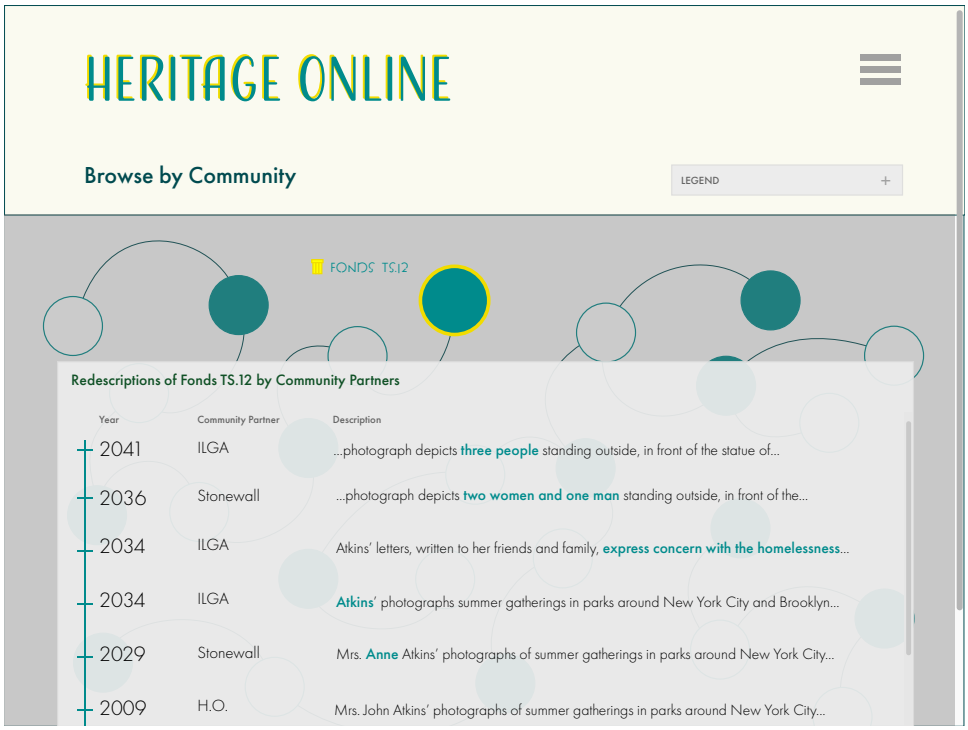


Figure 8.2: **Speculative cultural heritage catalog UI, network revisions view.** Employees and visitors to the online catalog could view revisions to collections' documentation, including and the group that made the revision and the date the revision was made.

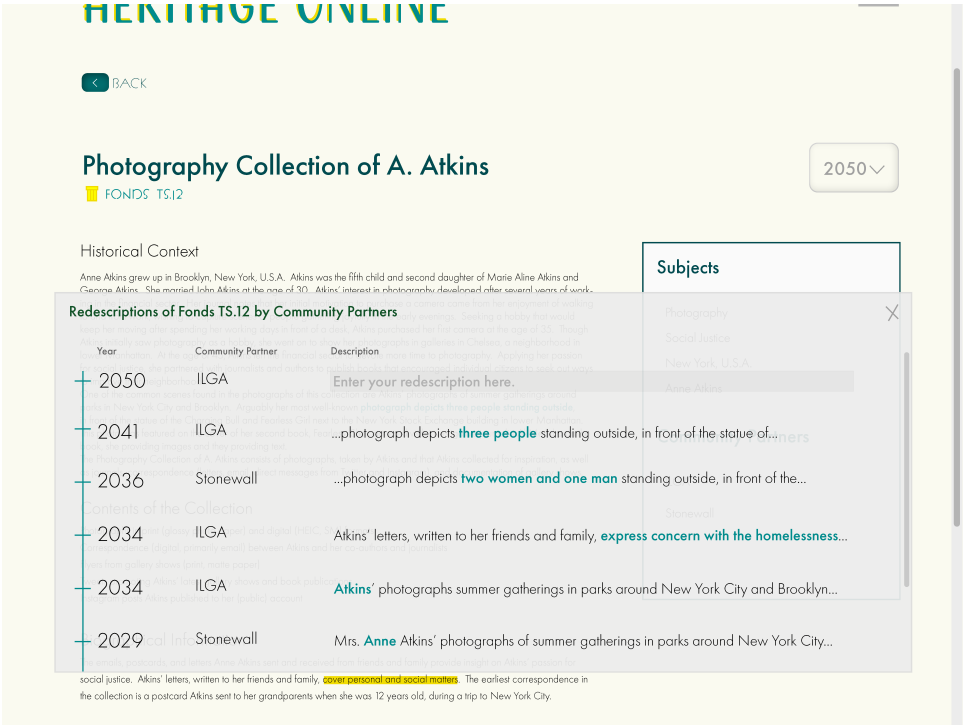


Figure 8.3: **Speculative cultural heritage catalog UI, record revisions view.** Employees and visitors to the online catalog could view revisions of a specific heritage record, including the group that made the revision and the date the revision was made.

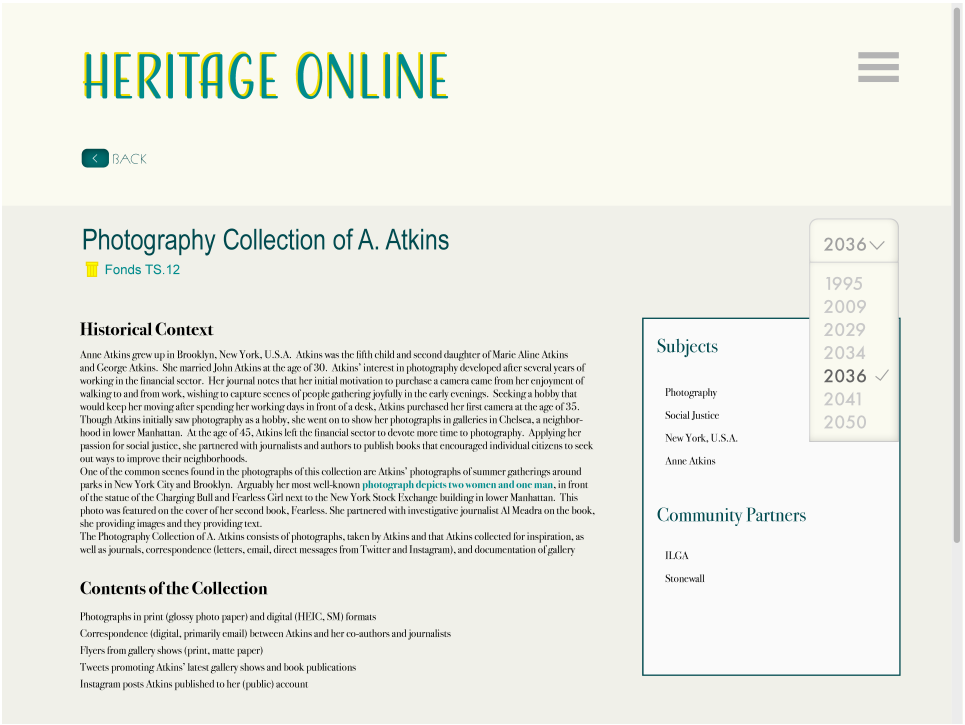


Figure 8.4: **Speculative cultural heritage catalog UI, record versions view.** Employees and visitors to the online catalog could view different versions of a specific heritage record.

Figure 8.5: **Speculative history textbook cover.** Using the dated and authored versions of records from the Heritage Online platform, language models would be trained to write historical narratives from distinct perspectives.

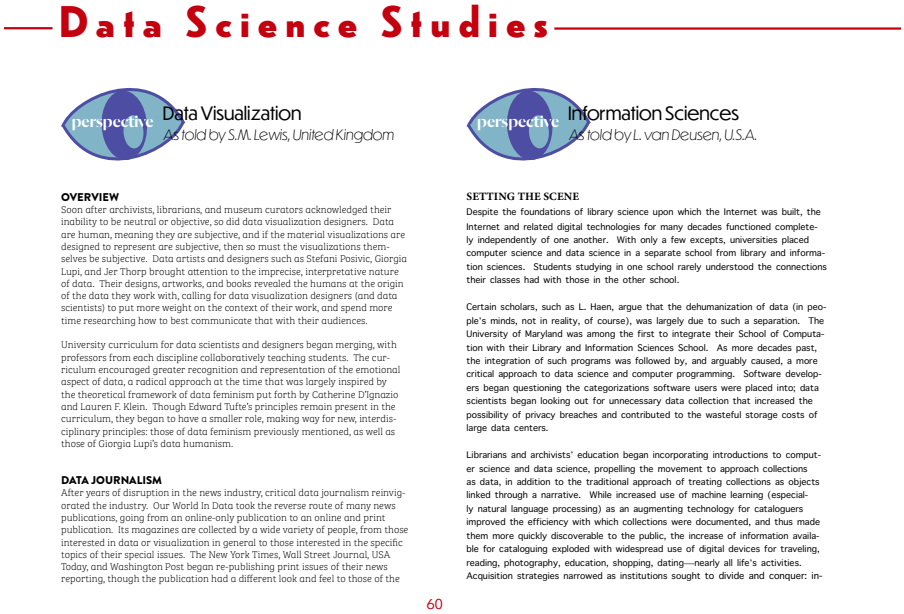


Figure 8.6: **Speculative history textbook page.** The history textbook would present historical narratives of the same event from different perspectives (written by different language models) side by side.

8.4.2 Collaboration with Stakeholders

Another promising direction for future work in GLAM and ML is in collaboration with ML systems and GLAM catalogs' stakeholders. GLAM catalogs were originally developed for those who worked at GLAM institutions to use in response to visitors' requests for records of cultural heritage (Library of Congress, 2017; Welsh, 2016). Thanks to digital technologies, however, GLAM have been able to provide users with direct access to their catalogs through online databases with a search engine interface (Welsh, 2016). GLAM have yet to undertake many user research studies, however, to understand how their visitors use search engine interfaces to their online catalogs and where they may be able to improve the user experience (Blouin and Rosenberg, 2011; Jaffe, 2020). Moreover, research on crowdsourcing and games for GLAM indicates that these interactive interfaces to heritage collections can increase public engagement with heritage (Flanagan and Carini, 2012; Flanagan et al., 2014; Ridge, 2013, 2016). In ML, collaboration with stakeholders is growing but is not yet well-established, with projects such as Masakhane (Nobata et al., 2016) and the work of Rodolfa et al. (2020) standing out as implementing participatory approaches to ML research from problem formulation through model evaluation.

Due to the large scale of GLAM catalogs and ML datasets, and the complexity of ML models' underlying mathematical calculations, I view data visualization for GLAM and ML as a valuable approach to facilitating collaborative data analysis, model development, and data and model evaluation (Havens, Bach, et al., 2022; Robertson et al., 2023; Spinner et al., 2019). Exploratory data visualizations are multifaceted, iterative, and open-ended, and thus represent and enable the exploration of multiple perspectives (Marchionini, 2006; White and Roth, 2009). Exploratory data visualizations also make analysis and evaluation processes accessible to a broader audience, because they rely on intuitive visual encodings, rather than specialized GLAM, History, ML, or Data Science knowledge (Havens, Bach, et al., 2022). Moreover, in addition to the visualization itself, the process of creating the visualization can be valuable (Hinrichs et al., 2017; Hinrichs et al., 2019). Data visualizations have been shown to facilitate an exchange of knowledge and understanding when created collaboratively with team members of distinct disciplinary

backgrounds (Hinrichs et al., 2017). As demonstrated in Chapter 7 with critiques of the *Occupation* label's application, engagement with stakeholders can bring unconscious biases to the forefront. The greater variety of perspectives included in a conversation, the more likely those biases can be surfaced and made explicit.

Chapter 9

Conclusion

In this chapter, I outline the contributions of my thesis (§9.1), summarize how I addressed my research questions (§9.2), reflect upon how I would now approach these questions differently (§9.3), and provide recommendations for researchers and practitioners in ML, GLAM, and History (§9.4).

9.1 Contributions

Working at the intersection of ML and GLAM, I developed five contributions for addressing social biases in data and ML models:

- The Bias-Aware Methodology (Chapter 4),
- The Taxonomy of Gendered and Gender Biased Language (Chapter 5),
- Datasets of archival documentation annotated for gender bias according to the Taxonomy (Chapter 5),
- Text classification models to automatically annotate gender biases in archival documentation (Chapter 6), and
- A human-centered approach to evaluating ML systems (Chapter 7).

I developed these contributions using interdisciplinary and participatory approaches, situating my work in a case study to account for the contextual nature of social biases. As I developed these contributions, I published and presented four papers at Natural Language Processing (NLP) and Digital Humanities venues:

- *Situated Data, Situated Systems: A Methodology to Engage with Power Relations in NLP* (Havens et al., 2020)
- *Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text* (Havens, Terras, et al., 2022)
- *Beyond Explanation: A Case for Exploratory Text Visualization of Non-Aggregated, Annotated Datasets* (Havens, Bach, et al., 2022)
- *Collaboration Across the Archival and Computational Sciences to Address Legacies of Gender Bias in Descriptive Metadata* (Havens et al., 2023; Appendix K)

I also presented in-progress work on my Ph.D. research at the:

- Association for Computers and Humanities Conference (Appendix I)
- Conference of the North American Chapter of the Association for Computational Linguistics's Student Research Workshop (Appendix J)

One additional publication reporting on aspects of my Ph.D. research, *Confronting Gender Biases in Heritage Catalogs: A Natural Language Processing Approach to Revisiting Descriptive Metadata*, will be published in the forthcoming *Routledge Handbook of Heritage and Gender* (Havens et al., 2024).

9.2 Summary

After describing the relevant research landscape (chapters 1-3), I began recounting the investigation of my first research question in Chapter 4, ***Can existing methods of identifying and categorizing gender biased language in NLP research be applied to archival documentation? Why or why not?***, while also contributing a new research methodology. I deemed Hitti et al.'s (2019) taxonomy for annotating gender biases in text applicable to my research context with the University of Edinburgh's Heritage Collections' (HC) archival documentation, revising and expanding it to create my Taxonomy of Gendered and Gender Biased Language (Chapter 5). Nonetheless, after conducting the literature review summarized in Chapter 4, I found that existing methods to identify and categorize gender bias in NLP (and, more broadly, ML) research could not be applied to my case study with the HC Archives.

Most NLP and ML work on gender bias aimed to eliminate, rather than identify and categorize, gender bias. I explained why the underlying assumption of this aim, that bias could be removed to create neutral data and models, was founded on overly simplistic conceptualizations of bias (§4.1.3, §4.2). I created the Bias-Aware Methodology to guide researchers towards a different approach to bias, an approach that focused on understanding rather than elimination, and that viewed ML systems as socio-technical rather than purely technical. My Bias-Aware Methodology acknowledges the inevitability and complexity of bias, and engages stakeholders throughout ML system creation to study and reflect upon manifestations of bias in the research context. The Bias-Aware Methodology served as my guiding research methodology for creating the remaining contributions of this thesis, and provides a guide for future work in ML on mitigating the harms from biases in ML systems.

In Chapter 5 I contributed the annotation taxonomy and annotated data used to create my classification models, further addressing my first research question and addressing my second research question: ***What types of gender bias are present in the language of archival metadata descriptions?*** With my Taxonomy of Gendered and Gender Biased Language, I proposed measuring gender biases at two levels: (1) at a high level, measuring the representation of genders through grammatically gendered terminology across a text corpus, and (2) at a low level, measuring the occurrence of stereotypical language and omitted information related to people's gender identity. My Taxonomy contains six annotation labels for types of gendered language, *Gendered Pronoun*, *Gendered Role*, *Feminine*, *Non-binary*, *Unknown*, and *Masculine*; and three for gender biased language, *Generalization*, *Stereotype*, and *Omission*. The Taxonomy also included a label for *Empowering*, to highlight the use of reclaimed terms, such as "queer," that indicate minoritized groups transforming a derogatory term into a term of pride and belonging (Bucholtz, 1999); as well as a label for *Occupation*, to enable further research on correlations between gender biases and work, such as the diachronic changes in stereotypical jobs for particular gender groups (Garg et al., 2018; Haines et al., 2016; M. Lewis and Lupyan, 2020). I applied the Taxonomy to HC Archives documentation, working with four hired annotators to label text spans with the Taxonomy's labels. The manual annotation process revealed that, for the subset of HC Archives documentation being annotated, text applicable the *Non-binary* and

Empowering labels was not present. Additionally, the *Generalization* label proved difficult to apply in practice, yielding the lowest Inter-Annotator Agreement (IAA) scores relative to the Taxonomy's other labels. The resulting annotated datasets are the first datasets built from GLAM documentation for the purpose of training ML models to detect social biases.

Chapter 6 contributed ML models, specifically text classification models, to annotate gender biased language. This chapter provided a quantitative approach to addressing the third research question, ***Can gender biased language in archival documentation be reliably annotated by domain experts to create data on which to train NLP classification models?*** The performance of the models relative to IAA scores suggest that certain types of gender biases in archival documentation can be reliably annotated to create model training data. Gender biased language in the form of *Omissions* and *Stereotypes* was classified reliably, while gender biased language in the form of *Generalization* was not. Classifying gendered language in the form of *Gendered Pronouns* and *Gendered Roles* was even more reliable, providing an approach to measuring the representation of different gender groups across an entire corpus of archival documentation. Classifying gendered language in the form of people's names (i.e. *Feminine*, *Masculine*, *Unknown*) was less reliable, though still more so than classifying text with the *Generalization* label. As in Chapter 5, I measured the reliability of annotations with standard NLP metrics, primarily F_1 score. The overall results of the model cascades suggested that gender biased language should be broken down into specific types of *Omissions* and *Stereotypes*, suggesting directions for future work on revising both the Taxonomy and model cascades.

Complementing the quantitative evaluations, Chapter 7 contributed a human-centered approach to qualitatively evaluating ML models' training data and performance. This chapter further addressed my third research question, ***Can gender biased language in archival documentation be reliably annotated by domain experts to create data on which to train NLP classification models?*** I partnered with Rachel Hosker, University Archivist and Research Collections Manager, to facilitate a workshop with the HC team to obtain their feedback on (1) the Taxonomy's application to HC Archives documentation and (2) the cross-collection measurements of gender bias that the classifiers from Chapter 6 enabled me to calculate. During the workshop

discussion, participants expressed uncertainty in the classifiers' capabilities, and spoke of their interest in comparing the classifiers' annotations of a metadata description to their own expectations of what should be annotated. Nonetheless, participants identified four areas of value that the classifiers could offer as a tool to support their existing workflows: (1) supporting collection reviews, (2) informing description-writing guidelines, (3) facilitating self-reflection, and (4) providing evidence of resource needs. These identified areas of value align with the Broussard (2019) and Irani's (2016) observation that human-ML collaboration outperforms fully autonomous ML systems, because only humans can reflect upon an ML model's interpretation of data relative to the ever-changing societal context in which that model functions.

Together, the contributions of this thesis illustrate an approach to recalibrating ML research. This recalibration acknowledges the inevitability of social biases and works to better understand them, so their consequences can be better anticipated and their potential harms more effectively mitigated.

9.3 Reflection

Reflecting upon my experience investigating the research questions of this thesis, I would now approach stakeholder collaboration, annotation, and human-centered evaluation differently. In an ideal situation, I would have collaborated not only with the HC team, but also with a community-run GLAM focused on representing minoritized gender communities.[1] Creating a dataset from these two types of GLAM institutions' documentation would likely lead to broader gender representation than the datasets in Chapter 5.

Regarding the dataset annotation, in an ideal situation I would engage in an iterative feedback process with members of the HC team. Rather than waiting until the manual annotation of HC Archives documentation was complete, I would seek feedback from members of the HC team throughout the annotation process. This iterative feedback would allow annotators (myself included) to adjust our approach to applying the Taxonomy of Gendered and Gender Biased Language (Chapter 5), informing our interpretation of the metadata descriptions, with the HC team members' experience with cultural heritage

[1] I did contact the Glasgow Women's Library during the first year of my research, but they were not interested in having a call to discuss a potential collaboration.

collections. This iterative feedback process would be particularly helpful for determining a suitable approach to annotating descriptions with the ***Person Name*** labels. The way in which *Unknown* was applied to most people's names in the descriptions prompted greatest debate during the human-centered evaluation of the Taxonomy and classifiers (Chapter 7).

Regarding the human-centered evaluation, I would have included an additional activity at the start of the workshop that asked the HC team members to apply the Taxonomy to the same descriptions used in *Activity 1*. This would provide an illustration of the variety of ways in which a person could interpret those descriptions and the Taxonomy's labels. Ideally, I would have also run an additional human-centered evaluation using an online platform that enabled the HC team to apply the models themselves. As was discussed during the workshop (Chapter 7), the HC team was interested in comparing their interpretation and the models' interpretation of gender biases in metadata descriptions of their choosing. The challenge with these expansions of the human-centered evaluation, as well as the iterative feedback process during manual annotation, is the additional time from HC team members that would be required. Future collaborative research projects should discuss suitable compensation and reasonable time commitments for asking GLAM domain experts to participate in the research.

9.4 Recommendations

Drawing on the process of creating and evaluating gender biased text classification models with archival documentation, I end this chapter with recommendations for researchers and practitioners in ML and GLAM, and for historians utilizing GLAM collections to inform narratives about the past.

9.4.1 Machine Learning

For ML researchers and practitioners, I recommend allocating additional time to project timelines for stakeholder collaboration and interdisciplinary engagement, echoing existing ML literature (Blodgett et al., 2020; Crawford, 2017; Devinney et al., 2022; Havens et al., 2020; Stańczak and Augenstein, 2021). More specifically, though, I recommend project timelines incorporate

stakeholder collaboration and interdisciplinary engagement end to end, from task formulation through to model evaluation. Biases may enter ML research and practice at any stage (Suresh and Guttag, 2021), so broader perspectives are needed throughout the entire ML system creation process. For guidance on collaboration methods to meaningfully and ethically partner with stakeholders, I recommend ML researchers and practitioners look to Aragon et al.'s *Human-Centered Data Science: An Introduction* (2022) for a summary of human-centered research methods relevant to data-driven work such as ML. For engaging with interdisciplinary theories and approaches, I recommend ML researchers and practitioners begin with a critical theory, such as critical discourse analysis (Bucholtz, 2003; Fairclough, 2003; van Leeuwen, 2009), and feminist theories, which emphasize the multiplicity and subjectivity of knowledge, such as D'Ignazio and Klein's *Data Feminism* (2020). For practical steps to collaborate with stakeholders and engage with interdisciplinary literature during ML system creation, I recommend applying the Bias-Aware Methodology (Chapter 4; Havens et al., 2020).

9.4.2 Galleries, Libraries, Archives, and Museums

I recommend GLAM researchers and practitioners continue to approach ML systems critically. The hype around ML systems exaggerates their capabilities (Bender et al., 2021; Posner, 2016; Raji et al., 2021; Verdegem, 2021). The GLAM community's approach to investigating potential benefits and risks of ML applications through small-scale projects (e.g. Baker, 2020) and case studies (e.g. Jaillant, 2022) demonstrates an important alternative to many corporations' approaches to deploying ML systems. In seeking collaborations with ML researchers and practitioners, GLAM researchers and practitioners should encourage a mutual exchange of approaches, methods, and theoretical frameworks. Just as ML systems offer benefits to GLAM, GLAM's approaches to classification and description offer benefits to ML (McGillivray et al., 2020). As Hauswedell et al. (2020) state, "The development of critical frameworks for scholarship with...digital cultural heritage materials that assist in helping researchers understand how and why they take the form that they do are paramount to ensuring that such tools can be used to study them rigorously and appropriately" (p. 142). Such critical frameworks are also paramount to

improving ML systems.

GLAM's aim to serve the public as memory institutions and information repositories (Library of Congress, 2017; Thomassen, 2002; Welsh, 2016; Welsh and Batley, 2009) position them to lead the way in ethical, socially just approaches to ML system creation and deployment (Jaillant, 2022; Noble, 2018). Moreover, thanks to GLAM practitioners' close work with "messy" heritage collection materials and data (Chapter 7), their understanding of the need for human-machine collaboration is informative for ML researchers and practitioners. I recommend national and otherwise hegemonic GLAM institutions partner with community-level GLAM, exploring and creating ML systems that improve existing cataloging practices. The perspectives documented in community GLAM catalogs need to be centered in GLAM and historical research and practice to complement the dominant perspectives in cultural heritage with alternative perspectives (Olson, 2001; Smith, 2006).

Additionally, Galleries and Museums can play a role in encouraging people to notice gender (as well as other social) biases, prompting critical reflection upon depictions of people that are both common and stereotypical (Hessel and Beard, 2022). However, the theory of reactance from Pyschology explains that encouraging people to think a particular way too forcefully can undermine the attempt (Brehm and Brehm, 1981). People tend to reject an idea proposed to them if they feel their freedom to come to their own conclusions is being threatened (ibid.). In the GLAM sector, the subtlety of an Embedded Design approach (Flanagan and Kaufman, 2017; Kaufman et al., 2021; Kaufman and Flanagan, 2015) could be replicated in exhibitions. For example, the Embedded Design practices of intermixing and obfuscating in game design could be adapted to an exhibition. Curators could intermix cultural heritage promoting stereotypical and antistereotypical perspectives throughout the exhibition galleries, aiming for a balanced representation of both perspectives; and curators could avoid explicitly stating the prosocial aim of an exhibition in descriptions on gallery walls and in the exhibition catalog. Bringing attention to datasets' biases poses a new challenge; collaboration across Human-Computer Interaction, Data Visualization, Art, and Design would likely be beneficial for adapting Embedded Design to ML dataset and model evaluations.

9.4.3 History

I recommend historians who study and report on the past be wary of relying on fully autonomous ML systems. While ML systems provide an approach to searching through the overwhelming amount of information online, these systems are trained on biased samples of data that influence their filtering and summarization. As stated for GLAM researchers and practitioners, historians should be aware that the hype around supposedly state-of-the-art models exaggerates ML systems' capabilities (Bender et al., 2021; Posner, 2016; Raji et al., 2021; Verdegem, 2021). Historians should engage with employees of GLAM who have familiarity with and expertise on cultural heritage collections and their documentation. GLAM employees can point historians towards areas of the past that have been under-documented or misrepresented, and are in need of further research. As with researchers and practitioners in ML and GLAM, I recommend historians look to critical studies that provide frameworks for questioning characterizations of power. Additionally, Graeber and Wengrow (2021) provide numerous examples of biased narratives of the past from Archaeology and Anthropology that deserve revisiting, and Beard's *Women and Power: A Manifesto* (2017) demonstrates how to begin questioning dominant narratives by centering a different perspective. I recommend historians work more closely with archivists, librarians, and curators to identify gaps in research on the past, and to write new narratives that can be incorporated into GLAM documentation.

I have proposed and demonstrated a recalibration of Machine Learning (ML) for social biases. Shifting ML from a top-down to a bottom-up approach, this thesis defines four recalibrations for ML research and practice: prioritizing quality over quantity, accuracy over efficiency, representativeness over convenience, and situated thinking over universal thinking. Through a case study with the University of Edinburgh's Heritage Collections team, and descriptive metadata from their Archives catalog, I demonstrated how to execute this recalibration. In so doing, I created five contributions: the Bias-Aware Methodology, the Taxonomy of Gendered and Gender Biased Language, datasets of archival documentation annotated for gender biases, gender biased text classification models, and a participatory approach to evaluating ML data and models. The Methodology provides practical guidance to ML researchers and practitioners for executing participatory and interdisciplinary work, a necessity for making the social biases of ML systems visible. The Taxonomy, datasets, models, and participatory evaluation provide ML, GLAM, and History researchers and practitioners with tools for studying gender biased language, to improve understandings of its variations and complexity. Together these contributions aim to empower minoritized gender communities, calling attention to their misrepresentation and omission so when future data and models are built, they incorporate those communities' perspectives.

Bibliography

Abercrombie, G., Cercas Curry, A., Pandya, M., & Rieser, V. (2021). Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 24–33. https://doi.org/10.18653/v1/2021.gebnlp-1.4

Adhikari, A., Ram, A., Tang, R., Hamilton, W. L., & Lin, J. (2020). Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification with DocBERT. *Proceedings of the 5th Workshop on Representation Learning for NLP*, 72–77. https://doi.org/10.18653/v1/2020.repl4nlp-1.10

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Rethinking Complex Neural Network Architectures for Document Classification. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4046–4051. https://doi.org/10.18653/v1/N19-1408

Adler, M. (2016). The Case for Taxonomic Reparations. *Knowledge Organization*, *43*(8), 630–640. https://doi.org/10.5771/0943-7444-2016-8-630

Adler, M. (2017). *Cruising the Library: Perversities in the Organization of Knowledge*. Fordham University Press. https://doi.org/10.2307/j.ctt1xhr79m

Adler, M., & Harper, L. M. (2018). Race and Ethnicity in Classification Systems: Teaching Knowledge Organization from a Social Justice Perspective. *Library Trends*, *67*(1), 52–73. https://doi.org/10.1353/lib.2018.0025

Agostinho, D., D'Ignazio, C., Ring, A., Thylstrup, N. B., & Veel, K. (2019). Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive.

Surveillance & Society, *17*(3/4), 422–441. https://doi.org/10.24908/ss.v17i3/4.12330

Alex, B., Grover, C., Shen, R., & Kabadjov, M. (2010). Agile Corpus Annotation in Practice: An Overview of Manual and Automatic Annotation of CVs. *Proceedings of the Fourth Linguistic Annotation Workshop*, 29–37. https://aclanthology.org/W10-1804

Ames, S., & Havens, L. (2022). Exploring National Library of Scotland Datasets with Jupyter Notebooks. *IFLA Journal*, *48*(1), 50–56. https://doi.org/10.1177/03400352211065484

Andriyansah, R., Bukhari, S. S., Jenckel, M., & Dengel, A. (2019). Using Balanced Training to Minimize Biased Classification. *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing - HIP '19*, 31–36. https://doi.org/10.1145/3352631.3352639

Andrus, M., Spitzer, E., Brown, J., & Xiang, A. (2021). What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 249–260. https://doi.org/10.1145/3442188.3445888

Angel, C. M., & Fuchs, C. (Eds.). (2018). *Organization, Representation and Description through the Digital Age: Information in Libraries, Archives and Museums*. Walter de Gruyter GmbH. https://doi.org/https://doi.org/10.1515/9783110337419

Antracoli, A. A., Berdini, A., Bolding, K., Charlton, F., & Ferrara, A. (2019). *Archives for Black Lives in Philadelphia: Anti-Racist Description Resources*. https://archivesforblacklives.files.wordpress.com/2019/10/ardr%5C_final.pdf

Aragon, C. R., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-Centered Data Science: An Introduction*. The MIT Press.

Aroyo, L., Lease, M., Paritosh, P., & Schaekermann, M. (2022). Data Excellence for AI: Why Should You Care? *Interactions*, *29*(2), 66–69. https://interactions.acm.org/archive/view/march-april-2022/data-excellence-for-ai

Aroyo, L., & Welty, C. (2015). Truth is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, *36*(1), 15–24. https://doi.org/10.1609/aimag.v36i1.2564

Artstein, R., & Poesio, M. (2008). Survey Article: Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596. https://doi.org/10.1162/coli.07-034-R2

Averkamp, S., Willette, K., Rudersdorf, A., & Meghan, F. (2021). *Humans-in-the-Loop: Recommendations Report*. https://labs.loc.gov/static/labs/work/reports/LC-Labs-Humans-in-the-Loop-Recommendations-Report-final.pdf

Awad, E., Dsouza, S., Kim, R., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine Experiment. *Nature*, *563*, 59–64. https://doi.org/10.1038/s41586-018-0637-6

Babcock, L., & Laschever, S. (2008). *Why Women Don't Ask: The High Cost of Avoiding Negotiation–and Positive Strategies for Change*. Piatkus.

Babcock, L., & Laschever, S. (2009). *Ask For It: How Women Can Use the Power of Negotiation to Get What They Really Want*. Bantam Dell.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., Rosenberg, M., Song, X., Stoica, A., Tiwary, S., & Wang, T. (2018). MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *Computing Research Repository*. https://doi.org/10.48550/arXiv.1611.09268

Baker, J. (2020). A Machine that Writes like Mary Dorothy George. https://cradledincaricature.com/2020/06/18/mary-dorothy-george/

Baker, J., & Salway, A. (2020). Curatorial Labour, Voice and Legacy: Mary Dorothy George and the Catalogue of Political and Personal Satires, 1930–54. *Historical Research: The Bulletin of the Institute of Historical Research*, *93*(262), 769–785.

Basile, V. (2022). The Perspectivist Data Manifesto [Online; accessed March 21, 2022]. https://pdai.info/

Basile, V., Cabitza, F., Campagner, A., & Fell, M. (2021). Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(6), 6860–6868. https://doi.org/10.1609/aaai.v37i6.25840

Basta, C., Costa-jussà, M. R., & Casas, N. (2020). Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings. *Neural Computing & Applications*, *33*(8), 3371–3384.

Beard, M. (2017). *Women & Power: A Manifesto*. Profile Books Ltd.

Beelen, K., Lawrence, J., McDonough, K., Westerling, K., & Wilson, D. C. S. (2023). The 'Environmental Scan' at Work: Radical Contextualisation of Newspaper Collections for New Historical Research. In A. Baillot, T. Tasovac, W. Scholger, & G. Vogeler (Eds.), *Digital Humanities 2023: Book of Abstracts* (pp. 408–409). Austrian Centre for Digital Humanities, University of Graz. https://doi.org/10.5281/zenodo.7961822

Beelen, K., Lawrence, J., Wilson, D. C. S., & Beavan, D. (2022). Bias and Representativeness in Digitized newspaper Collections: Introducing the Environmental Scan. *Digital Scholarship in the Humanities*, *38*(1), 1–22. https://doi.org/10.1093/llc/fqac037

Beelen, K., Nanni, F., Coll Ardanuy, M., Hosseini, K., Tolfo, G., & McGillivray, B. (2021). When Time Makes Sense: A Historically-Aware Approach to Targeted Sense Disambiguation. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2751–2761. https://doi.org/10.18653/v1/2021.findings-acl.243

Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, *6*, 587–604. https://doi.org/10.1162/tacl_a_00041

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity. https://ebookcentral.proquest.com/lib/ed/detail.action?docID=5820427

Bennett, C. L., & Keyes, O. (2020). What is the Point of Fairness?: Disability, AI and the Complexity of Justice. *ACM SIGACCESS Accessibility and Computing*, (125). https://doi.org/10.1145/3386296.3386301

Berardi, G., Esuli, A., Gordea, S., Marcheggiani, D., & Sebastiani, F. (2012). Metadata Enrichment Services for the Europeana Digital Library. In P. Zaphiris, G. Buchanan, E. Rasmussen, & F. Loizides (Eds.), *Theory and Practice of Digital Libraries* (pp. 508–511). Springer. https://doi.org/10.1007/978-3-642-33290-6_61

Berry, D. (2020). Conscious Editing: Enhancing Diversity and Discovery. https://youtu.be/XGCTtDgNty4

BigScience Workshop. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2211.05100

Bird, C., Ungless, E. L., & Kasirzadeh, A. (2023). Typology of Risks of Generative Text-to-Image Models. *AIES '23: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 396–410. https://doi.org/10.1145/3600211.3604722

Bird, S., Klein, E., & Loper, E. (2019). *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit* [Online; accessed 6-June-2023]. https://www.nltk.org/book

Birhane, A. (2020). Algorithmic Colonization of Africa. *SCRIPT-ed*, *17*(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184. https://doi.org/10.1145/3531146.3533083

Birhane, A., Kasirzadeh, A., Leslie, D., & Wachter, S. (2023). Science in the Age of Large Language Models. *Nature Reviews Physics*, *5*, 277–280. https://doi.org/10.1038/s42254-023-00581-4

Birhane, A., & Prabhu, V. U. (2021). Large Image Datasets: A Pyrrhic Win for Computer Vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

Bjorkman, B. M. (2017). Singular They and the Syntactic Representation of Gender in English. *Glossa (London)*, *2*(1).

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Blodgett, S. L., Green, L., & O'Connor, B. (2016). Demographic Dialectal Variation in Social Media: A Case Study of African-American English.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 1119–1130. https://doi.org/10.18653/v1/D16-1120

Blodgett, S. L., Lopez, G., Olteanu, A., Sim, R., & Wallach, H. (2021). Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1004–1015. https://doi.org/10.18653/v1/2021.acl-long.81

Blodgett, S. L., Madaio, M., O'Connor, B., Wallach, H., & Yang, Q. (Eds.). (2021). *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*. Association for Computational Linguistics. https://aclanthology.org/2021.hcinlp-1.0

Blouin, F. X., Jr, & Rosenberg, W. G. (2011). 8 The Archivist as Activist in the Production of (Historical) Knowledge. *Processing the Past: Contesting Authorities in History and the Archives* (Online, pp. 140–160). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199740543.001.0001

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 4356–4364. https://doi.org/10.18653/v1/2020.acl-main.485

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2108.07258

Bordia, S., & Bowman, S. R. (2019). Identifying and Reducing Gender Bias in Word-Level Language Models. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 7–15. https://doi.org/10.18653/v1/N19-3002

Borkan, D., Dixon, L., Sorensen, J., Thain, N., & Vasserman, L. (2019). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. *WWW '19: Companion Proceedings of The 2019 World Wide Web Conference*, 491–500. https://doi.org/10.1145/3308560.3317593

Bourgeois, D., Rappaz, J., & Aberer, K. (2018). Selection Bias in News Coverage: Learning it, Fighting it. *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, 535–543. https://doi.org/10.1145/3184558.3188724

Bowker, G. C. (2008). *Memory Practices in the Sciences*. The MIT Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*. The MIT Press.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A Large Annotated Corpus for Learning Natural Language Inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. https://doi.org/10.18653/v1/D15-1075

Brehm, S. S., & Brehm, J. W. (1981). *Psychological Reactance: A Theory of Freedom and Control*. Academic Press.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Broussard, M. (2019). Letting Go of Technochauvinism. *Public Books*, *Technology*(Co-Opting AI). https://www.publicbooks.org/letting-go-of-technochauvinism/

Broussard, M. (2023). *More Than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*. The MIT Press.

Brown, S. (2023). Why It's Time For "Data-Centric Artificial Intelligence". *MIT Sloan: Ideas Made to Matter*. https://mitsloan.mit.edu/ideas-made-to-matter/why-its-time-data-centric-artificial-intelligence

Bucholtz, M. (1999). Gender. *Journal of Linguistic Anthropology*, *9*(1-2), 80–83.

Bucholtz, M. (2003). Theories of Discourse as Theories of Gender: Discourse Analysis in Language and Gender Studies. *The Handbook of Language and Gender*, 43–68. https://doi.org/10.1002/9780470756942.ch2

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 1–15). PMLR. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burns, J., Cronquist, M., Huang, J., Murphy, D., Rawson, K., Schaefer, B., Simons, J., Watson, B. M., Williams, A., & Collective, T. T. M. (2022). Metadata Best Practices for Trans and Gender Diverse Resources. https://doi.org/10.5281/zenodo.6829167

Butler, J. (1990). *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-Like Biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Cao, Y. T., & Daumé, I., Hal. (2021). Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, *47*(3), 615–661. https://doi.org/10.1162/coli_a_00413

Cao, Y. T., & Daumé III, H. (2020). Toward Gender-Inclusive Coreference Resolution. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4568–4595). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.418

Carroll, L. (2019). *Alice's Adventures in Wonderland & Through the Looking-Glass*. Harper Design.

Caselli, T., Cibin, R., Conforti, C., Encinas, E., & Teli, M. (2021). Guiding Principles for Participatory Design-inspired Natural Language Processing. *Proceedings of the 1st Workshop on NLP for Positive Impact*, 27–35. https://doi.org/10.18653/v1/2021.nlp4posimpact-1.4

Casey, A., Bennett, M., Tobin, R., Grover, C., Walker, I., Engelmann, L., & Alex, B. (2021). Plague Dot Text: Text Mining and Annotation of Outbreak Reports of the Third Plague Pandemic (1894-1952). *Journal of Data Mining & Digital Humanities 2021*, *HistoInformatics*. https://doi.org/10.46298/jdmdh.6071

Caswell, M. (2016). 'The Archive' Is Not an Archives: On Acknowledging the Intellectual Contributions of Archival Studies. *Reconstruction: Studies in Contemporary Culture*, *16*(1). https://escholarship.org/uc/item/7bn4v1fk

Caswell, M. (2022). Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal. *Radical Empathy in Archival Practice* (Special issue). Journal of Critical Library; Information Studies. https://core.ac.uk/display/234708097?utm%5C_source=pdf%5C&utm%5C_medium=banner%5C&utm%5C_campaign=pdf-decoration-v1

Caswell, M., & Cifor, M. (2016). From Human Rights to Feminist Ethics: Radical Empathy in the Archives. *Archivaria*, 23–43. https://archivaria.ca/index.php/archivaria/article/view/13557

Caswell, M., & Cifor, M. (2019). Neither a Beginning Nor an End: Applying an Ethics of Care to Digital Archival Collections. *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites* (pp. 159–168). Routledge. https://doi.org/10.4324/9780429506765-11

Champion, E. M. (2016). Digital Humanities is Text Heavy, Visualization Light and Simulation Poor. *Digital Scholarship in the Humanities*, fqw053. https://doi.org/10.1093/llc/fqw053

Ciora, C., Iren, N., & Alikhani, M. (2021). Examining Covert Gender Bias: A Case Study in Turkish and English Machine Translation Models. *Proceedings of the 14th International Conference on Natural Language Generation*, 55–63. https://doi.org/https://doi.org/10.18653/v1/2021.inlg-1.7

Coffey, D. (2021). Māori are trying to save their language from Big Tech [Online; accessed 1-September-2023]. *Wired UK*. https://www.wired.co.uk/article/maori-language-tech

Coll Ardanuy, M., Nanni, F., Beelen, K., Hosseini, K., Ahnert, R., Lawrence, J., McDonough, K., Tolfo, G., Wilson, D. C., & McGillivray, B. (2020). Living

Machines: A Study of Atypical Animacy. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2005.11140

Collections Trust. (2023). Decolonising the Database [Online; accessed 16-August-2023]. https://collectionstrust.org.uk/decolonisation/

Combahee River Collective. (1979). The Combahee River Collective Statement. In Z. R. Eisenstein (Ed.), *Capitalist Patriarchy and the Case for Socialist Feminism*. Monthly Review Pr.

Cook, T. (2011). 'We Are What We Keep; We Keep What We Are': Archival Appraisal Past, Present and Future. *Journal of the Society of Archivists*, *32*(2), 173–189. https://doi.org/10.1080/00379816.2011.619688

Corbett, M. (1990). Clearing the Air: Some Thoughts on Gender-Neutral Writing. *IEEE Transactions on Professional Communication*, *33*(1), 2–6. https://doi.org/10.1109/47.49063

Cordell, R. (2020). Machine Learning + Libraries: A Report on the State of the Field. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig

Costanza-Chock, S. (2018). Design Justice, A.I. and Escape from the Matrix of Domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426

Cramer, J. S. (2010a). 1 - Introduction. *Logit Models from Economics and Other Fields*. Cambridge University Press. https://doi.org/10.1017/CBO9780511615412

Cramer, J. S. (2010b). 9 - The Origins and Development of the Logit Model. *Logit Models from Economics and Other Fields*. Cambridge University Press. https://doi.org/10.1017/CBO9780511615412

Crammer, K., Dekel, O., Keshet, J., Shalev-Schwartz, S., & Singer, Y. (2006). Online Passive-Aggressive Algorithms. *Journal of Machine Learning Research*, *7*(19), 551–585. https://jmlr.org/papers/v7/crammer06a.html

Crammer, K., Kulesza, A., & Dredze, M. (2013). Adaptive Regularization of Weight Vectors. *Journal of Machine Learning Research*, *91*, 155–187. https://doi.org/10.1007/s10994-013-5327-x

Crawford, K. (2017). The Trouble with Bias [Online; accessed 10-July-2020]. https://www.youtube.com/watch?v=fMym%5C_BKWQzk

Crawford, K. (2021). *Atlas of AI: Power, Politics and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crawford, K., & Paglen, T. (2019). *Excavating AI: The Politics of Training Sets for Machine Learning*. The AI Now Institute, NYU. https://excavating.ai

Crenshaw, K. (1989). Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, *1989*(1), 139–167. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics and Violence against Women of Color. *Stanford Law Review*, *43*(6), 60. https://doi.org/10.2307/1229039

Crumley, C. L. (1995). Heterarchy and the Analysis of Complex Societies. *Archaeological Papers of the American Anthropological Association*, *6*(1), 1–5. https://doi.org/10.1525/ap3a.1995.6.1.1

Davani, A. M., Dıaz, M., & Prabhakaran, V. (2022). Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, *10*, 92–110.

De Bonis, M., Minutella, F., Falchi, F., & Manghi, P. (2023). A Graph Neural Network Approach for Evaluating Correctness of Groups of Duplicates. In O. Alonso, H. Cousijn, G. Silvello, M. Marrero, C. Teixeira Lopes, & S. Marchesin (Eds.), *Linking Theory and Practice of Digital Libraries* (pp. 207–219). Springer. https://doi.org/10.1007/978-3-031-43849-3_18

De Toni, F., Akiki, C., de la Rosa, J., Fourrier, C., Manjavacas, E., Schweter, S., & van Strien, D. (2022). Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2204.05211

Deisenroth, M., Peter, A. A. F., & Ong, C. S. O. (2020). 1 Introduction and Motivation. *Mathematics for Machine Learning*. https://doi.org/10.1017/9781108679930

de Jong, S., & Koevoets, S. (2013). Introduction. *Teaching Gender with Libraries and Archives The Power of Information*.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., & Scheuerman, M. K. (2020). Bringing the People Back In: Contesting Benchmark

Machine Learning Datasets. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2007.07399

Department of Health, Education, and Welfare. (1979). Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. *Federal Register*, *44*(76), 23192–23197-.

de Vassimon Manela, D., Errington, D., Fisher, T., van Breugel, B., & Minervini, P. (2021). Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2232–2242. https://doi.org/10.18653/v1/2021.eacl-main.190

Devinney, H., Björklund, J., & Björklund, H. (2022). Theories of "Gender" in NLP Bias Research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability and Transparency (FAccT '22)*. https://doi.org/10.1145/3531146.3534627

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computing Research Repository*. https://doi.org/10.48550/arXiv.1810.04805

Diao, J., & Cao, H. (2016). Chronology in Cataloging Chinese Archaeological Reports: An Investigation of Cultural Bias in the Library of Congress Classification. *Cataloging & Classification Quarterly*, *54*(4), 244–262. https://doi.org/10.1080/01639374.2016.1150931

Diaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018). Addressing Age-Related Bias in Sentiment Analysis. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–14. https://doi.org/10.1145/3173574.3173986

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. The MIT Press. https://mitpressonpubpub.mitpress.mit.edu/data-feminism

Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2020). Queens are Powerful Too: Mitigating Gender Bias in Dialogue Generation. *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), 8173–8188. https://doi.org/10.18653/v1/2020.emnlp-main.656

Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., & Williams, A. (2020). Multi-Dimensional Gender Bias Classification. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 314–331. https://www.aclweb.org/anthology/2020.emnlp-main.23

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018). Measuring and Mitigating Unintended Bias in Text Classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES '18*, 67–73. https://doi.org/10.1145/3278721.3278729

Dotan, R., & Milli, S. (2020). Value-Laden Disciplinary Shifts in Machine Learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. https://doi.org/arxiv:1912.01172v1

Dotan, T. (2023). Microsoft Adds the Tech Behind ChatGPT to Its Business Software. *Wall Street Journal*, *Tech*. https://www.wsj.com/articles/microsoft-blends-the-tech-behind-chatgpt-into-its-business-software-c79f0e8d

Doughman, J., Abu Salem, F., & Elbassuoni, S. (2020). Time-Aware Word Embeddings for Three Lebanese News Archives. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4717–4725. https://aclanthology.org/2020.lrec-1.580

Doughman, J., Khreich, W., El Gharib, M., Wiss, M., & Berjawi, Z. (2021). Gender Bias in Text: Origin, Taxonomy, and Implications. *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, 34–44. https://doi.org/10.18653/v1/2021.gebnlp-1.5

Drabinski, E. (2013). Queering the Catalog: Queer Theory and the Politics of Correction. *The Library Quarterly*, *83*(2), 94–111. https://doi.org/10.1086/669547

Drenthe, G., & van der Sommen, M. (1998). *European Women's Thesaurus: A Structured List of Descriptors for Indexing and Retrieving Information in the Field of the Position of Women and Women's Studies* (M. Boere, Ed.; J. Vaughan, Trans.; 1st version). International Information Centre; Archives for the Women's Movement (IIAV).

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1383–1392. https://doi.org/10.18653/v1/P18-1128

Drucker, J. (2021). Visualization. In N. B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain Archives* (pp. 561–568). The MIT Press. https://doi.org/10.7551/mitpress/12236.003.0061

Duff, W. M., & Harris, V. (2002). Stories and Names: Archival Description as Narrating Records and Constructing Meanings. *Archival Science*, *2*(3–4), 263–85. https://doi.org/10.1007/BF02435625

Duncan, C. (2005). 1 The Art Museum as Ritual. *Civilizing Rituals: Inside Public Art Museums* (1st ed., pp. 7–20). Routledge. https://doi.org/10.4324/9780203978719

Dunne, A., & Raby, F. (2013). *Speculative Everything: Design, Fiction, and Social Dreaming*. The MIT Press.

Dunsire, G. (2018). Ethical Issues in Catalogue Content Standards. *Catalogue & Index*, *191*, 11–15.

Dunsire, G., & Willer, M. (2014). The Local in the Global: Universal Bibliographic Control from the Bottom Up. *Cataloguing with Bibliography, Classification & Indexing and UNIMARC Strategic Programme, World Library and Information Congress*. https://library.ifla.org/id/eprint/817/1/086-dunsire-en.pdf

Eisenstein, J. (2018). *Natural Language Processing* [Accessed 6-June-2023]. Open access edition. https://cseweb.ucsd.edu/~nnakashole/teaching/eisenstein-nov18.pdf

Elejalde, E., Ferres, L., & Herder, E. (2017). The Nature of Real and Perceived Bias in Chilean Media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 95–104. https://doi.org/10.1145/3078714.3078724

Eubanks, V. (2017). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (First Edition). St. Martin's Press.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. Routledge.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, *9*, 1871–1874. https://doi.org/10.5555/1390681.1442794

Filgueira, R., Grover, C., Karaiskos, V., Alex, B., Van Eyndhoven, S., Gotthard, L., & Terras, M. (2021). Extending defoe for the Efficient Analysis of Historical Texts at Scale. *2021 IEEE 17th International Conference on eScience (eScience)*, 21–29. https://doi.org/10.1109/eScience51609.2021.00012

Filgueira, R., Jackson, M., Roubickova, A., Krause, A., Ahnert, R., Hauswedell, T., Nyhan, J., Beavan, D., Hobson, T., Coll Ardanuy, M., Colavizza, G., Hetherington, J., & Terras, M. (2019). defoe: A Spark-Based Toolbox for Analysing Digital Historical Textual Data. *2019 15th International Conference on eScience (eScience)*, 235–242. https://doi.org/10.1109/eScience.2019.00033

Firth, J. (1957). A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis*, 10–32. https://cir.nii.ac.jp/crid/1574231874045325568

Flanagan, M. (2009). *Critical Play: Radical Game Design*. The MIT Press.

Flanagan, M. (2017). [help me know the truth]. https://maryflanagan.com/help-me-know-the-truth

Flanagan, M., & Carini, P. (2012). How Games Can Help Us Access and Understand Archival Images. *The American Archivist*, *75*(2), 514–537. https://www.jstor.org/stable/43489634

Flanagan, M., & Jakobsson, M. (2023). *Playing Oppression: The Legacy of Conquest and Empire in Colonialist Board Games*. The MIT Press.

Flanagan, M., & Kaufman, G. (2017). Shifting Implicit Biases with Gaming Using Psychology: The Embedded Design Approach. In Y. B. Kafai, G. T. Richard, & B. M. Tynes (Eds.), *Diversifying Barbie and Mortal Kombat: Intersectional Perspectives and Inclusive Designs in Gaming* (pp. 219–233). ETC Press.

Flanagan, M., Punjasthitkul, S., Seidman, M., Kaufman, G., & Carini, P. (2014). Citizen Archivists at Play: Game Design for Gathering Metadata for Cultural Heritage Institutions. *DiGRA &#3913 - Proceedings of the 2013 DiGRA International Conference: DeFragging Game Studies*.

https://www.digra.org/wp-content/uploads/digital-library/paper_418.compressed.pdf

Flinn, A., & Alexander, B. (2015). "Humanizing an Inevitability Political Craft:" Introduction to the Special Issue on Archiving Activism and Activist Archiving. *Archival Science*, *15*, 329–335. https://doi.org/10.1007/s10502-015-9260-6

Flinn, A., Stevens, M., & Shepherd, E. (2009). Whose Memories, Whose Archives? Independent Community Archives, Autonomy and the Mainstream. *Archival Science*, *9*(1–2), 71–86. https://doi.org/10.1007/s10502-009-9105-2

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Furner, J. (2007). Dewey Deracialized: A Critical Race-Theoretic Perspective. *Knowledge Organization*, *34*(3), 144–168. https://doi.org/10.5771/0943-7444-2007-3-144

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635–E3644. https://doi.org/10.1073/pnas.1720347115

Garimella, A., Banea, C., Hovy, D., & Mihalcea, R. (2019). Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3493–3498. https://doi.org/10.18653/v1/P19-1339

Garnerin, M., Rossato, S., & Besacier, L. (2020). Gender Representation in Open Source Speech Resources. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6599–6605. https://aclanthology.org/2020.lrec-1.813

Gaver, W. W., Beaver, J., & Benford, S. (2003). Ambiguity As a Resource for Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 233–240. https://doi.org/10.1145/642611.642653

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., & Crawford, K. (2018). Datasheets for Datasets. *Computing Research Repository*. https://doi.org/10.48550/arXiv.1803.09010

Gee, J. P., & Handford, M. (2014). Introduction. In J. P. Gee & M. Handford (Eds.), *The Routledge Handbook of Discourse Analysis*. Routledge.

Geraci, N. (2019). Programmatic Approaches to Bias in Descriptive Metadata [Online; accessed 28-May-2020]. https://www.youtube.%20com/watch?v=7mdMtukvtxc%5C%5C%5C&list=PLw-ls5JXzeNYcmotU2peVxu27nH2qIrV6%5C%5C%5C&%20index=1

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 1–13). The MIT Press.

Glaser, B. G., & Strauss, A. L. (1980). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.

Gleibs, I. H. (2017). Are All "Research Fields" Equal? Rethinking Practice for the Use of Data From Crowdsourcing Market Places. *Behavior Research Methods*, *49*(4), 1333–1342. https://doi.org/10.3758/s13428-016-0789-y

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., & Lopez, A. (2021). Intrinsic Bias Metrics Do Not Correlate with Application Bias. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1926–1940. https://doi.org/10.18653/v1/2021.acl-long.150

Goldwater, S. (2021). Lecture notes from Accelerated Natural Language Processing 2020-2021[SEM1.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. https://doi.org/10.18653/v1/N19-1061

Goree, S., & Crandall, D. (2023). Situated Cameras, Situated Knowledges: Towards an Egocentric Epistemology for Computer Vision. https://doi.org/10.48550/arxiv.2307.00064

Goree, S., Khoo, W., & Crandall, D. J. (2023). Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models. *AAAI Conference on Artificial Intelligence*.

Gov.uk. (2022). National Minimum Wage and National Living Wage rates. https://www.gov.uk/national-minimum-wage-rates

Graeber, D., & Wengrow, D. (2021). *The Dawn of Everything: A New History of Humanity*. Penguin Books Ltd.

Graeff, E. (2020). The Responsibility to Not Design and the Need for Citizen Professionalism. *Tech Otherwise*. https://doi.org/10.21428/93b2c832.c8387014

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

Greenburg, J., Sprugin, K., Crystal, A., Cronquist, M., & Wilson, A. (2005). Final Report for the AMeGA (Automatic Metadata Generation Applica-tions) Project. https://www.loc.gov/catdir/bibcontrol/lc%5C_amega%5C_final%5C_report.pdf

Guo, M., Hwa, R., Lin, Y.-R., & Chung, W.-T. (2020). Inflating Topic Relevance with Ideology: A Case Study of Political Ideology Bias in Social Topic Detection Models. *Proceedings of the 28th International Conference on Computational Linguistics*, 4873–4885. https://doi.org/10.18653/v1/2020.coling-main.428

Hagey, K., & Cherney, M. (2023). ChatGPT Owner Vows to Improve Its AI Tools After Sam Altman's World Tour. *Wall Street Journal*, *WSJ News Exclusive*. https://www.wsj.com/articles/chatgpt-owner-vows-to-improve-its-ai-tools-after-sam-altmans-world-tour-e0466dfd

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The Times They Are a-Changing . . . or Are They Not? A Comparison of Gender Stereotypes, 1983-2014. *Psychology of Women Quarterly*, *40*(3), 353–363. https://doi.org/10.1177/0361684316634081

Hanson, V. L., Cavender, A., & Trewin, S. (2015). Writing About Accessibility. *Interactions*, *22*(6), 62–65. https://doi.org/https://dx.doi.org/10.1145/2828432

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575. https://doi.org/10.2307/3178066

Harding, S. (1995). "Strong objectivity": A Response to the New Objectivity Question. *Synthese*, *104*(3). https://doi.org/10.1007/BF01064504

Harper, C. A. (2016). Metadata Analytics, Visualization and Optimization: Experiments in Statistical Analysis of the Digital Public Library of America (DPLA). *Code4Lib Journal*, (33). https://journal.code4lib.org/articles/11752

Harris, V. (2002). The Archival Sliver: Power, Memory, and Archives in South Africa. *Archival Science*, *2*(1–2), 63–86. https://doi.org/10.1007/BF02435631

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hauswedell, T., Nyhan, J., Beals, M. H., Terras, M., & Bell, E. (2020). Of Global Reach Yet of Situated Contexts: An Examination of the Implicit and Explicit Selection Criteria that Shape Digital Archives of Historical Newspapers. *Archival Science*, *20*(2), 139–165. https://doi.org/10.1007/s10502-020-09332-1

Havens, L. (2021). An Information Space with More Than a Search Bar for Discovery. *ACH2021 Conference*. https://lucyhavens.com/more-than-a-search-bar-for-discovery

Havens, L., Bach, B., Terras, M., & Alex, B. (2022). Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP LREC2022*, 73–82. https://aclanthology.org/2022.nlperspectives-1.10

Havens, L., Hosker, R., Bach, B., Terras, M., & Alex, B. (2023). Collaboration Across the Archival and Computational Sciences to Address Legacies of Gender Bias in Descriptive Metadata. *Digital Humanities 2023: Book of Abstracts*, 267–270. https://zenodo.org/record/7961822

Havens, L., Terras, M., Bach, B., & Alex, B. (2024). *Routledge Handbook of Heritage and Gender*. Routledge.

Havens, L., Terras, M., Bach, B., & Alex, B. (2020). Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 107–124. https://aclanthology.org/2020.gebnlp-1.10

Havens, L., Terras, M., Bach, B., & Alex, B. (2022). Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated

Datasets of British English Text. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 30–57. https://doi.org/10.18653/v1/2022.gebnlp-1.4

Heritage Collections. (2018). Collection: Papers and Artwork of Yolanda Sonnabend Relating to Her Collaboration with C.H. Waddington [Online; accessed 19 May 2022]. https://archives.collections.ed.ac.uk/repositories/2/resources/84761

Hessel, K. (2023a). 'Blatant Sexism:' Why is a Great Painter Who Lived to 101 Still Defined by a Man She Left in the 1950s? *The Guardian*. https://www.theguardian.com/artanddesign/2023/jun/12/blatant-sexism-great-painter-francoise-gilot

Hessel, K. (2023b). *The Story of Art Without Men* (First American edition). W.W. Norton & Company, Inc.

Hessel, K. (2023c). Why Do We Still Define Female Artists as Wives, Friends and Muses? *The Guardian*. https://www.theguardian.com/artanddesign/2023/feb/20/why-do-we-still-define-female-artists-as-wives-friends-and-muses

Hessel, K., & Beard, M. (2022). Mary Beard on Classical Women (100th Episode Special!)

Hicks, M. (2018). Introduction: Britain's Computer "Revolution". *Programmed Inequality: How Britain Discarded Women Technologists and Lost its Edge in Computing* (pp. 1–17). The MIT Press.

Hicks, M. (2021). Sexism is a Feature, Not a Bug. *Your Computer is on Fire* (pp. 135–158). The MIT Press.

Hill Collins, P. (2000). *Black Feminist Thought: Knowledge, Consciousness and the Politics of Empowerment*. Routledge.

Hinrichs, U., Alex, B., Clifford, J., Watson, A., Quigley, A., Klein, E., & Coates, C. M. (2015). Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqv046

Hinrichs, U., El-Assady, M., Bradley, A. J., Forlini, S., & Collins, C. (2017). Risk the Drift! Stretching Disciplinary Boundaries through Critical Collaborations between the Humanities and Visualization. *Proceedings of the 2nd Workshop on Visualization for the Digital Humanities*.

Hinrichs, U., Forlini, S., & Moynihan, B. (2019). In Defense of Sandcastles: Research Thinking Through Visualization in Digital Humanities. *Digital Scholarship in the Humanities*, *34*(Supplement_1), i80–i99. https://doi.org/10.1093/llc/fqy051

Hitti, Y., Jang, E., Moreno, I., & Pelletier, C. (2019). Proposed Taxonomy for Gender Bias in Text; A Filtering Methodology for the Gender Generalization Subtype. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 8–17. https://doi.org/10.18653/v1/W19-3802

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). Training Compute-Optimal Large Language Models. *Computing Research Repository*. https://doi.org/10.48550/arXiv.2203.15556

Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3290605.3300830

Hosseini, K., Wilson, D. C. S., Beelen, K., & McDonough, K. (2022). MapReader: A Computer Vision Pipeline for the Semantic Exploration of Maps at Scale. *Proceedings of the 6th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, 8–19. https://doi.org/10.1145/3557919.3565812

Hovy, D., & Prabhumoye, S. (2021). Five Sources of Bias in Natural Language Processing. *Language and Linguistics Compass*, *15*(8). https://doi.org/10.1111/lnc3.12432

Hube, C. (2017). Bias in Wikipedia. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*, 717–721. https://doi.org/10.1145/3041021.3053375

Hube, C., & Fetahu, B. (2019). Neural Based Statement Classification for Biased Language. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 195–203. https://doi.org/10.1145/3289600.3291018

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*, *Artificial Intelligence (AI)*. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Husse, S., & Spitz, A. (2022). Mind Your Bias: A Critical Review of Bias Detection Methods for Contextual Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4212–4234. https://aclanthology.org/2022.findings-emnlp.311

Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social Biases in NLP Models as Barriers for Persons with Disabilities. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. https://doi.org/10.18653/v1/2020.acl-main.487

Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., Barnes, P., & Mitchell, M. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575. https://doi.org/10.1145/3442188.3445918

Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., Hansen, H., Roussel, N., & Eiderbäck, B. (2003). Technology Probes: Inspiring Design For and With Families. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 17–24. https://doi.org/10.1145/642611.642616

Iacovino, L. (2010). Rethinking Archival, Ethical and Legal Frameworks for Records of Indigenous Australian Communities: A Participant Relationship Model of Rights and Responsibilities. *Archival Science*, *10*(4), 353–372. https://doi.org/10.1007/s10502-010-9120-3

ICA. (2011). ISAD(G): General International Standard Archival Description - Second Edition. https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

ICA. (2021). *Records in Contexts: Introduction to Archival Description*.

Irani, L. (2015). Difference and Dependence Among Digital Workers: The Case of Amazon Mechanical Turk. *South Atlantic Quarterly*, *114*(1), 225–234. https://doi.org/10.1215/00382876-2831665

Irani, L. (2016). The Hidden Faces of Automation. *XRDS: Crossroads, The ACM Magazine for Students*, *23*(2), 34–37. https://doi.org/10.1145/3014390

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency*, 375–385. https://doi.org/10.1145/3442188.3445901

Jaffe, R. (2020). Rethinking Metadata's Value and How It is Evaluated. *Technical Services Quarterly*, *37*(4), 432–443. https://doi.org/10.1080/07317131.2020.1810443

Jaillant, L. (Ed.). (2022). *Archives, Access and Artificial Intelligence: Working with Born-Digital and Digitized Archival Collections*. Bielefeld University Press. https://doi.org/10.14361/9783839455845

Jentzsch, S., & Turan, C. (2022). Gender Bias in BERT - Measuring and Analysing Biases Through Sentiment Rating in a Realistic Downstream Classification Task. *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 184–199. https://doi.org/10.18653/v1/2022.gebnlp-1.20

Jiang, M., & Fellbaum, C. (2020). Interdependencies of Gender and Race in Contextualized Word Embeddings. *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, 17–25. https://aclanthology.org/2020.gebnlp-1.2

Jin, X., Barbieri, F., Kennedy, B., Mostafazadeh Davani, A., Neves, L., & Ren, X. (2021). On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3770–3783. https://doi.org/10.18653/v1/2021.naacl-main.296

Jo, E. S., & Gebru, T. (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning. *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, 306–316. https://doi.org/10.1145/3351095.3372829

Joos, M. (1950). Description of Language Design. *The Journal of the Acoustical Society of America*, *22*(6), 701–707. https://doi.org/10.1121/1.1906674

Junginger, P., & Dörk, M. (2021). Categorizing Queer Identities: An Analysis of Archival Practices Using the Concept of Boundary Objects. *Journal of Feminist Scholarship*, *19*(19). https://doi.org/10.23860/jfs.2021.19.05

Jurafsky, D., & Martin, J. H. (2023). *Speech & Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition* [Online; accessed 20-May-2023]. https://web.stanford.edu/~jurafsky/slp3/

Kalluri, P. (2020). Don't Ask if Artificial Intelligence is Good or Fair, Ask How it Shifts Power. *Nature*, *583*(169). https://doi.org/10.1038/d41586-020-02003-2

Kaneko, M., & Bollegala, D. (2019). Gender-Preserving Debiasing for Pre-trained Word Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1641–1650. https://doi.org/10.18653/v1/P19-1160

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models. *Computing Research Repository*, *abs/2001.08361*. https://doi.org/10.48550/arxiv.2001.08361

Karen, H. (2023). What is ChatGPT? What to Know About the AI Chatbot. *Wall Street Journal*, *Tech*. https://www.wsj.com/articles/chatgpt-ai-chatbot-app-explained-11675865177

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356. https://doi.org/10.1145/3514094.3534188

Kasirzadeh, A., & Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning Language Models with Human Values. *Philosophy & Technology*, *36*(27). https://doi.org/10.1007/s13347-023-00606-x

Kaufman, G., Flanagan, M., & Seidman, M. (2021). 5 Creating Stealth Game Interventions for Attitude and Behavior Change: An 'Embedded Design' Model. In T. La Hera, J. Jansz, J. Raessens, & B. Schouten (Eds.), *Persuasive Gaming in Context*. Amsterdam University Press. https://doi.org/10.5117/9789463728805

Kaufman, G., & Flanagan, M. (2015). A Psychologically 'Embedded' Approach to Designing Games for Prosocial Causes. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, Special Issue on Videogames*, *9*(3). https://doi.org/10.5817/CP2015-3-5

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016). When (ish) is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5092–5103. https://doi.org/10.1145/2858036.2858558

Keskustalo, H., Korkeamäki, L., Vanamo, S., Kettunen, K., & Kumpulainen, S. (2023). Analyzing Gender Clues in War-Time Letters. *Digital Scholarship in the Humanities*, *38*(1), 209–223. https://doi.org/10.1093/llc/fqac035

Keyes, O. (2018). The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–22. https://doi.org/10.1145/3274357

Keyes, O., Hitzig, Z., & Blell, M. (2021). Truth from the Machine: Artificial Intelligence and the Materialization of Identity. *Interdisciplinary Science Reviews*, *46*(1–2), 158–175. https://doi.org/10.1080/03080188.2020.1840224

Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. (2021). Digital Cultural Colonialism: Measuring Bias in Aggregated Digitized Content Held in Google Arts and Culture. *Digital Scholarship in the Humanities*, *36*(3), 607–640. https://doi.org/10.1093/llc/fqaa055

KK, R. (2013). Coloring Outside the Lines [Online; accessed 3-Oct-2023]. https://www.youtube.com/watch?v=BXPGk0HXDpk

Koene, A., Perez, E., Ceppi, S., Rovatsos, M., Webb, H., Patel, M., Jirotka, M., & Lane, G. (2017). Algorithmic Fairness in Online Information Mediating Systems. *Proceedings of the 2017 ACM on Web Science Conference*, 391–392. https://doi.org/10.1145/3091478.3098864

Korobov, M. (2015). sklearn-crfsuite [Online; accessed 6-June-2023]. https://sklearn-crfsuite.readthedocs.io/en/latest

Krause, H. (2019). An Introduction to the Data Biography [Online; accessed 17-October-2020]. *We All Count*. https://weallcount.com/2019/01/21/an-introduction-to-the-data-biography/

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S.,

Suarez, P. O., ... Adeyemi, M. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50–72. https://doi.org/10.1162/tacl_a_00447

Krippendorff, K. (2018). *Content Analysis: An Introduction to its Methodology*. SAGE.

Kruppa, M. (2023). Google Launches Bard AI Chatbot to Counter ChatGPT. *Wall Street Journal*, *Tech*. https://www.wsj.com/articles/google-launches-bard-ai-chatbot-to-counter-chatgpt-2200c357

Kurita, K., Vyas, N., Pareek, A., Black, A. W., & Tsvetkov, Y. (2019). Measuring Bias in Contextualized Word Representations. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. https://doi.org/10.18653/v1/W19-3823

Kushner, K. E., & Morrow, R. (2003). Grounded Theory, Feminist Theory, Critical Theory: Toward Theoretical Triangulation. *Advances in Nursing Science*, *26*, 30–43. https://doi.org/10.1097/00012272-200301000-00006

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., & Petrov, S. (2019). Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, *7*, 452–466. https://doi.org/10.1162/tacl_a_00276

Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289. https://doi.org/10.5555/645530.655813

Lakoff, R. (1989). *Language and Woman's Place*. Harper & Row.

Lamb, W., Alex, B., & Sinclair, M. (2022). Handwriting Recognition for Scottish Gaelic. *Proceedings of the 4th Celtic Language Technology Workshop within LREC2022*, 60–70. https://aclanthology.org/2022.cltw4-1.9

Leavy, P. (2017a). Community-Based Participatory Research Design. *Research Design: Quantitative, Qualitative, Mixed Methods, Arts-Based and Community-Based Participatory Research Approaches* (pp. 224–254). Guilford Publications. https://ebookcentral.proquest.com/lib/ed/reader.action?docID=4832778%5C&ppg=244

Leavy, P. (2017b). Qualitative Research Design. *Research Design: Quantitative, Qualitative, Mixed Methods, Arts-Based and Community-Based Participatory Research Approaches* (pp. 131–163). Guilford Publications. https://ebookcentral.proquest.com/lib/ed/reader.action?docID=4832778%5C&ppg=244

Leavy, S. (2018). Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning. *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, 14–16. https://doi.org/10.1145/3195570.3195580

Lee, P. (2016). Learning from Tay's Introduction. *Microsoft Official Blog*. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

Lewis, J. E., Philip, N., Arista, N., Pechawis, A., & Kite, S. (2018). Making Kin with the Machines. *Journal of Design and Science*. https://doi.org/10.21428/bfafd97b

Lewis, M., & Lupyan, G. (2020). Gender Stereotypes are Reflected in the Distributional Structure of 25 Languages. *Nature: Human Behaviour*, *4*(10), 1021–1028. https://doi.org/10.1038/s41562-020-0918-6

Library of Congress (Ed.). (2017). *The Card Catalog: Books, Cards, and Literary Treasures*. Chronicle Books.

Library of Congress. (2021). Library of Congress Subject Headings PDF Files. https://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html

Loper, E., & Bird, S. (2002). NLTK: The Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, 63–70. https://doi.org/10.3115/1118108.1118117

Lorde, A. (1984). *Sister Outsider: Essays and Speeches*. Crossing Press.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2020). Gender Bias in Neural Natural Language Processing. Springer. https://doi.org/10.1007/978-3-030-62077-6_14

Luccioni, A., & Viviano, J. (2021). What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 182–189. https://doi.org/10.18653/v1/2021.acl-short.24

Lucy, L., & Bamman, D. (2021). Gender and Representation Bias in GPT-3 Generated Stories. *Proceedings of the Third Workshop on Narrative Understanding*, 48–55. https://doi.org/10.18653/v1/2021.nuse-1.5

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. https://doi.org/10.1147/rd.14.0309

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. https://aclanthology.org/P11-1015

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An Extensive Experimental Comparison of Methods for Multi-Label Learning. *Pattern Recognition*, *45*(9), 3084–3104. https://doi.org/10.1016/j.patcog.2012.03.004

Malik, V., Dev, S., Nishi, A., Peng, N., & Chang, K.-W. (2022). Socially Aware Bias Measurements for Hindi Language Representations. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1041–1052. https://doi.org/10.18653/v1/2022.naacl-main.76

Manjavacas, E., & Fonteyn, L. (2022). Adapting vs. Pre-Training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, *NLP4DH*(Digital Humanities in Languages), 1–19. https://doi.org/10.46298/jdmdh.9152

Manzo, C., Kaufman, G., Punjasthitkul, S., & Flanagan, M. (2015). By the People, For the People: Assessing the Value of Crowdsourced, User-Generated Metadata. *Digital Humanities Quarterly*, *9*(1).

Marchionini, G. (2006). Exploratory Search: From Finding to Understanding. *Communications of the ACM*, *49*(4), 41–46. https://doi.org/10.1145/1121949.1121979

Markl, N. (2022a). Language Variation and Algorithmic Bias: Understanding Algorithmic Bias in British English Automatic Speech Recognition. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 521–534. https://doi.org/10.1145/3531146.3533117

Markl, N. (2022b). Mind the Data Gap(s): Investigating Power in Speech and Language Datasets. *Proceedings of the Second Workshop on*

Language Technology for Equality, Diversity and Inclusion, 1–12. https://doi.org/10.18653/v1/2022.ltedi-1.1

Markl, N., & Lai, C. (2021). Context-Sensitive Evaluation of Automatic Speech Recognition: Considering User Experience & Language Variation. *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, 34–40. https://aclanthology.org/2021.hcinlp-1.6

Marshall, J. (2016). *Direcotry of Collections: University of Edinburgh* (H. Bowen, Ed.). Third Millenium Publishing.

Marston, G. (2000). Metaphor, Morality and Myth: Critical Discourse Analysis of Public Housing Policy in Queensland. *Critical Social Policy*, *20*(3), 249–274.

Martin, B., & Hanington, B. (2012a). 11 Case studies. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas and Design Effective Solutions* (p. 28). Rockport Publishers.

Martin, B., & Hanington, B. (2012b). 60 Participatory Action Research. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas and Design Effective Solutions* (pp. 126–127). Rockport Publishers.

Martin, B., & Hanington, B. (2012c). 67 Questionnaires. *Universal Methods of Design: 100 Ways to Research Complex Problems, Develop Innovative Ideas and Design Effective Solutions* (p. 140). Rockport Publishers.

Martinková, S., Stańczak, K., & Augenstein, I. (2023). Measuring Gender Bias in West Slavic Language Models. *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, 146–154. https://aclanthology.org/2023.bsnlp-1.17

McCradden, M., Mazwi, M., Joshi, S., & Anderson, J. A. (2020). When Your Only Tool is a Hammer: Ethical Limitations of Algorithmic Fairness Solutions in Healthcare Machine Learning. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (p. 109). Association for Computing Machinery. https://doi.org/10.1145/3375627.3375824

McGillivray, B., Alex, B., Ames, S., Armstrong, G., Beavan, D., Ciula, A., Colavizza, G., Cummings, J., De Roure, D., Farquhar, A., Hengchen, S., Lang, A., Loxley, J., Goudarouli, E., Nanni, F., Nini, A., Nyhan, J., Osborne, N., Poibeau, T., … Wilcox, P. (2020). The Challenges and

Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2020-08/humanities%5C_and%5C_data%5C_science%5C_white%5C_paper%5C_-%5C_updated.pdf

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

McPherson, T. (2012). Why are the Digital Humanities So White? Or Thinking the Histories of Race and Computation. *Debates in the Digital Humanities*, 139–160. https://doi.org/10.5749/minnesota/9780816677948.003.0017

Mehdi, Y. (2023). Reinventing Search with a New AI-Powered Microsoft Bing and Edge, your Copilot for the Web. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer Sentinel Mixture Models. *Computing Research Repository*. https://doi.org/10.48550/arxiv.1609.07843

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

Montréal Declaration for a Responsible Development of Artificial Intelligence. (2018). https://declarationmontreal-iaresponsable.com/

Moore, N. (2018). *A Cat's Cradle of Feminist and Other Critical Approaches to Participatory Research* [Online; accessed 24-July-2020]. University of Bristol/AHRC Connected Communities Programme. https://connected-communities.org/index.php/connected-communities-foundation-series/

Morrison, R. R. (2021). Flesh. In N. B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain Archives: Critical Keywords for Big Data* (pp. 249–258). The MIT Press. https://doi.org/10.7551/mitpress/12236.003.0027

Muntean, R., Antle, A. N., Matkin, B., Hennessy, K., Rowley, S., & Wilson, J. (2017). Designing Cultural Values into Interaction. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6062–6074. https://doi.org/10.1145/3025453.3025908

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabongo, S., Osei, S., Freshia, S., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Meressa, M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., ... Bashir, A. (2020). Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2144–2160). Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.findings-emnlp.195

Ng, R. (2023). Learn as You Search (and Browse) Using Generative AI. https://blog.google/products/search/google-search-generative-ai-learning-features/

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., & Chang, Y. (2016). Abusive Language Detection in Online User Content. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 145–153. https://doi.org/10.1145/2872427.2883062

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research. *Archival Science*, *22*(3), 367–392. https://doi.org/10.1007/s10502-022-09397-0

Noonan, P. (2023). A Six-Month AI Pause? No, Longer is Needed. *Wall Street Journal*, *Opinion*. https://www.wsj.com/articles/a-six-month-ai-pause-no-longer-is-needed-civilization-danger-chat-gpt-chatbot-internet-big-tech-4b66da6e

OCLC. (2023). Dewey Services: Improve the Organization of Your Materials. https://www.oclc.org/en/dewey.html

Odumosu, T. (2020). The Crying Child: On Colonial Archives, Digitization, and Ethics of Care in the Cultural Commons. *Current Anthropology*, *61*(S22), S289–S302. https://doi.org/10.1086/710062

OED. (n.d.). datum, n. [Online; accessed 29-August-2022]. *Oxford English Dictionary Online*. www.oed.com/view/Entry/47434

OED. (2013a). Classism [Online; accessed 21-August-2020]. *Oxford English Dictionary Online*.

OED. (2013b). Discourse [Online; accessed 17-October-2020]. *Oxford English Dictionary Online*.

OED. (2013c). Racism [Online; accessed 21-August-2020]. *Oxford English Dictionary Online*.

OED. (2013d). Sexism [Online; accessed 21-August-2020]. *Oxford English Dictionary Online*.

OED. (2023). Model [Online; accessed 6-June-2023]. *Oxford English Dictionary Online*. https://www.oed.com/view/Entry/120577?rskey=XvWzzB%5C&result=1%5C#eid

Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., & Matusik, W. (2019). Speech2Face: Learning the Face Behind a Voice. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Olson, H. A. (2001). The Power to Name: Representation in Library Catalogs. *Signs: Journal of Women in Culture and Society*, *26*(3), 639–668. https://doi.org/10.1086/495624

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Penguin Books.

Onuoha, M. (2016). The Library of Missing Datasets. https://mimionuoha.com/the-library-of-missing-datasets

Onuoha, M. (2017). On Missing Datasets. https://github.com/mimionuoha/missing-datasets

OpenAI. (2023). GPT-4 Technical Report. *Computing Research Repository*. https://doi.org/10.48550/arXiv.2303.08774

Orgad, H., Goldfarb-Tarrant, S., & Belinkov, Y. (2022). How Gender Debiasing Affects Internal Model Representations and Why It Matters. *Proceedings of the 2022 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, 2602–2628. https://doi.org/10.18653/v1/2022.naacl-main.188

Ørngreen, R., & Levinson, K. (2017). Workshops as a Research Methodology. *The Electronic Journal of eLearning*, *15*(1), 70–81. https://files.eric.ed.gov/fulltext/EJ1140102.pdf

Ortolja-Baird, A., & Nyhan, J. (2022). Encoding the Haunting of an Object Catalogue: On the Potential of Digital Technologies to Perpetuate or Subvert the Silence and Bias of the Early-Modern Archive. *Digital Scholarship in the Humanities*, *37*(3), 844–867. https://doi.org/10.1093/llc/fqab065

Padilla, T. (2017). On a Collections as Data Imperative. *UC Santa Barbara Previously Published Works*. https://escholarship.org/uc/item/9881c8sv

Padilla, T. (2019). Responsible Operations: Data Science, Machine Learning and AI in Libraries. *OCLC Research*, 38. https://doi.org/10.25333/xk7z-9g97

Pal, K., Avery, N., Boston, P., Campagnolo, A., De Stefani, C., Matheson-Pollock, H., Panozzo, D., Payne, M., Schüller, C., Sanderson, C., Scott, C., Smith, P., Smither, R., Sorkine-Hornung, O., Stewart, A., Stewart, E., Stewart, P., Terras, M., Walsh, B., ... Weyrich, T. (2016). Digitally Reconstructing the Great Parchment Book: 3D Recovery of Fire-Damaged Historical Documents. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqw057

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in Word Embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability and Transparency*, 446–457. https://dl.acm.org/doi/abs/10.1145/3351095.3372843

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and Its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns*, *2*(11). https://doi.org/10.1016/j.patter.2021.100336

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Perez, C. C. (2019). *Invisible Women: Exposing Data Bias in a World Designed for Men*. Vintage.

Pilcher, J., & Whelehan, I. (2004). *50 Key Concepts in Gender Studies* (1st ed.). SAGE.

Posner, M. (2016). What's Next: The Radical, Unrealized Potential of Digital Humanities. *Debates in the digital humanities 2016* (pp. 32–41). University of Minnesota Press. https://doi.org/10.5749/j.ctt1cn6thb.6

Raji, D., Denton, E., Bender, E. M., Hanna, A., & Paullada, A. (2021). AI and the Everything in the Whole Wide World Benchmark. In J. Vanschoren & S. Yeung (Eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran. https://datasets-benchmarks-proceedings.neurips.cc/paper%5C_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf

RDA Steering Committee. (2022). About RDA. https://www.rda-rsc.org/content/about-rda

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier Chains for Multi-Label Classification. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, 254–269. https://doi.org/10.1007/978-3-642-04174-7_17

Reason, P., & Bradbury-Huang, H. (2007). Introduction. *The SAGE Handbook of Action Research: Participative Inquiry and Practice* (pp. 1–10). SAGE Publications Ltd. https://doi.org/10.4135/9781848607934

Reid, C., & Frisby, W. (2008). Continuing the Journey: Articulating Dimensions of Feminist Participatory Action Research (FPAR). In P. Reason & H. Bradbury (Eds.), *The SAGE Handbook of Action Research* (pp. 93–105). SAGE Publications Ltd. https://doi.org/10.4135/9781848607934.n12

Ridge, M. (2013). From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing. *Curator: The Museum Journal*, *56*(4), 435–450. https://doi.org/10.1111/cura.12046

Ridge, M. (Ed.). (2016). *Crowdsourcing our Cultural Heritage*. Routledge. https://doi.org/10.4324/9781315575162

Rieke, A., & Bogen, M. (2018). *Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias*. https://upturn.org/work/help-wanted/

Risam, R. (2015). Beyond the Margins: Intersectionality and the Digital Humanities. *Digital Humanities Quarterly*, *9*(2), 14.

Risam, R. (2021). Digital Humanities. In N. B. Thylstrup, D. Agostinho, A. Ring, C. D'Ignazio, & K. Veel (Eds.), *Uncertain Archives: Critical Keywords for Big Data*. The MIT Press.

Ritchie, J., & Lewis, J. (2003). In-Depth Interviews. *Qualitative Research Practice: A Guide for Social Science Students and Researchers* (pp. 138–168). SAGE Publications Ltd.

Robertson, S., Wang, Z. J., Moritz, D., Kery, M. B., & Hohman, F. (2023). Angler: Helping Machine Translation Practitioners Prioritize Model Improvements. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3544548.3580790

Robson, C., & McCartan, K. (2016). The Analysis and Interpretation of Qualitative Data. *Real World Research: A Resource for Users of Social Research Methods in Applied Settings* (pp. 459–486). Wiley-Blackwell.

Rodolfa, K. T., Salomon, E., Haynes, L., Mendieta, I. H., Larson, J., & Ghani, R. (2020). Case Study: Predictive Fairness to Reduce Misdemeanor Recidivism Through Social Service Interventions. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 142–153. https://doi.org/10.1145/3351095.3372863

Rogers, A. (2021). Changing the World by Changing the Data. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2182–2194. https://doi.org/10.18653/v1/2021.acl-long.170

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. https://doi.org/10.18653/v1/N18-2002

Sahoo, N., Gupta, H., & Bhattacharyya, P. (2022). Detecting Unintended Social Bias in Toxic Language Datasets. *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, 132–143. https://aclanthology.org/2022.conll-1.10

Salway, A., & Baker, J. (2020). Investigating Curatorial Voice with Corpus Linguistic Techniques. *Museum and Society*, *18*(2), 151–169.

Sambasivan, N., Arnesen, E., Hutchinson, B., & Prabhakaran, V. (2020). Non-Portability of Algorithmic Fairness in India. https://doi.org/10.48550/arxiv.2012.03659

Samorani, M., Harris, S. L., Blount, L. G., Lu, H., & Santoro, M. A. (2022). Overbooked and Overlooked: Machine Learning and Racial Bias in Medical Appointment Scheduling. *Manufacturing & Service Operations Management*, *24*(6), 2825–2842. https://doi.org/10.1287/msom.2021.0999

Sandberg, S. (2015). *Lean In: Women, Work, and the Will to Lead*. WH Allen.

Sang, Y., & Stanton, J. (2022). The Origin and Value of Disagreement Among Data Labelers: A Case Study of Individual Differences in Hate Speech Annotation. *Information for a Better World: Shaping the Global Future* (pp. 425–444). Springer International Publishing.

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The Risk of Racial Bias in Hate Speech Detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. https://doi.org/10.18653/v1/P19-1163

Scheuerman, M. K., Paul, J. M., & Brubaker, J. R. (2019). How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–33. https://doi.org/10.1145/3359246

Scheuerman, M. K., Spiel, K., Haimson, O. L., Hamidi, F., & Branham, S. M. (2020). HCI Guidelines for Gender Equity and Inclusion. www.morgan-klaus.com/gender-guidelines.html

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–35. https://doi.org/10.1145/3392866

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Schulz, M. R. (2000). The Semantic Derogation of Women. In L. Burke, T. Crowley, & A. Girvin (Eds.), *The Routledge Language and Cultural Theory Reader*. Routledge.

Schwartz, J. M., & Cook, T. (2002). Archives, Records, and Power: The Making of Modern Memory. *Archival Science*, *2*(1–2), 1–19. https://doi.org/10.1007/BF02435628

scikit-learn developers. (2023a). 1.11 Ensemble Methods - scikit-learn 1.2.2 documentation [Online; accessed 8-June-2023]. https://scikit-learn.org/stable/modules/ensemble.html

scikit-learn developers. (2023b). Choosing the Right Estimator - scikit-learn 1.2.2 documentation [Online; accessed 8-June-2023]. https://scikit-learn.org/stable/tutorial/machine%5C_learning%5C_map/index.html

scikit-learn developers. (2023c). sklearn.linear_model.LogisticRegression - scikit-learn 1.2.2 documentation [Online; accessed 8-June-2023]. https://scikit-learn.org/stable/modules/generated/sklearn.linear%5C_model.LogisticRegression.html

scikit-learn developers. (2023d). Stochastic Gradient Descent - scikit-learn 1.2.2 documentation [Online; accessed 8-June-2023]. https://scikit-learn.org/stable/modules/sgd.html

Seetharaman, D. (2023). Elon Musk, Other AI Experts Call for Pause in Technology's Development. *Wall Street Journal*, *Tech*. https://www.wsj.com/articles/elon-musk-other-ai-bigwigs-call-for-pause-in-technologys-development-56327f

Shah, C., & Bender, E. M. (2022). Situating Search. *ACM SIGIR Conference on Human Information Interaction and Retrieval*, 221–232. https://doi.org/10.1145/3498366.3505816

Sharma, S., Dey, M., & Sinha, K. (2021). Evaluating Gender Bias in Natural Language Inference. *Computing Research Repository*. https://doi.org/10.48550/arXiv.2105.05541

Shopland, N. (2020). *A Practical Guide to Searching LGBTQIA Historical Records*. Taylor & Francis Group. https://doi.org/10.4324/9781003006787

Skloot, R. (2011). *The Immortal Life of Henrietta Lacks*. Pan.

Smith, L. (2006). *Uses of Heritage*. Routledge.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. https://aclanthology.org/D13-1170

Søgaard, A., Johannsen, A., Plank, B., Hovy, D., & Martínez Alonso, H. (2014). What's in a p-value in NLP? *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, 1–10. https://doi.org/10.3115/v1/W14-1601

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Spencer, D. (2000). Language and Reality: Who Made the World? (1980). In L. Burke, T. Crowley, & A. Girvin (Eds.), *The Routledge Language and Cultural Theory Reader*. Routledge.

Spiel, K., Keyes, O., & Barlas, P. (2019). Patching Gender: Non-Binary Utopias in HCI. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11. https://doi.org/10.1145/3290607.3310425

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2019). explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2019.2934629

Stańczak, K., & Augenstein, I. (2021). A Survey on Gender Bias in Natural Language Processing. *Computing Research Respository*, *abs/2112.14168*, 1–35. https://doi.org/10.48550/arXiv.2112.14168

Stanovsky, G., Smith, N. A., & Zettlemoyer, L. (2019). Evaluating Gender Bias in Machine Translation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. https://doi.org/10.18653/v1/P19-1164

Steed, R., Panda, S., Kobren, A., & Wick, M. (2022). Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models. *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), 3524–3542. https://doi.org/10.18653/v1/2022.acl-long.247

Stenetorp, P., Pyysalo, S., Topić, G., Tomoko Ohta, S. A., & Tsujii, J. (2012). brat: A Web-based Tool for NLP-Assisted Text Annotation. In F. Segond (Ed.), *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107). Association for Computational Linguistics. https://aclanthology.org/E12-2021

Stoler, A. L. (2002). Colonial Archives and the Arts of Governance. *Archival Science*, *2*(1–2), 87–109. https://doi.org/10.1007/BF02435632

Storvang, P., Mortensen, B., & Clarke, A. H. (2018). Using Workshops in Business Research: A Framework to Diagnose, Plan, Facilitate and Analyze Workshops. In P. V. Freytag & L. Young (Eds.), *Collaborative Research Design: Working with Business for Meaningful Findings* (pp. 155–174). Springer. https://doi.org/10.1007/978-981-10-5008-4_7

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645–3650. https://doi.org/10.18653/v1/P19-1355

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., & Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1630–1640. https://doi.org/10.18653/v1/P19-1159

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. https://doi.org/10.1145/3465416.3483305

Swantz, M. L. (2008). Participatory Action Research as Practice. *The SAGE Handbook of Action Research* (pp. 31–48). SAGE Publications Ltd. https://doi.org/10.4135/9781848607934.n8

Sweeney, C., & Najafian, M. (2019). A Transparent Framework for Evaluating Unintended Demographic Bias in Word Embeddings. *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, 1662–1667. https://doi.org/10.18653/v1/P19-1162

Sweeney, L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM*, *56*(5), 44–54. https://doi.org/10.1145/2447976.2447990

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019). What are the Biases in My Word Embedding? *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society*, 305–311. https://doi.org/10.1145/3306618.3314270

Szymański, P., & Kajdanowicz, T. (2018). A scikit-based Python Environment for Performing Multi-Label Classification. *Computing Research Repository*, *abs/1702.01460*, 1–22. https://doi.org/10.48550/arXiv.1702.01460

Tahaei, M., Constantinides, M., Quercia, D., Kennedy, S., Muller, M., Stumpf, S., Liao, Q. V., Baeza-Yates, R., Aroyo, L., Holbrook, J., Luger, E., Madaio, M., Blumenfeld, I. G., De-Arteaga, M., Vitak, J., & Olteanu, A. (2023). Human-Centered Responsible Artificial Intelligence: Current and Future Trends. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3544549.3583178

Tahaei, M., Constantinides, M., Quercia, D., & Muller, M. (2023). A Systematic Literature Review of Human-Centered, Ethical and Responsible AI. *Computing Research Repository*. https://doi.org/10.48550/arXiv.2302.05284

Tai, J. (2021). Cultural Humility as a Framework for Anti-Oppressive Archival Description. *Journal of Critical Library and Information Studies*, *3*(2). https://doi.org/10.24242/jclis.v3i2.120

Talbot, M. (2003). Gender Stereotypes: Reproduction and Challenge. *The Handbook of Language and Gender*, 468–486. https://doi.org/10.1002/9780470756942.ch20

Tan, Y. C., & Celis, L. E. (2019). Assessing Social and Intersectional Biases in Contextualized Word Representations. *Proceedings of the 2019 Advances in Neural Information Processing Systems Conference*, *32*, 1–12. https://proceedings.neurips.cc/paper_files/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf

Tanselle, G. T. (2002). The World as Archive. *Common Knowledge*, *8*(2), 402–406.

Taylor, A. (2018). The Automation Charade. *Logic(s) Magazine*, *Failure*(5). https://logicmag.io/failure/the-automation-charade/

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., & Yuan, A. (2020). The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 107–118.

Terras, M. (2015). Opening Access to Collections: The Making and Using of Open Digitised Cultural Content (G. Gorman & P. Professor Jennifer, Eds.). *Online Information Review*, *39*(5), 733–752. https://doi.org/10.1108/OIR-06-2015-0193

Terras, M., Baker, J., Hetherington, J., Beavan, D., Zaltz Austwick, M., Welsh, A., O'Neill, H., Finley, W., Duke-Williams, O., & Farquhar, A. (2018). Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High-Performance Computing, and Transforming Access to British Library Digital Collections. *Digital Scholarship in the Humanities*, *33*(2), 456–466. https://doi.org/10.1093/llc/fqx020

Terras, M., Nyhan, J., & Vanhoutte, E. (2013). *Defining Digital Humanities: A Reader*. Ashgate Publishing. https://doi.org/10.4324/9781315576251

Thatcher, J., O'Sullivan, D., & Mahmoudi, D. (2016). Data Colonialism Through Accumulation by Dispossession: New Metaphors for Daily Data. *Environment and Planning D: Society and Space*, *34*(6), 990–1006. https://doi.org/10.1177/0263775816633195

The pandas development team. (2023). *pandas-dev/pandas: Pandas* (Version 1.5.3). Zenodo. https://doi.org/10.5281/zenodo.7549438

Thomassen, T. (2002). A First Introduction to Archival Science. *Archival Science*, *1*, 373–385.

Thornton, P. (2016). {Poem}.py: a critique of linguistic capitalism. https://pipthornton.com/2016/06/12/poem-py-a-critique-of-linguistic-capitalism/

Thornton, P. (2017). Geographies of (Con)text: Language and Structure in a Digital Age. *Computational Culture*, (6). https://computationalculture.net/geographies-of-context-language-and-structure-in-a-digital-age/

Thylstrup, N. B. (2022). The Ethics and Politics of Data Sets in the Age of Machine Learning: Deleting Traces and Encountering Remains. *Media, Culture & Society*, *44*(4), 655–671. https://doi.org/10.1177/01634437211060226

Thylstrup, N. B., Agostinho, D., Ring, A., D'Ignazio, C., & Veel, K. (Eds.). (2021). *Uncertain Archives: Critical Keywords for Big Data*. The MIT Press.

Trask, A., Michalak, P., & Liu, J. (2015). sense2vec - A Fast and Accurate Method for Word Sense Disambiguation in Neural Word Embeddings. *Computing Research Repository*, *abs/1511.06388*, 1–9. https://doi.org/10.48550/arXiv.1511.06388

Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.-C., Carroll, A., Lau, C., Tanno, R., Ktena, I., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D., Mansfield, P., Prakash, S., Wong, R., Virmani, S., ... Natarajan, V. (2023). Towards Generalist Biomedical AI. *Computing Research Repository*. https://doi.org/10.48550/arxiv.2307.14334

Vainapel, S., Shamir, O. Y., Tenenbaum, Y., & Gilam, G. (2015). The Dark Side of Gendered Language: The Masculine-Generic Form as a Cause for Self-Report Bias. *Psychological Assessment*, *27*(4), 1513–1519. https://doi.org/10.1037/pas0000156

van den Berg, E., & Markert, K. (2020). Context in Informational Bias Detection. *Proceedings of the 28th International Conference on Computational Linguistics*, 6315–6326. https://doi.org/10.18653/v1/2020.coling-main.556

van Leeuwen, T. (2009). Discourse as the Recontextualization of Social Practice: A Guide. *Methods for Critical Discourse Analysis*.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth; Co. https://www.dcs.gla.ac.uk/Keith/Preface.html

Verdegem, P. (2021). Introduction: Why We Need Critical Perspectives on AI. In P. Verdegem (Ed.), *AI for Everyone?* (pp. 1–18). University of Westminster Press. https://www.jstor.org/stable/j.ctv26qjjhj.3

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific

Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. https://doi.org/10.18653/v1/W18-5446

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Computing Research Repository*. https://doi.org/10.48550/arXiv.1810.05201

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., . . . Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Welsh, A. (2016). The Rare Books Catalog and the Scholarly Database. *Cataloging & Classification Quarterly*, *54*(5–6), 317–337. https://doi.org/10.1080/01639374.2016.1188433

Welsh, A., & Batley, S. (2009). *Practical Cataloguing: AACR, RDA and MARC21*. Facet.

Welty, C., Paritosh, P., & Aroyo, L. (2019). Metrology for AI: From Benchmarks to Instruments. *Computing Research Repository*, *abs/1911.01875*. https://doi.org/10.48550/arXiv.1911.01875

Wetli, A. (2019). Addressing Cultural Insensitivity in Archival Description: A Literature Review Examining Collaborative Approaches. *Journal of New Librarianship*, *4*(2), 505–515. https://doi.org/10.21173/newlibs/8/3

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., & Wilson, J. (2019). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, *26*, 56–65. https://doi.org/10.1109/TVCG.2019.2934619

White, R. W., & Roth, R. A. (2009). *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Publishers.

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, *9*(1), 1–16.

Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. https://doi.org/10.18653/v1/N18-1101

Wood, S., Carbone, K., Cifor, M., Gilliland, A., & Punzalan, R. (2014). Mobilizing Records: Re-Framing Archival Description to Support Human Rights. *Archival Science*, *14*, 397–419. https://doi.org/10.1007/s10502-014-9233-1

Wurl, J. (2005). Ethnicity as Provenance: In Search of Values and Principles for Documenting the Immigrant Experience. *Archival Issues*, *29*(1), 65–76. https://www.jstor.org/stable/41102095

Yale, E. (2015). The History of Archives: The State of the Discipline. *Book History*, *18*(1), 332–359. https://doi.org/10.1353/bh.2015.0007

Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H., & Miklau, G. (2018). A Nutritional Label for Rankings. *Proceedings of the 2018 International Conference on Management of Data*, 1773–1776. https://doi.org/10.1145/3183713.3193568

Yilmazel, O., Finneran, C. M., & Liddy, E. D. (2004). Metaextract: An NLP System to Automatically Assign Metadata. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries - JCDL '04*. https://doi.org/10.1145/996350.996405

Young, I. M. (2011). *Responsibility for Justice* (M. Nussbaum, Ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Zadrozny, B., & Elkan, C. (2002). Transforming Classifier Scores into Accurate Multiclass Probability Estimates. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699. https://doi.org/10.1145/775047.775151

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. *Proceedings of the ACM Conference on Health, Inference and Learning*, 110–120. https://doi.org/10.1145/3368555.3384448

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-Level Convolutional Networks for Text Classification. In C. Cortes, N. Lawrence, D. Lee, M.

Sugiyama, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper%5C_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., & Chang, K.-W. (2019). Gender Bias in Contextualized Word Embeddings. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 629–634. https://doi.org/10.18653/v1/N19-1064

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2979–2989. https://doi.org/10.18653/v1/D17-1323

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2018). Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. https://doi.org/10.18653/v1/N18-2003

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K.-W. (2018). Learning Gender-Neutral Word Embeddings. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4847–4853. https://doi.org/10.18653/v1/D18-1521

Zheng, K., Wang, H., Qi, Z., Li, J., & Gao, H. (2017). A Survey of Query Result Diversification. *Knowledge and Information Systems*, *51*, 1–36. https://doi.org/10.1007/s10115-016-0990-4

Appendix A

Data Statement: Unannotated Data

Dataset of Select Catalog Metadata Descriptions from the University of Edinburgh's Archives

October 2020

A.1 Curation Rationale

I extracted metadata descriptions from the University of Edinburgh's Heritage Collections (HC) Archives' catalog[1] to create an annotated dataset for training text classification models to detect gender biased language. I define gender biased language as:

written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity; and privileging other people through words or phrases that favor their identity (Havens et al., 2020).

I extracted text from four descriptive metadata fields for all collections, subcollections, and items in the HC Archives' online catalog. The "Title" field is the name of the archival record, which either documents a single

[1]The online catalog can be visited at: archives.collections.ed.ac.uk

or group of archival material. The "Biographical / Historical" field contains information about the people, time period, and places associated with the collection, subcollection, or item being described. The "Scope and Contents" field summarizes the contents of the collection, subcollection, or item to which the field belongs. Though not all records include the "Processing Information" field, those that do typically record the person who wrote the description for the collection, subcollection, or item's descriptive metadata fields, and the date the description was written.

We curated this dataset of extracted archival catalog metadata descriptions to training discriminative classification models to identify types of contextual gender bias. Additionally, the dataset will serve as a source of annotated, historical text to complement already existing datasets for NLP primarily composed of contemporary texts (i.e. from social media, Wikipedia, news articles). We chose to use archival metadata descriptions as a data source because:

1. Metadata descriptions in the Archives' catalog (and most GLAM catalogs) are freely, publicly available online

2. GLAM metadata descriptions have yet to be analyzed at large scale using NLP methods and, as records of cultural heritage, the descriptions have the potential to provide historical insights on changes in language and society (Welsh, 2016)

3. GLAM metadata standards are freely, publicly available, often online, meaning we can use historical changes in metadata standards used in the Archive to guide large-scale text analysis of changes in the language of the metadata descriptions over time

4. The HC Archives' policy acknowledges its responsibility to address legacy descriptions in its catalogs that use language considered biased or otherwise inappropriate today[2]

[2]The HC Archives is not alone; across the GLAM sector, institutions acknowledge and are exploring ways to address legacy language in their catalogs' descriptions. The "Note" in We Are What We Steal provides one example: dxlab.sl.nsw.gov.au/we-are-what-we-steal/notes/.

A.2 Language Variety

The metadata descriptions extracted from the HC Archives' catalog are written in British English.

A.3 Producer Demographic

I am American and identify as a ciswoman. My Ph.D. supervisors, who provided feedback throughout the dataset curation process, are of German and Scots nationalities, and identify as two women and one man. We all work primarily as academic researchers in the disciplines of natural language processing, data science, data visualization, human-computer interaction, digital humanities, and digital cultural heritage. Additionally, one of us has audited an online course on feminist and social justice studies.

A.4 Annotator Demographic

Not applicable (the dataset is not annotated).

A.5 Speech or Publication Situation

The metadata descriptions extracted from the HC Archives' online catalog using Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH). For OAI-PMH, an institution (in this case, the HC Archives) provides a URL to its catalog that displays its catalog metadata in XML format. A member of our research team wrote scripts in Python to extract three descriptive metadata fields for every collection, subcollection, and item in the Archive's online catalog (the metadata is organized hierarchically). Using Python and its Natural Language Toolkit (NLTK) library, the researcher removed duplicate sentences and calculated that the extracted metadata descriptions consist of a total of 2,754,044 tokens, an estimated 1,273,237 words (alphabetic tokens), and 156,124 sentences across 1,081 collections. Per collection, the length of descriptive text ranges from eight to 63,458 tokens.

Please refer to the Provenance Appendix (§A.9) for information on the Speech or Publication Situation of all of the HC Archives' metadata descriptions.

A.6 Data Characteristics

Upon extracting the metadata descriptions using OAI-PMH, the XML tags were removed so that the total words and sentences of the metadata descriptions could be calculated to ensure the text source provided a sufficiently large dataset. A member of our research team has grouped all the extracted metadata descriptions by their collection (the *fonds* level in the XML data), preserving the context in which the metadata descriptions were written and will be read by visitors to the Archive's online catalog.[3]

A.7 Data Quality

As a member of our research team extracts and filters metadata descriptions from the Archive's online catalog, they write assertions and tests to ensure as best as possible that metadata isn't being lost or unintentionally changed.

Please refer to the Provenance Appendix (§A.9) for information on the Data Quality of all of the HC Archives' metadata descriptions.

A.8 Other

The dataset can be downloaded from the University of Edinburgh's DataShare platform at: https://doi.org/10.7488/ds/2953.

A.9 Provenance Appendix

Data Statement of the Heritage Collections Archives Catalog at the University of Edinburgh

August 2023

[3]The code for the extraction and data cleaning processes is at: github.com/thegoose20/annot-prep.

A.9.1 Curation Rationale

The Heritage Collections (HC) Archives' policy describes a commitment to develop collections that are as inclusive and diverse as possible, keeping up with social changes and looking for opportunities to better represent communities of people. Additionally, the HC Archives' policy states that the institution aims to make its collections accessible to as many people as possible.

A.9.2 Language Variety

The HC Archives' metadata descriptions are written primarily in British English. As of 2023, the material in the *Manuscripts of the Islamicate World and South Asia* collection have titles written in Arabic or Persian, along with English, and can be searched for in those languages.

A.9.3 Producer Demographic

People who write metadata descriptions to document the HC Archives' collections include employees, interns, and volunteers. Employees have received professional training in archival documentation, in addition to training at the HC Archives. Interns and volunteers are typically students studying Information Sciences, Museology, History, or related disciplines who have also received training at the HC Archives. The institution began in the 16th century, so the metadata descriptions in its online catalog date from that time period up through the present day (the HC Archives continues to collect and document cultural heritage records).

Additional demographic information on all those who have written the HC Archives' metadata descriptions is limited, however the HC Archives is based in the United Kingdom, meaning the perspectives of those who wrote the descriptions is most likely English, Irish, Scottish, British, or European. The HC Archives is closely associated with a research university, the University of Edinburgh, so interns and volunteers who write the HC Archives' catalog metadata descriptions are likely to have received, or be in the process of receiving, higher education degrees.

A.9.4 Annotator Demographic

Not applicable.

A.9.5 Speech or Publication Situation

The metadata descriptions in the HC Archives' online catalog document collections created by a university associated with the Archive and acquired or donated from other people and organizations. The HC Archives' earliest metadata descriptions were written in the 16th century; metadata descriptions continue to be written today.

The goal of the metadata descriptions is to help people find primary source material in the HC Archives. At the time most of the HC Archives' metadata descriptions were written, the descriptions were intended for employees of the Archive, who would help visitors locate primary source material. Circa 2015, employees of the HC Archives began writing metadata descriptions with visitors included in their intended audience.

Current employees at the HC Archives have stated that they would be happy for the metadata descriptions they write to be viewed as works in progress, because the HC Archives could never have enough time to document all its collection items completely. Moreover, often information about collections items is impossible to know due to their historical nature and lack of accompanying documentation, so the metadata descriptions will always be incomplete.

The metadata descriptions include information available from the cultural heritage records they describe, from any available documentation that accompanied those records when the HC Archives acquired them, from authorities such as the Library of Congress Subject Headings, and from other documentation resources considered trustworthy in the Archives sector.

A.9.6 Data Characteristics

Beginning circa 2017, people documenting collections in the HC Archives have written metadata descriptions according to the General International Standard Archival Description (ISAD(G)). Past metadata descriptions were written according to library metadata standards. Metadata descriptions may include

contextual information about the people, places, and time periods relevant to the collection items, as well as the date a description was written and who wrote the description. Though all of this descriptive information ideally exists for a collection item, some collection items do not have this complete of a description.

A.9.7 Data Quality

The metadata descriptions in the HC Archives' online catalog consists of manually entered data, some of which was initially written in digital form, and some of which was initially written on paper and has since been manually typed into digital form.

A.9.8 Other

Not applicable.

A.9.9 Provenance Appendix

Not applicable.

Appendix B

Power Relations Document

Stakeholder Power Relations in NLP Research on Bias in Archival Metadata Descriptions

The Stakeholders

Identification:

1. The research team
2. Employees of the Heritage Collections (HC) Archives (current and former) who wrote the metadata descriptions that serve as this research's text source
3. The HC Archives and its associated university, the University of Edinburgh, as institutions that provide access to the metadata descriptions
4. People represented in the metadata descriptions
5. Visitors to the HC Archives, as they will read the metadata descriptions used as this research's text source when using the Archives' online catalog

Limitations: Due to the length of the text and the historical nature of the metadata descriptions we use from the Archives' catalog, we do not have access to every person represented in the metadata descriptions. However,

the Archives does have a take-down policy that we will follow with our text source to respect the people represented in metadata descriptions as best as possible: if a person requests that information about them or someone they are connected to be removed from or anonymized in the catalog, the Archives will comply. To the best of our ability, we will make sure that the metadata descriptions we use as the text source for our research do not include information that a visitor has requested the Archives take down.

Power Relations Questions

Who or what is included in the research?

Who:

- Current employees of the HC Archives: To account for intragroup differences, we include employees with different years of experience and employees working in several positions within the hierarchy of job roles in the HC Archives.
- Us (the research team): The size of the team is small enough that all members are included, meaning intragroup differences are accounted for by default.

To Do: Find visitors to the HC Archives who I can speak to about their experience reading its catalog's metadata descriptions. To account for intragroup differences among visitors, we will seek out a selection of visitors with as diverse of identity characteristics as possible.

What: Ongoing work includes conducting historical research to understand the context in which the metadata descriptions were written. For example, employees at the HC Archives stated that for many years, people wrote metadata descriptions with the aim of being as neutral and objective as possible, however the latest generation of archivists is challenging this, arguing that neutrality isn't possible and encouraging transparency instead.

Who or what is excluded from the research?

Who:

- Past employees of the HC Archives
- People represented in the HC Archives' cultural heritage records
- The majority of the HC Archives' visitors (the research only has the capacity to include a selection of visitors in user research and participatory action research activities)

What: The historical context of metadata descriptions written before my lifetime

To Do: Determine if policy guidelines for the HC Archives since its beginnings in the 16th century are available to understand how it perceived itself and what drove its collection and documentation practices. Otherwise, the historical existence of the HC Archives is also excluded form the research.

How will the research define knowledge?

The research will define knowledge as multifaceted. We (the research team) will draw on the disciplines of gender studies and linguistics to manually identify and annotate types of contextual gender bias in metadata descriptions. The research will share the annotated dataset as one interpretation of gender bias, recognizing that different people have different experiences of oppression that cause variations in attitudes towards words or phrases.

We will use the annotated dataset to train a discriminative classification algorithm. The types of gender bias that the algorithm identifies will be presented as potentially biased text, requiring verification from a person working with the text to decide whether the text should be considered biased.

Who has agency and who can be empowered?

We (the research team) have agency as the people applying NLP methods to the HC Archives' metadata descriptions.

The employees of the HC Archives can be empowered through participatory action research, with collaborative activities in which we situate the employees as partners in the research and as experts on archival practices and metadata.

The employees of the HC Archives have determined that people who do not identify as male are underrepresented in the HC Archives' collections and thus

those collections' metadata descriptions. We focus our bias identification and classification efforts on gender bias to explore how we can empower people who do not identify as male through the process and outputs of our NLP research.

Appendix C

Data Biography

Dataset of Catalog Metadata from the University of Edinburgh's Archives

Where was the Data Collected or Created?

I collected the data from the Archives' online catalog[1] using the Open Access Initiative - Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH provides access to the catalog as eXtensible Markup Language (XML) data in Encoded Archival Description (EAD) format, which I converted to Plaintext and Comma-separated Values formats. I obtained the descriptive text from the following metadata fields: EAD Identifier (EADID), Title, Biographical / Historical, Scope and Contents, Processing Information, Date (of archival material), Language (of archival material), and Geography (of archival material).

Employees, interns, and volunteers at the Archives who wrote the metadata descriptions collected information to include in the descriptions from documentation accompanying the cultural heritage record(s) they were describing, from the cultural heritage records themselves, from authorities such as Library of Congress Subject Headings, and from other metadata standards, thesauri, and description resources for archival documentation. The Archives is a part of Heritage Collections at the University of Edinburgh, within Library & University Collections. At the time of data collection, Heritage Collections was referred to as the Centre for Research Collections.

[1] archives.collections.ed.ac.uk

Where possible, we provide dates associated with the descriptions, and the material the description describe, to contextualize their text in relation to historical time periods and historical changes in metadata structures. For example, the metadata standard Library of Congress Subject Headings (LCSH) once used the term "Jewish Question" instead of the current term "Jews," so GLAM who use LCSH may have descriptions in their catalogs that use the historical term now considered biased.

Who Collected or Created the Data?

The Archives and the university to which it is associated, the University of Edinburgh, collected some of the cultural heritage records and the accompanying documentation that informs the records' metadata descriptions. For other cultural heritage records and their accompanying documentation, individual collectors gathered the records and wrote their documentation, which employees, interns, and volunteers used to write descriptive metadata for the records in the Archives' catalog.

The Archives has existed since the 16th century, so its directors will each have established different policies and goals for acquiring and documenting cultural heritage records (Marshall, 2016). The latest policy document for the Archives includes a statement about diversity, inclusion, and accessibility that describes the Archives' commitment to providing representative collections for local, national, and international audiences.[2]

Why was the Data Collected or Created?

The Archives' policy explains that it documents cultural heritage records in its catalog so that researchers can find the records and use them as primary source material to guide their work. Current employees of the Archives reiterated the goal of discoverability as the main reason for writing metadata descriptions.

Individuals and institutions who have donated their collections to the Archives had personal reasons motivating their choices of records to save. A directory of the Archives' collections contains information about select individuals and institutions that suggest their reasons for saving the records

[2] www.ed.ac.uk/information-services/about/policies-and-regulations/operational-policies/collections

they did. Information in the metadata descriptions themselves may also provide insight on why their associated records were collected.

When was the Data Collected or Created?

Among the metadata descriptions we extracted that include a year documenting when they were written, the years show that the descriptions were written from the 19th century up through the 21st century. Further research is needed to determine how early the extracted metadata descriptions without a year were written. I collected the archival metadata descriptions data using OAI-PMH in April 2020.

Appendix D

Annotation Instructions

The annotation instructions were written to guide annotators in applying the Taxonomy of Gendered and Gender Biased Language to the annotation corpus of archival documentation. Prior to beginning the annotation process, an annotation pilot was undertaken with three participants to test the clarity of the Taxonomy. The pilot led to revisions of the instructions: more examples were added and annotators were explicitly instructed to read and interpret the descriptions from their contemporary perspective.

The annotation instructions below contain a slightly different annotation Taxonomy than the final annotation Taxonomy of Gendered and Gender Biased Language in Chapter 5. This is due to the fact that during and after the annotation process, the Taxonomy was revised based on the data that was being annotated. The definitions of Gendered Role and Generalization proved to be difficult to distinguish in practice, so the definitions were revised during the dataset aggregation process. Additionally, we realized during the annotation process that *Woman* and *Man* were inaccurate labels based on what we could learn about gender from text, so we changed these labels to *Feminine* and *Masculine*, respectively, for the final Taxonomy.

Instructions

Step 1: As you read and label the archival metadata descriptions displayed on the screen, including text that quotes from source material, meaning text surrounded in quotation marks that reproduces something written in a letter, manuscript, or other text-based record from an archival collection.

NOTE: If you are unsure about an annotation, please make a note the file name and your question so that we can discuss it and decide on the way to annotate that sort of language moving forward!

Step 2: Please note that *Gendered Pronouns*, *Gendered Roles*, and *Occupations* have been pre-annotated. If any of these three categories of language have been annotated incorrectly, please correct them by clicking on the annotation label, deleting it, and making the correct annotation. If any of these three categories of language have been missed in the pre-annotation process, please annotate them yourself.

Step 3: Read the archival metadata descriptions displayed and while reading:

- Use your mouse to highlight a selection of text or click on a word that uses gendered language according to the schema in the table on the next page.
- Using the keyboard shortcuts (see the table) or your mouse, select the type of gendered language you've identified. Please select the most specific label possible (listed as (a), (b), (c), or (d))! Please only select Person-Name, Linguistic or Contextual if you do not feel their subcategories are suitable to the gendered language you would like to annotate.
- If you select a subcategory of Contextual gendered language, please write a brief note explaining what you've annotated as gendered in the "Notes" section of the "New/Edit Annotation" pop-up window.
- If you used your mouse to open the pop-up window, press the Enter/Return key or the "OK" button to make the annotation.
- You may make overlapping annotations, meaning a single word or phrase may have multiple gendered language annotations.
- Please annotate all instances of a particular type of gendered language used for a specific person or people in the text.

- Please note that the labels to annotate with as defined below are intended to guide your interpretation of the text through a contemporary lens (not a historical lens).

The examples provided in the schema below are highlighted according to the words, phrases or sentences that should be highlighted or clicked in brat. If in doubt about how much to annotate, please annotate more words rather than less!

1. ***Person Name***: the name of a person including any pre-nominal titles they have (i.e. Professor, Mrs., Sir)

   NOTE 1: Please annotate every instance of a name in brat only (do not use a spreadsheet anymore). This means that each person may have multiple person-name labels annotating the same form of their name or different forms of their name.

   NOTE 2: Use the pronouns and roles that occur within the descriptive field in which the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") to determine whether the annotation label should be Woman, Man, Non-binary, or Unknown. Please do not use the occupation, name, or other information that implies a gender to determine the annotation label; only use explicit terms such as gender-marking pronouns (him, her, he, she, himself, herself, etc.) and gender-marking roles (mother, father, daughter, wife, husband, son, Mrs., Ms., Mr., etc.).

   (a) *Woman*: the pronouns (i.e. she, her) or roles (i.e., mother, wife, daughter, grandmother, Mrs., Ms., Queen, Lady, Baroness) or use of term *nee [Last Name]* indicating a maiden name within the descriptive field in which the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") of the named person suggest they are a woman

   Example: Mrs. Jane Bennet went to Huntsford.

   (b) *Men*: the pronouns, roles, or titles of the named person suggest they are a man

Example: Conrad Hal Waddington lived in Edinburgh and he published scientific papers.

(c) *Non-binary*: the pronouns or roles of the named person within the descriptive field in which this instance of the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") suggest they are non-binary

NOTE: a preliminary search of the text returned no results for exclusively non-binary pronouns such as Mx, so most likely any non-binary person would be indicated with "they"); if the gender of a person is named and it's not a woman or man, please note this gender in the "Notes" section of the annotation pop-up window

Example: Francis McDonald went to the University of Edinburgh where they studied law.

(d) *Unknown*: there are no pronouns or roles for the named person within the descriptive field in which this instance of the name appears (either "Title," "Scope and Contents," "Biographical / Historical," or "Processing Information") that suggest their gender identity

Example: Jo McMahon visited Edinburgh in 1900.

2. ***Linguistic***: gender marked in the way a sentence references a person or people, assigning them a specific gender that does not account for all genders possible for that person or group of people (Keyboard shortcut: L)

(a) *Generalization*: use of a gender-specific term to refer to a group of people (including the job title of a person) that could identify as more than the specified gender (Keyboard shortcut: G)

Example 1: The chairman of the university was born in 1980. Explanation: Chair would be the gender-neutral form of chairman

Example 2: Readers, scholars, and workmen Explanation: readers and scholars are gender-neutral, while workers would be the gender-neutral form of workmen

Example 3: Housewife

(b) *Gendered Pronoun*: explicitly marking the gender of a person or people through the use of the pronouns he, him, his, her, and she (Keyboard shortcut: P)

Example 1: She studied at the University of Edinburgh. In 2000, she graduated with a degree in History.

Example 2: This manuscript belonged to Sir John Hope of Craighill. Sir John Hope was a judge. He lived in Scotland.

(c) *Gendered Role*: use of a title or word denoting a person's role that marks either a masculine or feminine gender (Keyboard shortcut: R)

Example 1: Sir Robert McDonald, son of Sir James McDonald

Example 2: Mrs. Jane Do

Example 3: Sam is the sister of Charles

Example 4: Sir Robert McDonald, son of Sir James McDonald

3. ***Contextual***: gender bias that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (i.e. sentence structure, word choice) (Keyboard shortcut: C)

(a) *Occupation*: occupations, whether or not they explicitly communicate a gender, should be annotated, as statistics from external data sources can be used to estimate the number of people of different genders who held such occupations; please label words as occupations if they'd be a person's job title and are how the person would make money, but not if the words are used as a title (Keyboard shortcut: J)

Example 1: minister

Example 2: Sergeant-Major-General

(b) *Stereotype*: language that communicates an expectation of a person or group of people's behaviors or preferences that does not reflect the reality of all possible behaviors/preferences that person or group of people may have, or language that focuses on a particular aspect of a person that doesn't represent that person holistically; for example, women described in relation to their family and home, and men in relation to their careers and

workplace; men more associated with science and women more associated with liberal arts (Keyboard shortcut: S)

NOTE: Please label whichever words, phrases, or sentences you feel communicate the stereotype. Three different examples are shown below for how this may look. Include names being turned into ways of thought (e.g. Bouldingism, Keynsian).

Example 1: The event was sports-themed for all the fathers in attendance. *Explanation: The assumption here is that all fathers and only fathers would enjoy a sports-themed event. A neutral alternative sentence could read: The event was sports-themed for all the former athletes in attendance*

Example 2: A programmer works from his computer most of the day. *Explanation: The assumption here is that any programmer must be a man, since the indefinite article "A" is used with the pronoun "his"*

Example 3: A man with no doctorate degree being known as Dr. Jazz *Explanation: Women often receive negative attention for using titles such as Dr (see the WSJ op-ed on Dr Jill Biden for a recent example) while men typically do not*

(c) *Omission*: focusing on the presence, responsibility, or contribution of a single gender in a situation in which more than one gender has a presence, responsibility or contribution; or defining a person's identity in terms of their relation to another person (Keyboard shortcut: O)

NOTE: If initials are provided, consider that enough of a name that it doesn't need to be labeled as an omission!

Example 1: Mrs. John Williams lived in Edinburgh. *Explanation: Mrs. John Williams is, presumably, referred to by her husband's first and last name rather than her given name*

Example 2: Mr. Arthur Cane and Mrs. Cane were married in 1850. *Explanation: Mrs. Cane is not referred to by her given name*

Example 3: Mrs. Elizabeth Smith and her husband went to Scotland. *Explanation: The husband is not named, being referred to only by his relationship to Mrs. Elizabeth Smith*

Example 4: His name was Edward Kerry, son of Sir James Kerry. *Explanation: paternal relations only, no maternal relations*

Example 5: The novelist, Mrs. Oliphant, wrote a letter. *Explanation: Mrs. Oliphant is referred to by the last name she shares with her husband without including her given name*

(d) *Empowering*: use of gendered language to challenge stereotypes or norms that reclaims derogatory terms, empowering a minoritized person or people; for example, using the term queer in an empowering rather than a derogatory manner (Keyboard shortcut: E)

Example: "Queer" being used in a self-affirming, positive manner to describe oneself

Step 4: If you would like to change an annotation you have made, double click the annotation label. If you would like to remove the annotation, click the "Delete" button in the pop-up window. If you would like to change the annotation, click the label you would like to change to and then click the "OK" button.

Step 5: Click the right arrow at the top left of the screen to navigate to the next archival metadata description (if you would like to return to a previous description, click the left arrow).

Step 6: If the screen does not advance when you click the right arrow, you've reached the end of the folder you're currently in. To move onto the next file, please hover over the blue bar at the top of the screen and click the "Collection" button. Click the first list item in the pop-up window "../" to exit your current folder and then double click the next folder in the list. Double click the first file in this next folder to begin annotating its text.

Step 7: Repeat from step 1.

Appendix E

Data Statement: Annotated Data

Dataset of Archival Metadata Descriptions Manually Annotated for Gender Bias

July 2022

E.1 Curation Rationale

These datasets were created from a corpus of 1,460 files of archival catalog metadata descriptions (from the Archives of Heritage Collections, University of Edinburgh, referred to as HC Archives) totaling circa 15,419 sentences and 255,943 words. That corpus is the first 20% of text from the corpus described in the data statement in Appendix A, annotated for gender bias according the the Taxonomy of Gendered and Gender Biased Language (see §E.8). 73 of files (10% of the text) were triply annotated; the remaining 1,387 files (90% of the text) were doubly annotated. There are six instances of the annotated corpus: one for each of the five annotators and one that aggregates all annotators' labels. Participatory action research with archivists led the project to choose four metadata fields were chosen in the archival catalog to extract for annotation: "Title," "Scope and Contents," "Biographical / Historical," and "Processing Information."

The five annotated datasets were merged[1] into a single aggregated dataset

[1]The code documenting the merging of the five individual annotator datasets, and the datasets themselves, is available at: github.com/thegoose20/annot/tree/main/notebooks/aggregating_data.

for classifier training and evaluation, so comparisons could be made on classifiers' performances after training on an individual annotator's dataset versus on the aggregated dataset. The merging process began with a one-hour manual review of each annotator's labels to identify patterns and common mistakes in their labeling, which informed the subsequent steps for merging the five annotated datasets.

The second step of the merging process was to manually review the 97,861 disagreeing labels, defined as annotations with the same or overlapping text spans with different labels, and determine which labels to add to the aggregated dataset. Disagreeing labels for the same text span were reviewed for all ***Person Name***, ***Linguistic***, and ***Contextual*** categories of labels. For ***Person Name*** and ***Linguistic*** labels, where three annotators labeled the same span of text, majority voting determined the correct label: if two out of the three annotators used one label and the other annotator used a different label, the label used by the two annotators was deemed correct and added to the aggregated dataset. For ***Contextual*** labels, unless an obvious mistake was made, the union of all three annotators' labels was included in the aggregated dataset.

Thirdly, the *Occupation* and *Gendered Pronoun* labels were reviewed. A unique list of the text spans with these labels was generated and incorrect text spans were removed from this list. The *Occupation* and *Gendered Pronoun* labels in the annotated datasets with text spans in the unique lists of valid text spans were added to the aggregated dataset. Fourthly, the remaining ***Linguistic*** labels (*Gendered Pronoun*, *Gendered Role*, and *Generalization*) not deemed incorrect in the annotated datasets were added to the aggregated dataset. Due to common mistakes in annotating ***Person Name*** labels with one annotator, only data from the other two annotators who annotated with ***Person Name*** labels was added to the aggregated dataset.

Fifthly, the 100,659 agreeing annotations, defined as annotations with the same or overlapping text spans and the same label, were reviewed. Among the 2,327 overlapping annotations, the annotation with the longest text span in each group of overlaps was automatically chosen as correct and added to the aggregated dataset. The 98,332 matching annotations were automatically chosen as correct and added to the aggregated dataset. The sixth and final step to constructing the aggregated dataset was to take the union of the remaining

Contextual labels (*Stereotype*, *Omission*, *Occupation*, and *Empowering*) not deemed incorrect in the three annotated datasets with these labels and add them to the aggregated dataset.

E.2 Language Variety

The metadata descriptions extracted from the HC Archives' catalog are written primarily in British English, with the occasional word in another language such as French or Latin.

E.3 Producer Demographic

The producing research team are of American, German, and Scots nationalities, and are three women and one man. We all work primarily as academic researchers in the disciplines of Natural Language Processing, Data Science, Data Visualization, Human-Computer Interaction, Digital Humanities, and Digital Cultural Heritage. Additionally, one of us is audited an online course on feminist and social justice studies.

E.4 Annotator Demographic

The five annotators are of American and European nationalities and identify as women. Four annotators were hired by the lead annotator for their experience in gender studies and archives. The four annotators worked 72 hours each over eight weeks in 2022, receiving £1,333.44 each (£18.52 per hour). As lead annotator, I completed the work for this Ph.D. research, which totaled to 86 hours of work over 16 weeks.

E.5 Speech or Publication Situation

The archival metadata descriptions describe material about a range of topics, such as teaching, research, town planning, music, and religion. The materials described also vary, from letters and journals to photographs and audio recordings. The descriptions in this project's dataset with a known date (which

describe 38.5% of the HC Archives' records) were written from 1896 through 2020.

The annotated dataset will be published with a forthcoming paper detailing the methodology and theoretical framework that guided the development of the annotation taxonomy and the annotation process, accompanied by analysis of patterns and outliers in the annotated dataset.

E.6 Data Characteristics

The datasets were organized for annotation in a web-based annotation platform, the brat rapid annotation tool Stenetorp et al., 2012. Consequently, the data formats conform to the brat formats: plain text files that end in '.txt' contain the original text and plain text files that end in '.ann' contain the annotations. The annotation files include the starting and ending text span of a label, the actual text contained in that span, the label name, and any notes annotators recorded about the rationale for applying the label they did. The names of all the files consist of the name of the *fonds* (the archival term for a collection) and a number indicating the starting line number of the descriptions. Descriptions from a single fonds were split across files so that no file contained more than 100 lines, because brat could not handle the extensive length of certain fonds' descriptions. Table E.1 displays the total number of annotations per label, from the Taxonomy of Gendered and Gender Biased Language (§E.8), in dataset.

E.7 Data Quality

A subset of annotations were applied automatically with a grep script and then corrected during the manual annotation process. All three categories of the annotation taxonomy were manually applied by the annotators. The lead annotator then manually checked the labels for accuracy. That being said, due to time constraints, mistakes are likely to remain in the application of labels (for example, the starting letter may be missing from a labeled text span or a punctuation mark may have accidentally been included in a labeled text span).

category	label	total
Linguistic	Gendered Pronoun	3682
Linguistic	Gendered Role	2785
Linguistic	Generalization	1293
Person Name	Feminine	1655
Person Name	Masculine	5586
Person Name	Unknown	10511
Contextual	Occupation	2958
Contextual	Omission	5597
Contextual	Stereotype	1279

Table E.1: **Annotation Totals in the Aggregated Dataset.** The total annotations per label from the Taxonomy of Gendered and Gender Biased Language in the final aggregated dataset. The "category" column refers to the Taxonomy's category, the "label" column refers to the label annotators' gave a text span, and the "total" column refers to the total number of annotations with the associated label.

E.8 Other: Annotation Schema

The detailed schema that guided the annotation process, the Taxonomy of Gendered and Gender Biased Language (Havens, Terras, et al., 2022), is listed below with examples for each label. In each example, the labeled text is underlined. All examples are taken from the dataset except for labels 1.1, *Non-binary*, and 3.4, *Empowering*, as the annotators did not find any text to which the provided label definitions applied. The annotation instructions permitted labels to overlap as each annotator saw fit, and asked annotators to read and annotate from their contemporary perspective. The categories of labels from the annotation taxonomy were divided among annotators: two hired annotators labeled with categories 1 and 2, two hired annotators labeled with category 3, and the lead annotator labeled with all categories.

The annotation taxonomy includes labels for *gendered* language, rather than only explicitly gender-biased language, because measuring the use of gendered words across an entire archives' collection provides information about gender bias at the overall collections' level. For example, using a gendered pronoun such as "he" is not inherently biased, but if the use of this masculine gendered pronoun far outnumbers the use of other gendered pronouns in our dataset, we can observe that the masculine is over-represented, indicating a masculine

bias in the HC Archives' collections overall. Labeling gender-biased language focuses on the individual description level. For example, the stereotype of a wife playing only or primarily a supporting role to her husband comes through in the following description:

> *Jewel took an active interest in her husband's work, accompanying him when he travelled, sitting on charitable committees, looking after missionary furlough houses and much more. She also wrote a preface to his Baptism and Conversion and a foreward [sic] to his A Reasoned Faith.* (Fonds Identifier: Coll-1036)

1. ***Person Name:*** the name of a person, including any pre-nominal titles (i.e., Professor, Mrs., Sir, Queen), when the person is the primary entity being described (rather than a location named after a person, for example)

    1.1 *Non-binary:** the pronouns or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are Non-binary

    Example 1.1: Francis McDonald went to the University of Edinburgh where they studied law.
    Note: the annotation process did not find suitable text on which to apply this label in the dataset.

    1.2 *Feminine:* the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are feminine

    Example 1.2: "Jewel took an active interest in her husband's work..." (Fonds Identifier: Coll-1036)

    1.3 *Masculine:* the pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are masculine

    Example 1.3: "Martin Luther, the man and his work." (Fonds Identifier: BAI)

1.4 *Unknown:* any pronouns, titles, or roles of the named person within the descriptive field in which this instance of the name appears (either Title, Scope and Contents, Biographical / Historical, or Processing Information) are gender neutral, or no such pronouns or roles are provided within the descriptive field

Example 1.4: "Testimonials and additional testimonials in favour of Niecks, candidacy for the Chair of Music, 1891" (Fonds Identifier: Coll-1086)

2. ***Linguistic:*** gender marked in the way a word, phrase or sentence references a person or people, assigning them a specific gender that does not account for all genders possible for that person or people

2.1 *Generalization:* use of a gender-specific term (i.e. roles, titles) to refer to a group of people that could identify as more than the specified gender

Example 2.1: "His classes included Anatomy, Practical Anatomy, ... Midwifery and Diseases of Women, Therapeutics, Neurology, ... Public Health, and Diseases of the Skin." (Fonds Identifier: Coll-1118)

2.2 *Gendered Role:* use of a title or word denoting a person's role that marks either a Non-binary, feminine, or masculine gender

Example 2.2: "New map of Scotland for Ladies Needlework, 1797" (Fonds Identifier: Coll-1111)

2.3 *Gendered Pronoun:* explicitly marking the gender of a person or people through the use of pronouns (e.g., he, him, himself, his, her, herself, and she)

Example 2.3: "He obtained surgical qualifications from Edinburgh University in 1873 ([M.B.])." (Fonds Identifier: Coll-1096)

3. ***Contextual:*** expectations about a gender or genders that comes from knowledge about the time and place in which language is used, rather than from linguistic patterns alone (i.e., sentence structure or word choice)

3.1 *Stereotype:* a word, phrase, or sentence that communicates an expectation of a person or group of people's behaviors or preferences that

does not reflect the reality of all their possible behaviors or preferences; or a word, phrase, or sentence that focuses on a particular aspect of a person that doesn't represent that person holistically

Example 3.1: "The engraving depicts a walking figure (female) set against sunlight, and holding/releasing a bird." (Fonds Identifier: Coll-1116)

3.2 *Omission:* focusing on the presence, responsibility, or contribution of a single gender in a situation in which more than one gender has a presence, responsibility or contribution; or defining one person's identity in terms of their relation to another person

Example 3.2: "This group portrait of Laurencin, Apollinaire, and Picasso and his mistress became the theme of a larger version in 1909 entitled *Apollinaire and his friends*." (Fonds Identifier: Coll-1090).

3.3 *Occupation:* a word or phrase that refers to a person or people's job title (singular or plural) for which the person or people received payment; do not annotate occupations used as a pre-nominal title (for example, "Colonel Sir Thomas Francis Fremantle" should not have an occupation label)

Example 3.3: "He became a surgeon with the Indian Medical Service." (Fonds Identifier: Coll-1096).

3.4 *Empowering:* reclaiming derogatory words or phrases to empower a minoritized person or people

Example 3.4: a person describing themself as queer in a self-affirming, positive manner
*Note: the annotation process did not find enough text on which to apply this label in the dataset to include it when training a classifier. One annotator used the label according to a different definition.***

*The *Non-binary* label was not used by the annotators. That being said, this does not mean there were not people who would identify as Non-binary represented in the text of the annotation corpus. When relying only on descriptions written by people other than those represented in the descriptions, knowledge about people's gender identity remains incomplete Shopland, 2020. Additional linguistic research informed by a knowledge of terminology

for the relevant time period may identify people who were likely to identify as Non-binary in the corpus of archival metadata descriptions. For example,Shopland (2020) finds that focusing on actions that people were described doing can help to locate people of minoritized genders (and sexualities) in historical texts, but also cautions researchers against assuming too much. A full understanding of a person's gender often remains unattainable from the documentation that exists about them.

**One annotator used the *Empowering* label in the following instances:

- When a person referenced with feminine terms was described as the active party in marriage
- Honor or achievement held by a woman (as indicated in the text)

  Note: Honors and achievements held by men were labeled as stereotypes, as there was a consistent focus on this type of detail about people, which involved spheres of life historically dominated by men in the UK. Spheres of life historically dominated by women in the UK were described with greater vagueness, eliminating the possibility of honors or achievements in these spheres to be identified.
- The fate of a wife is mentioned in an entry predominantly about the life of a husband
- Family members referenced with feminine terms are prioritized (i.e. they are listed first or more detail is given about them than those referenced with masculine terms)
- A gender-neutral term is used instead of gendered term

All annotators were encouraged to use the annotation tool's notes field to record their rationale for particular label choices, especially for text labeled with *Generalization*, *Stereotype*, or *Omission*. The work intends these notes to lend transparency to the annotation process, providing anyone who wishes to use the data with insight onto the annotator's mindset when labeling the archival documentation.

E.9 Provenance Appendix

The data documented above is an aggregation of five manually-annotated datasets. The aggregated and disaggregated datasets can be downloaded from the University of Edinburgh's DataShare platform at: https://doi.org/10.7488/ds/7540.

Please refer to the data statement in Appendix A for the documentation of the corpus of archival metadata descriptions extracted from the HC Archives catalog.

Appendix F

Inter-Annotator Agreement Detail

The tables on the following pages display Inter-Annotator Agreement (IAA) measures among annotators, as well as agreement measures between annotators and the aggregated dataset. The reported metrics include True Positive (TP), False Positive (FP), and False Negative (FN) counts, and precision (prec.), recall, and F_1 scores. The metrics are calculated between the annotations listed in the expected (exp.) and predicted (pred.) columns; either an annotator (0, 1, 2, 3, or 4) or the aggregated dataset (Agg). The "files" column reports the total number of ".txt" files of archival metadata descriptions that were compared to make the agreement calculations.

exp.	pred.	label	TP	FP	FN	prec.	recall	F_1	files
0	1	Unknown	5031	1524	4268	0.768	0.541	0.634	584
0	2	Unknown	2776	537	432	0.838	0.865	0.851	170
1	2	Unknown	1048	1421	315	0.424	0.769	0.547	72
0	1	Masculine	2367	2372	1079	0.499	0.687	0.578	584
0	2	Masculine	728	111	146	0.868	0.833	0.850	170
1	2	Masculine	380	169	411	0.692	0.480	0.567	72
0	1	Feminine	627	427	642	0.595	0.494	0.540	584
0	2	Feminine	724	128	178	0.850	0.803	0.826	170
1	2	Feminine	287	496	279	0.367	0.507	0.426	72
0	1	Non-binary	0	0	0	-	-	-	584
0	2	Non-binary	0	0	0	-	-	-	170
1	2	Non-binary	0	0	0	-	-	-	72
0	1	Gendered Role	1802	306	882	0.854	0.671	0.752	584
0	2	Gendered Role	1404	162	257	0.897	0.84527	0.870	170
1	2	Gendered Role	438	292	52	0.600	0.894	0.718	72
0	1	Gendered Pronoun	3398	101	190	0.971	0.947	0.959	584
0	2	Gendered Pronoun	869	70	60	0.925	0.935	0.930	170
1	2	Gendered Pronoun	518	7	11	0.987	0.979	0.983	72
0	1	Generalization	37	35	262	0.514	0.124	0.199	584
0	2	Generalization	74	51	63	0.592	0.540	0.565	170
1	2	Generalization	2	50	7	0.0385	0.222	0.066	72

Table F.1: ***Person Name* and *Linguistic* IAA.** Inter-annotator agreement measures for annotators who used the ***Person Name*** and ***Linguistic*** categories of labels to annotate archival documentation. No annotators applied *Non-binary*.

exp.	pred.	label	TP	FP	FN	prec.	recall	F_1	files
Agg	0	Unknown	10561	36	1900	0.997	0.848	0.916	714
Agg	1	Unknown	6608	0	4511	1.000	0.594	0.746	597
Agg	2	Unknown	15140	117	679	0.992	0.957	0.974	444
Agg	0	Masculine	3963	18	2446	0.995	0.618	0.763	714
Agg	1	Masculine	4749	1	1099	1.000	0.812	0.896	597
Agg	2	Masculine	1007	5	525	0.995	0.657	0.792	444
Agg	0	Feminine	1454	19	523	0.987	0.735	0.843	714
Agg	1	Feminine	1076	0	707	1.000	0.603	0.753	597
Agg	2	Feminine	994	12	410	0.988	0.708	0.824	444
Agg	0	Non-binary	0	0	0	-	-	-	714
Agg	1	Non-binary	0	0	0	-	-	-	597
Agg	2	Non-binary	0	0	0	-	-	-	444
Agg	0	Gendered Role	3108	697	330	0.817	0.904	0.858	714
Agg	1	Gendered Role	1924	218	716	0.898	0.729	0.805	597
Agg	2	Gendered Role	1471	652	230	0.693	0.865	0.769	444
Agg	0	Gendered Pronoun	3933	160	165	0.961	0.960	0.960	714
Agg	1	Gendered Pronoun	3498	3	190	0.999	0.948	0.973	597
Agg	2	Gendered Pronoun	1016	1	41	0.999	0.961	0.979	444
Agg	0	Generalization	405	1	1370	0.998	0.228	0.371	714
Agg	1	Generalization	69	4	1123	0.945	0.058	0.109	597
Agg	2	Generalization	127	0	862	1.000	0.128	0.228	444

Table F.2: ***Person Name* and *Linguistic* annotator agreement with aggregated data**. Agreement between the aggregated dataset and annotators for the ***Person Name*** and ***Linguistic*** categories of labels to annotate archival documentation. No annotators applied *Non-binary*.

exp.	pred.	label	TP	FP	FN	prec.	recall	F_1	files
Agg	0	Occupation	2725	23	571	0.992	0.827	0.902	631
Agg	3	Occupation	2320	290	873	0.889	0.727	0.780	508
Agg	4	Occupation	1746	147	253	0.922	0.873	0.897	450
Agg	0	Omission	5916	12	1187	0.998	0.833	0.908	631
Agg	3	Omission	2310	13	3475	0.994	0.399	0.570	508
Agg	4	Omission	1876	5	967	0.997	0.660	0.794	450
Agg	0	Stereotype	1748	11	1058	0.994	0.623	0.766	631
Agg	3	Stereotype	1089	9	279	0.992	0.796	0.883	508
Agg	4	Stereotype	1400	2	715	0.999	0.662	0.796	450
Agg	0	Empowering	0	0	0	-	-	-	631
Agg	3	Empowering	0	80	0	0.000	-	0.000	508
Agg	4	Empowering	0	0	0	-	-	-	450

Table F.3: ***Contextual* annotator agreement with aggregated data.** Agreement between the aggregated dataset and annotators for the ***Contextual*** category of labels to annotate archival metadata descriptions. Only Annotator 3 applied *Empowering*.

exp.	pred.	label	TP	FP	FN	prec.	recall	F_1	files
0	3	Occupation	1988	613	724	0.764	0.733	0.74835	485
0	4	Occupation	738	396	240	0.651	0.755	0.699	149
3	4	Occupation	422	327	134	0.563	0.759	0.647	57
0	3	Omission	1376	914	3259	0.601	0.297	0.397	485
0	4	Omission	416	317	875	0.568	0.322	0.411	149
3	4	Omission	215	315	155	0.406	0.581	0.478	57
0	3	Stereotype	505	539	227	0.484	0.690	0.569	485
0	4	Stereotype	507	525	600	0.491	0.458	0.474	149
3	4	Stereotype	34	60	161	0.362	0.174	0.235	57
0	3	Empowering	0	80	0	0.000	-	0.000	485
0	4	Empowering	0	0	0	-	-	-	149
3	4	Empowering	0	0	80	-	0.000	0.000	57

Table F.4: ***Contextual* IAA.** IAA measures for annotators who used the ***Contextual*** category of labels to annotate archival metadata descriptions. Only Annotator 3 applied *Empowering*.

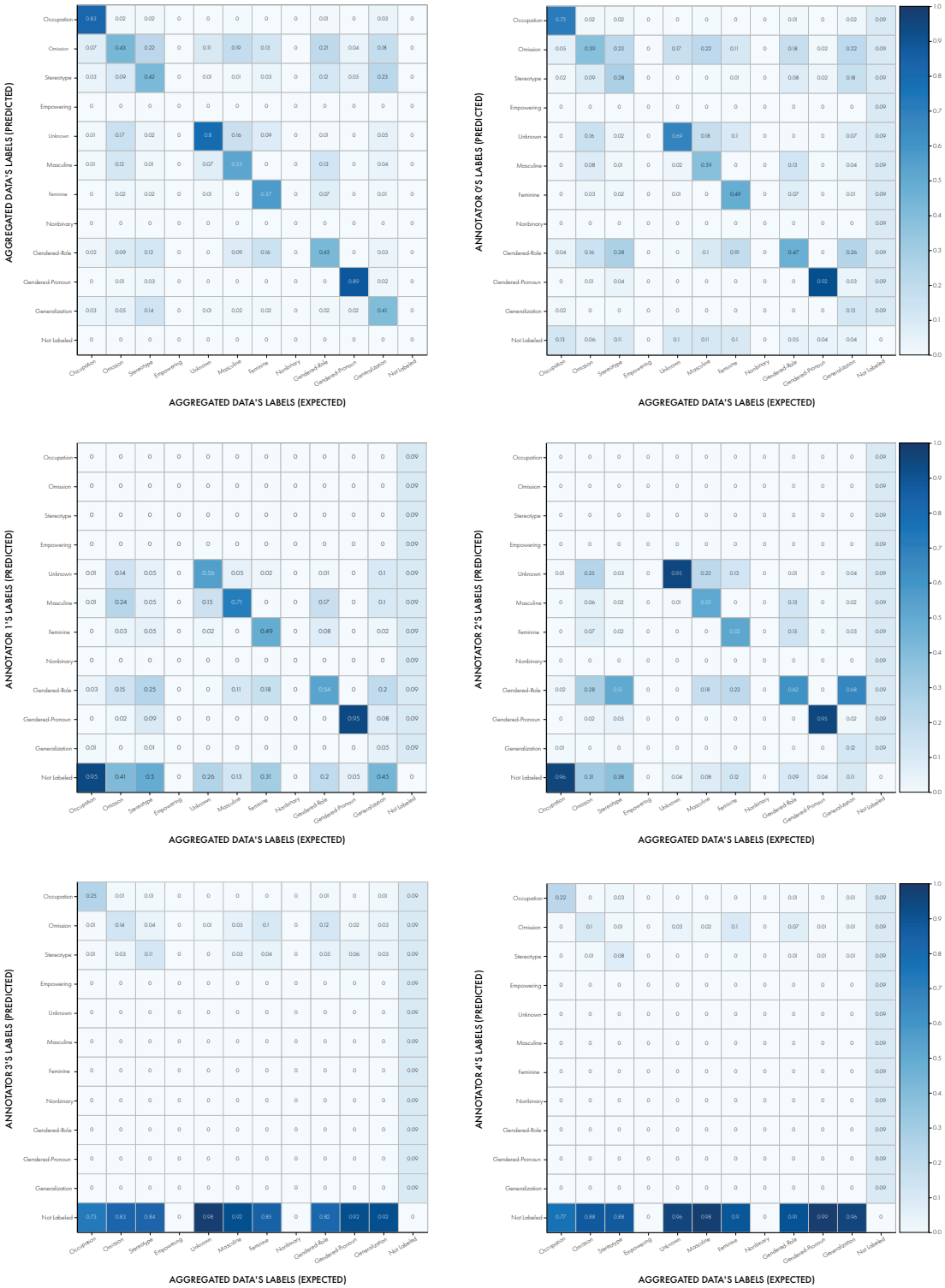


Figure F.1: **IAA confusion matrices.** Confusion matrices normalized with a weighted average on the aggregated data's labels, so class imbalances are taken into account. The top left matrix displays intersections between the aggregated datasets labels, illustrating where the same text spans have more than one label. The remaining matrices display agreement between an annotator (Y axis) and the aggregated data (X axis). All matrices have the same Y axis scale.

Appendix G

Classification Experiments Detail

Tables report False Negatives (FN), False Positives (FP), and True Positives (TP); and macro, micro, and per label precision, recall, and F_1 scores.

G.1 Linguistic Classification

Scores are reported for multilabel token classification of ***Linguistic*** labels (*Gendered Pronoun*, *Gendered Role*, *Generalization*).

Word Representation Experiments Detail

label	FN	FP	TP	precision	recall	F_1
Gendered Pronoun	84	93	675	0.879	0.889	0.884
Gendered Role	340	295	353	0.545	0.509	0.526
Generalization	275	298	110	0.270	0.286	0.277
macro				0.564	0.561	0.563
micro				0.624	0.619	0.622

Table G.1: **Performance of Classifier Chain model with Random Forest (`random_state = 22`) and without word embeddings.**

label	FN	FP	TP	precision	recall	F_1
Gendered Pronoun	24	178	735	0.805	0.968	0.879
Gendered Role	258	161	435	0.730	0.628	0.675
Generalization	319	44	66	0.600	0.171	0.267
macro				0.712	0.589	0.607
micro				0.763	0.673	0.715

Table G.2: **Performance of Classifier Chain model with Random Forest (`random_state = 22`) and with custom fastText word embeddings of 100 dimensions.**

Algorithm Experiments Detail

label	FN	FP	TP	precision	recall	F_1
Gendered Pronoun	321	109	438	0.801	0.577	0.671
Gendered Role	514	189	179	0.486	0.258	0.337
Generalization	384	86	1	0.011	0.003	0.004
macro				0.433	0.279	0.337
micro				0.617	0.336	0.435

Table G.3: **Performance of Classifier Chain model with Passive Aggressive and 100-dimension custom fastText embeddings.**

See Table G.2 for the Classifier Chain model with the Random Forest and 100-dimension custom fastText embeddings.

G.2 Person Name and Occupation Classification

Scores are reported for multiclass sequence classification of ***Person Name*** (*Feminine*, *Masculine*, *Unknown*) and *Occupation* labels' tags.

Word Representation Experiment Detail

tag	FN	FP	TP	precision	recall	F_1
B-Feminine	52	51	358	0.875	0.873	0.874
I-Feminine	121	50	810	0.942	0.870	0.905
B-Masculine	349	155	646	0.806	0.649	0.719
I-Masculine	475	267	898	0.771	0.654	0.708
B-Unknown	375	219	1642	0.882	0.814	0.847
I-Unknown	590	309	2746	0.899	0.823	0.859
B-Occupation	28	21	722	0.972	0.963	0.967
I-Occupation	34	21	792	0.974	0.959	0.966
macro				0.890	0.826	0.856
micro				0.887	0.810	0.847

Table G.4: **CRF model performance with AROW (`variance = 1`, `max_iterations = 50`) and without word embeddings.**

tag	FN	FP	TP	precision	recall	F_1
B-Feminine	44	48	474	0.908	0.915	0.912
I-Feminine	124	51	990	0.951	0.889	0.919
I-Masculine	441	282	1392	0.832	0.759	0.794
B-Masculine	296	179	1042	0.853	0.779	0.814
B-Unknown	404	205	1994	0.907	0.832	0.868
I-Unknown	644	295	3246	0.917	0.834	0.874
B-Occupation	22	29	738	0.962	0.971	0.967
I-Occupation	35	25	846	0.971	0.960	0.966
macro				0.913	0.867	0.889
micro				0.906	0.842	0.873

Table G.5: **CRF model performance with AROW (`variance = 1`, `max_iterations = 50`) and with 100-dimension fastText word embeddings.**

Algorithm Experiments Detail

The suggested parameter values for algorithms provided in sklearn-crfsuite (Korobov, 2015) were experimented with to determine which algorithm and parameter combination would yield the highest-scoring CRF model for the multiclass sequence classification of ***Person Name*** (*Unknown*, *Feminine*, and *Masculine*) and *Occupation* labels. Empty cells indicate that a parameter was not applicable for that row's algorithm. Macro and micro precision (prec.), recall, and F_1 scores are reported for `("B-[LABELNAME]"` and `("I-[LABELNAME]"` tags. Tokens are represented using 50-dimension custom fastText embeddings. The CRF model using AROW with `variance = 1` yielded the highest performance when measured with macro and micro F_1 scores.

algorithm	c1	c2	pa_type	variance	macro prec.	macro recall	macro F_1	micro prec.	micro recall	micro F_1
lbfgs	0				0.404	0.071	0.118	0.436	0.070	0.120
lbfgs	0.1				0.648	0.305	0.402	0.621	0.295	0.400
lbfgs	0.1	0.2			0.657	0.404	0.497	0.639	0.384	0.480
l2sgd		1			0.681	0.261	0.349	0.322	**0.611**	0.322
l2sgd		0.2			0.662	0.245	0.343	0.629	0.218	0.324
ap					0.679	0.334	0.427	0.413	0.303	0.413
pa			0		0.695	0.378	0.478	0.662	0.371	0.475
pa			1		0.694	0.385	0.483	**0.663**	0.371	0.476
pa			2		**0.697**	0.376	0.477	0.661	0.369	0.474
arow				1	0.538	**0.504**	**0.513**	0.506	0.480	**0.492**
arow				0.5	0.505	0.479	0.490	0.497	0.438	0.466

Table G.6: **Multiclass sequence classification performance: 11 CRF models.**

G.3 All Labels' Classification

Document Classification Model Experiments

Each document is a description from HC Archives documentation represented as a TF-IDF matrix.

label	FN	FP	TP	precision	recall	F_1
Feminine	104	5	24	0.828	0.188	0.306
Masculine	188	41	335	0.891	0.641	0.745
Unknown	467	350	1976	0.850	0.809	0.829
Generalization	161	25	107	0.811	0.399	0.535
Gendered Pronoun	108	13	147	0.919	0.576	0.708
Gendered Role	212	19	170	0.899	0.445	0.595
Occupation	274	10	134	0.931	0.328	0.486
Omission	409	61	395	0.866	0.491	0.627
Stereotype	124	6	191	0.970	0.606	0.746
macro				0.885	0.498	0.620
micro				0.868	0.630	0.730

Table G.7: **Multilabel document classification with Logistic Regression.**

labels	FN	FP	TP	precision	recall	F_1
Feminine	109	2	19	0.905	0.148	0.255
Masculine	214	27	309	0.920	0.591	0.719
Unknown	471	353	1972	0.848	0.807	0.827
Generalization	187	10	81	0.890	0.302	0.451
Gendered Pronoun	122	9	133	0.937	0.522	0.670
Gendered Role	250	8	132	0.943	0.346	0.506
Occupation	266	13	142	0.916	0.348	0.504
Omission	427	40	377	0.904	0.469	0.618
Stereotype	142	3	173	0.983	0.549	0.705
macro				0.916	0.454	0.584
micro				0.878	0.604	0.716

Table G.8: **Multilabel document classification with Random Forest (`random_state = 22`).**

labels	FN	FP	TP	precision	recall	F_1
Feminine	83	12	45	0.789	0.352	0.486
Masculine	112	61	411	0.871	0.786	0.826
Unknown	378	394	2065	0.840	0.845	0.843
Generalization	139	23	129	0.849	0.481	0.614
Gendered Pronoun	46	16	209	0.929	0.820	0.871
Gendered Role	132	30	250	0.893	0.654	0.755
Occupation	259	12	149	0.925	0.365	0.524
Omission	376	74	428	0.853	0.532	0.655
Stereotype	92	18	223	0.925	0.708	0.802
macro				0.875	0.616	0.709
micro				0.859	0.707	0.776

Table G.9: **Multilabel document classification with SVMs.**

G.4 Classification Model Cascades Detail

Reported scores are calculated strictly, meaning a model's annotation is only considered a TP if it exactly matches a manual annotation. All models represent tokens with 100-dimension fastText work embeddings.

label	**FN**	**FP**	**TP**	**precision**	**recall**	**F_1**
Gendered Pronoun	77	990	3654	0.787	0.979	0.873
Gendered Role	1064	862	2192	0.718	0.673	0.695
Generalization	1728	168	294	0.636	0.145	0.237
macro				0.714	0.599	0.601
micro				0.752	0.682	0.715

Table G.10: **Linguistic Classifier performance, strictly evaluated.** Multilabel token classifier performance with *Gendered Pronoun*, *Gendered Role*, and *Generalization* labels.

label	FN	FP	TP	precision	recall	F_1
B-Feminine	372	485	647	0.572	0.635	0.602
I-Feminine	874	825	1385	0.627	0.613	0.620
B-Masculine	1887	1548	1166	0.430	0.382	0.404
I-Masculine	2610	2744	1662	0.377	0.389	0.383
B-Unknown	4057	3360	5407	0.617	0.571	0.593
I-Unknown	6509	4864	8434	0.634	0.564	0.597
B-Occupation	1343	935	1563	0.626	0.538	0.578
I-Occupation	1920	1242	1518	0.550	0.442	0.490
macro				0.554	0.517	0.533
micro				0.576	0.527	0.550

Table G.11: **Baseline Person Name and Occupation Classifier performance, strictly evaluated.** Multiclass sequence classifier performance with *Feminine*, *Masculine*, *Unknown*, and *Occupation* tags.

label	FN	FP	TP	precision	recall	F_1
B-Feminine	397	585	453	0.436	0.533	0.480
I-Feminine	1035	1516	1138	0.429	0.524	0.472
B-Masculine	2176	1695	837	0.331	0.278	0.302
I-Masculine	3369	3812	1396	0.268	0.293	0.280
B-Unknown	4131	3183	3008	0.486	0.421	0.451
I-Unknown	7317	4876	5388	0.525	0.424	0.469
B-Occupation	1304	901	1407	0.610	0.519	0.561
I-Occupation	2054	1180	1266	0.518	0.381	0.439
macro				0.450	0.422	0.432
micro				0.456	0.406	0.430

Table G.12: **Person Name and Occupation Classifier performance with *Linguistic* labels, strictly evaluated.** Multiclass sequence classifier performance with *Feminine*, *Masculine*, *Unknown*, and *Occupation* tags.

Appendix H

Workshop Documents

H.1 Participant Information Sheet

Project title: Studying the Language of Descriptive Metadata for Cultural Heritage Collections
Principal investigator: Beatrice Alex
Researcher collecting data: Lucy Havens
Funder (if applicable): *Not applicable*
This study was certified according to the Informatics Research Ethics Process, RT number 2019/81479. Please take time to read the following information carefully. You should keep this page for your records.

Who are the researchers?
Beatrice Alex, Benjamin Bach, Lucy Havens, Melissa Terras

What is the purpose of the study?
The purpose of the study is to learn about the process of writing descriptive metadata and biases that may exist in descriptive metadata. The study is conducted as part of Lucy Havens' PhD research, which seeks to identify and classify bias in cultural heritage catalogues' metadata. Interviews, surveys, annotation tasks, and workshops conducted during the study will provide insight on types of bias and patterns in language that the PhD research could analyze (such as with Natural Language Processing) and display (such as with data visualization). Results from interviews, surveys, and workshops may inform the design of follow-up interviews, surveys, or workshops.

Why have I been asked to take part?

You have been asked to take part because the research target group is individuals who work at cultural heritage institutions.

Do I have to take part?

No – participation in this study is entirely up to you. You can withdraw from the study at any time, without giving a reason. Your rights will not be affected. If you wish to withdraw, contact the PI. We will stop using your data in any publications or presentations submitted after you have withdrawn consent. However, we will keep copies of your original consent and withdrawal request.

What will happen if I decide to take part?

If you take part in this research, you will participate in an interview, survey, or workshop about your cataloguing training process(es), your perception of cataloguing training and processes across the cultural heritage sector, and the presence of human biases in cultural heritage catalogues. Data gathered during your participation will be documented, analyzed, and may appear in research publications.

The method of documenting your participation will follow the preference you note in this study's Participant Consent Form. If, in question 3, you tick "Yes," that you consent to being audio recorded, an audio recording of your participation will be made. If, in question 3, you tick "No," that you do not consent to being audio recorded, the researcher will manually write notes to document your participation. If you produce documents during your participation, the researcher will collect or make copies of those documents. For example, if you make a poster in a workshop, the researcher may photograph the poster. Your personal information (including images of yourself) will never appear in the researcher's documentation of this study.

In all cases, data gathered during your participation will be anonymized. You will be assigned a unique participant number that will be used in publications should the researchers wish to reference your work or quote you. You will never be named or otherwise personally identified in a research publication produced from the study. After your participation in the study, the researcher will disseminate the results of the study to you.

Are there any risks associated with taking part?
There are no significant risks associated with participation.

Are there any benefits associated with taking part?
There are no significant benefits associated with participation in an interview, survey or workshop.

Participation in annotation tasks will result in compensation for the work through the form of a voucher, payment at an hourly rate, or co-authorship of a paper in which the annotation tasks will be reported. Compensation for annotation will be agreed upon between the participant and research team prior to any annotation tasks being undertaken.

What will happen to the results of this study?
The results of this study may be summarized in published articles, reports and presentations. Quotes or key findings will be anonymized: We will remove any information that could, in our assessment, allow anyone to identify you. With your consent, information can also be used for future research. Your data (from the interview and the consent form) may be archived for a minimum of 2 years.

The researcher will disseminate the findings of the study to you. Data protection and confidentiality.

Your data will be processed in accordance with Data Protection Law. All information collected about you will be kept strictly confidential. Your data will be referred to by a unique participant number rather than by name. Your data will only be viewed by the research team: Beatrice Alex, Benjamin Bach, Lucy Havens, and Melissa Terras.

All electronic data will be stored on a password-protected encrypted computer, on the School of Informatics' secure file servers, or on the University's secure encrypted cloud storage services (DataShare, ownCloud, or Sharepoint) and all paper records will be stored in a locked filing cabinet in the PI's office. Your consent information will be kept separately from your responses in order to minimize risk.

What are my data protection rights?
The University of Edinburgh is a Data Controller for the information you

provide. You have the right to access information held about you. Your right of access can be exercised in accordance Data Protection Law. You also have other rights including rights of correction, erasure and objection. For more details, including the right to lodge a complaint with the Information Commissioner's Office, please visit www.ico.org.uk. Questions, comments and requests about your personal data can also be sent to the University Data Protection Officer at dpo@ed.ac.uk.

Who can I contact?
If you have any further questions about the study, please contact the lead researcher, Lucy Havens, by email at: lucy.havens@ed.ac.uk. If you wish to make a complaint about the study, please contact inf-ethics@inf.ed.ac.uk. When you contact us, please provide the study title and detail the nature of your complaint.

Updated information.
If the research project changes in any way, an updated Participant Information Sheet will be made available on https://web.inf.ed.ac.uk/infweb/research/study-updates. In this situation the PI would also notify the Ethics panel who previously reviewed and approved this study.

Alternative formats.
To request this document in an alternative format, such as large print or on colored paper, please contact Lucy Havens at: lucy.havens@ed.ac.uk.

General information.
For general information on how we use your data, go to: edin.ac/privacy-research

H.2 Participant Consent Form

Project title: Studying the Language of Descriptive Metadata for Cultural Heritage Collections (Reference number: 2019/81479)
Principal investigator (PI): Beatrice Alex
Researcher: Lucy Havens

PI contact details: balex@staffmail.ed.ac.uk
Please tick yes or no for each of these statements.

1. I confirm that I have read and understood the Participant Information Sheet for the above study, that I have had the opportunity to ask questions, and that any questions I had were answered to my satisfaction.

   Yes / No

2. I understand that my participation is voluntary, and that I can withdraw at any time without giving a reason. Withdrawing will not affect any of my rights.

   Yes / No

3. I agree to being audio recorded.

   Yes / No

4. I consent to my anonymized data being used in academic publications and presentations.

   Yes / No

5. I understand that my anonymized data can be stored for a minimum of two years

   Yes / No

6. I allow my data to be used in future ethically approved research.

   Yes / No

7. I agree to take part in this study.

   Yes / No

Name of person giving consent:
Date:
Signature:

Name of person taking consent:
Date:
Signature:

H.3 Metadata Bias Workshop Agenda

Date: April 20, 2023

Time: 2:00-4:00 PM

Description

This 2-hour workshop will seek feedback on the presentation of findings from a machine learning model trained to flag potentially gender biased language in archival metadata descriptions. The workshop will ask participants to review sample outputs from the model and explain to what extent they find the outputs useful, and what actions they would take upon being presented with the outputs.

Prior to the workshop, participants are asked to reflect on encounters they've had with biased language in catalogues or the material they describe. The workshop will close with a discussion of the variations of biases that surface when cataloguing collections, and potential computing tools that could be developed to support participants in their efforts to mitigate harms from biased language.

This workshop is being held as part of Lucy Havens' PhD research and has been approved by the Informatics Research Ethics Process (RT number 2019/81479). All participants will be asked to read and fill in the attached participant information sheet and consent form at the beginning of the workshop. All participants' anonymity will be maintained in research outputs based on this workshop.

Worksheets distributed to participants during the workshop are for participants' use however they choose (e.g. writing notes on, highlighting). The worksheets will be collected at the end of the workshop to inform Lucy Havens' reporting on the workshop.

The audio of the workshop will be recorded for the researchers to reference only. The audio recording will not be shared outside the research team (Lucy Havens, Ben Bach, Bea Alex, Melissa Terras).

Agenda

Welcome (2:00-2:15 PM):

Introduction to research, workshop aims
Questions
Participant information and consent

Activity 1 (2:15-3:00 PM):

Review and discuss example outputs from a machine learning model that flag potentially gender biased language in metadata descriptions from the Archives' catalog.
Questions:

1. *Do you agree or disagree with, or are you unsure about, the labels on the descriptions? Why?*

2. *What would you do with this information? What information is missing that you would need to respond to what you're seeing?*

Break (3:00-3:10 PM)

Activity 2 (3:10-3:55 PM):

Review and discuss example summary information about a machine learning model's findings across a subset of the Archives' catalog.
Questions:

1. *What do you understand from the charts in the dashboard? What questions do you have about the information you're seeing?*

2. *How would you use this information? What information is missing that you would need to respond to what you're seeing?*

Wrap up (3:55-4:00 PM)

Participant questions
Final thoughts
Thank you

H.4 Workshop Transcript

The following transcript was automatically generated through Zoom and then manually cleaned using an audio recording of the workshop. I kept time markers throughout the transcript for ease of reference with the audio recording. When one participants' voice was inaudible, I put "[...]" to indicate that that was not picked up by the microphone. When multiple participants were talking and their words could not be distinguished, I put "[multiple participants talking]." For every comment and question, participants are referenced numerically ("P1," "P2," "P3," etc.) where I could recognize their voices and by "P" when I could not recognize who was speaking; I reference myself as "LH."

START: 00:59.00

LH: Thank you all so much for coming. I'm really excited that this worked out to have scheduled. For those of you who are familiar with what I've been researching so far, I basically- I started my PhD about 3 years ago, and I've been interested in seeing if I can support the kind of collection review process, particularly around gender-biased language obviously bias is about more than just gender, but to make the problem a narrow enough scope for a PhD, I focus specifically on gender bias. And Rachel is not something that the Archives are already interested in, so it kind of fit in with some of the goals that you guys already had.

Essentially what I've been looking at is seeing if I can use language technology to try to automate the identification of language that potentially has gender biases in it, and that way help potentially with kind of the prioritization of collection reviews or thinking about like, which is our different kinds of applications for funding for particular projects, being able to support different aspects of what I think is now Heritage Collections, right?

P: Yeah.

LH: So the data I've been working with is from the Archives catalog, and I met with a group people earlier, some of whom are here today-.

[another participant enters]

So I first worked on figuring out what descriptions to extract. So you know the title, that was recommended to extract, the biographical and historical information, scope and contents information, and processing information. And so what I did was take all those individual descriptions. And then I, along with a few other Ph.D. students we, manually labeled instances of gendered language so things like gendered pronouns, for example, as well as gender biased language, so things like a stereotype. And the idea is that although there's something inherently wrong with using gender language, if you're looking at how like as a whole, that could be one way to measure potential gender imbalances in a catalog.

So then with those, with that manually annotated data, what you can do is use a process called supervised learning, which essentially is taking all those labels and the data together and feeding it into the machine learning model, and seeing if that model can learn to label language in the same way that the humans did. So if there are patterns and like the same words that end up being annotated or the same parts of speech, things like that. So what you'll see today in a couple of worksheets I'll pass around, and-

P1: Is it fair to say, talking about kind of ... with the focus on gender we have to talk about intersectionality as well and how it isn't just gender that you're finding that there are problems where, or issues where [...] lenses as well.

LH: Yeah. And thank you, because that's a really good point, I think as well because although I've been very as I said very narrowed into gender obviously bias is about more than that, so if there are other types of biases that come to mind, please feel free to talk about those too, or comment on how you think it's not being considered, or or how it could be considered in some of the different visual representations of data that that we'll talk through today. And one thing to note as well so the worksheets and the information on them are samples from the catalog. So I have from the manual annotation process, labeled about 20% of the catalog. And the idea is that if you can. if I can create these models and the rest can be automated. So you kind of minimize the amount of time and resources and people that are needed to do that work. So the information

you see at the moment it's about the first 20% of the catalogue as it was in April 2020. And then what i'll do from the end of the my Phd, at the end of September essentially, is to run these models over the entire catalog. And so what I'd like to do today is figure out how I can present what the models find in a way that's useful to all of you.

And so that's mainly what I'm interested in today is I'm trying to present different different outputs from the model that I can get at both the detailed views in the descriptions. And then secondly we'll look at more sort of higher level views of the catalog and I'm curious about what information you find useful or what might be missing, and what you would do with this information.

Any questions at this point?

Okay.

P3: I've got kind of a technical question. So how like, what format did you export the data out of the catalog to run it through the model?

LH: I, so I used the Open Archives Initiative – Protocol for Metadata Harvesting. Encoded Archival Description is the format as XML is how the data is presented in. And then I went through and looked at the, those 4 metadata fields that I named for all the different levels, like fonds, series, file, item, and just extracted the text and then I use the text to represent them for the models, so for models you use numeric representations called vectors. And so they, the meanings basically get represented in a way that allows the model to like, calculate a bunch of statistics. because the vectors are based on relationships between different words. So they, they represent the meaning in a way, by looking at what words surround a particular word.

P3: So how much work was it to get the data into the condition that you wanted it in for the model?

00:07:18.270

LH: So the initial extraction was actually pretty easy. I hadn't worked with xml a lot before, but it wasn't too hard just to pick out those specific descriptions because they're pretty cleanly labeled and he format is really consistent in

XML. The challenge was, kind of-for different labels I was looking at different models required slightly for input. So for some I was looking at words, some I was looking at sentences, and for some I was looking at the full description. So it was more the kind of going back and forth between those format that got a bit complicated.

P2: Yeah.

LH: Does that help?

P3: Yeah, so did you maintain a structure so that you could say, most of this bias occurs in this type of description, or it's sort of equally distributed across these types of metadata, like titles or whatever.

LH: Yeah, so I have for every description there's, a-so I just assigned a unique ID, and then I have the metadata field that it came from.

P3: Right okay. So you can calculate that. I was just wondering.

LH: Yeah, no, that's a good question. So there's different ways to roll up the data, too, so. And that's where I wasn't sure what would be most useful to you guys based on [...]

P1: Do you want to go around the room so we know who we all are?

LH: Yeah, thank you.

[introductions are made as to participant roles]
LH: All right, so let's take a look at this first worksheet REDACTED passed around. This is on the first side, an overview of what I call the Taxonomy of Gender and Gender Biased language. The bold words are the labels that I used to create this annotated data set that I'm then using the teach models to try to pick up on gendered and gender biased language. And so I put in a few examples, and try to explain how those are applied. If you have any questions, or if it's not clear, do let me know. On the back side of the paper are

three examples of just excerpts of descriptions with kind of colored overlays for how these labels would be applied. So a lot of the labels are that kind of an individual word level, excuse me, except for stereotype and omission, so those are 2 that were applied at the description level. So that's why the kind of color coding of that looks slightly different.

00:14:32.280

LH: So for the first part workshop if everybody could work individually for about 10 minutes, or, if it if it takes less time, that's fine, and just think and, maybe, if you could just note down on the worksheets, and I can, I can collect them at the end, if you agree or disagree with something, if something's confusing to you, if you think something's not really useful in the way that it's been defined in the Taxonomy, any sort of initial reactions basically to how the taxonomy is applied.

P7: Do the colors mean something?

LH: It's, yeah, so they are based on the different categories of labels so Person Name labels are in green, the Contextual labels are blue, and then the Linguistic labels are yellow, for the higher-level categories of the label Taxonomy.

P8: What was it? Linguistic yellow?

LH: Yeah.

P8: And the other ones?

LH: Linguistic, yellow; Contextual, blue; and then Person Names are green.

P7: And you want to see if we're happy with them? Or want to change them or?

LH: Yeah, and if you find them useful or not useful or confusing, or if you would have applied them differently.

P7: So [...] so you're going to have to talk me through it. See that first one there. So, Henry Duncan, I'd say that would be a Masculine name, right? But that's marked as being Unknown. Why?

LH: Yeah, so a good question. So the idea with the Unknown label comes largely from some of the limitations of language technology research in that often there's a lot of assumptions of gender based on someone's name or jobs, or sometimes if it's speech, the sound of somebody's voice. And so you- that arguably is just engineering stereotypes into the system. If someone's gender identity isn't actually known. And so with the Unknown label or with the Person Name labels overall, the way we applied them was, if there's a pronoun that's referring to a problem or a role that has a grammatical gender that is referring to a name, so like in this first example, "John Baillie" is later referred to as "he" then the name would get Masculine, or Feminine if it was "she." And if it was a neo pronoun or a singular "they," it would get Non-binary, although we didn't have examples of that in the data we annotated. But otherwise it would get an Unknown label.

P7: Right. So if the sentence had read, "John Baillie was a preacher who gave a biographical talk," that would then make it Unknown.

LH: Yeah.

P7: Hmm. I feel kind of weird, though with it, in that now I can't say what their genders are. How can I catalog- So if they have a wife, then, then they're Masculine. But how can we say who they are if we can't assume that?

00:18:46.570

P6: I think [...]

P7: Yeah, yeah, yeah, okay, okay.

LH: Yeah, I think figuring out how to how to apply the Person Name labels with something that during the annotation process I changed. I initially had it

within an entire collection, trying to keep track of if a name had been referred to by a grammatical gender. And then, because of the size of some collections it just started to get very unwieldy to keep track of whether someone was referred to by a grammatically gendered word or not. And so- and like I said it was motivated by some of the limitations that people have been critiquing with language technology approaches to gendering people. But I know from previous conversations with people in Heritage Collections like you guys that there's a lot of useful information that can be gathered from someone's gender, and that can indicate a lot about social relationships. And I think REDACTED you had talked about interviews-

P6: Yeah it's things like people being called by Mrs. MacDonald, if you lose Mrs., you lose all of their identity rather than enhancing their identity. I think that's, that's the biggest one. But if somewhere in the description if we know that they were Mrs. Ian MacDonald then that identifies them even further. Technically all issues of common [...] then you could actually, I think you'd have a better chance of placing someone, but if later on, someone says you know oh Maggie, you shouldn't have said that, then you might be able to add all that together, but it's, it's [...] if you remove titles all together, or you just make everybody Mx./Ms., then that's a problem because all you have all you have is Mx./Ms. MacDonald.

LH: So the idea that I had here with these labels wasn't to remove anything which I should have said, for me it was more about trying to be transparent. I'm very much of the opinion that everything is always good on the bias to a certain extent, and there are some things that you can try to even out. But I think because we all come to things with our own perspective, our own training-

P: We're assuming a lot.

LH: Yeah, there's always going to be bias. And so, rather than- Again, kind of in the language technology space, a lot of people try to remove bias, and just neutralize data and drawing on a lot of Archival Science literature, actually and other- and literature in the cultural heritage sector that's talking about how there's no truly neutral position and thinking back to feminist theories as

well, I've been thinking about how we could make the biases more transparent. So it's more about trying to highlight who is being represented, and where, like, if you have a Mrs. MacDonald, for example, emphasizing that like, there is a woman here, but we don't fully know her identity and, and can that maybe point to future research directions-

P: Yes.

LH: Something like that.

00:22:20.910

P7: In the library cataloging context, the original object, and the text of the original object has the primacy, and you are largely transcribing what is in front of you. Because that is actually with library books such as multiples, it's actually the only way you can be sure that you're describing the same thing. So that means that you get the names of the identities in whatever form they are presented. But it is then the, the separate fields in the database which are the access points which unpick it so you might well get, like, novels which are by Mrs. Mullsworth [...] but it is then somewhere else in in some cases, in a Note field, that you then explain that this was whatever her actual name was, and the, the rules for constructing the authority index names, it would actually go in under her own name.

P6: Really? So how would you do that?

P7: Well, well, unless you know you really, really, really don't know it. But if you actually, if you know that she was usually known as- I've forgotten what Mrs. Mullsworth's Christian name was but anyway if it was known that she was a, a, an Anne or whatever it was-

P: Margaret.

P7: Thank you! So she'll go into the catalog as Margaret Mullsworth. And somewhere there will be across reference from her maiden name as well. So it

should actually all join it up.

P: We did it that way with a painting, and it was known as Mr. and Mrs. REDACTED, and so we had his name and her, her name is Mrs. REDACTED [husband's name used in place of wife's name], and I said, can we please look at her? So we found the marriage certificate, and found out that actually it was the wealth that they'd acquired, and the reason that we had this picture in this collection of art was that she'd a very kind of rich dad. And so her her name is actually REDACTED [wife's name]. So we were able to to kind of say what her name was that I thought so, yes, in some context, contexts, it's so important to know who the husband was. That is absolute context.

But I think when you're looking for bias and gaps when the search are identifying, particularly where a Mrs. is, and saying that's potentially a partner that needs further exploration of the [...] really like interested in how many of them are in there. But my concerns are with this is that you know what REDACTED was saying, you know this is absolutely just copying exactly what the record says, I believe, the way that they've got like "sics" in there. So it's not clear, then, if he gave a biographical talk, is part of that literal copying, or whether it's an assumption that the archivist has made. So I can see that, that's a tricky one. And the same for example 3, but, example 3 is slightly different, because it looks like that's been done by a person much much earlier, rather than relative- because there is a, a spelling mistake there as well. So it's really hard to unpick.

00:25:42.840

P1: It's- as, as people were talking one of the things I noted down was about, it's what something students brought up with me a while ago about, it's about dating descriptions and not overwriting them, saying when descriptions were made, and does that actually help in terms of the content or nature of the language. That is, has been, that has been used, and by that, you know if we know something was a catalog, was, you know, I'm not saying it unpicks everything then that gives us a clue, you know, to think about what kind of context, what kind of environment that was created in. Because we know [...] cataloging collections in very different ways. So we've got all these shelf

marks and all these different ways of doing things, and now we've got loads of problems with descriptions and [...] another thing on top of that. But that idea about date and context [...] As you were saying, it's a way of investigating that then allows us to flag, and not solve it, but to say is this something for investigation? And if there's a way of pulling [...] that, [...] there might be some patterns that allow us to either consider projects or [...] or focused work that we have, priorities, help us think about that.

LH: So that was a time thing to one thing. I was reading. Also, there's a woman, Shopland, who talks about researching LGBTQIA+ history, and gives recommendations. They're practical for people trying to sift through this information and one of the things she talks about is that people need to be careful not to assign terms that are used today that wouldn't have existed when somebody else was alive.

Multiple participants: Yeah.

LH: And I feel like that's another challenge with the whole archive, and I mean any cultural heritage collections really, they cover such a big time span, the language is constantly evolving. I don't really know how you work with that.

[laughter]

P6: A good example of that in Scottish Studies. The Scottish government has recognized Scottish Gypsy Roma Travelers, and but we have, say, archives relating to travelers of various sort of ethnic origins, if you like, in Scotland, and somebody who is a traveler, from the traveler community got in touch with us, and they said I would like to change that, because I'm a Highland Traveler, and I'd like to change that to Scottish Gypsy Roma Traveler. Except that the person who's described as a Highland Traveler, that's how they described themselves in the recording. So you then get into layers of, well who's, who's say is it, is it the current community, or is it the person who's own voice it is? Because you might even have a traveler from back in fifties, calling himself Tinker, which is completely pejorative now, so who gets precedence, and when do you change it? Should you have the opportunity to change it? Or do you

have to reflect all of these things, and then you get to a clunky-looking, or difficult to read description. So I think it's, it's definitely an issue about time.

00:29:40.220

P9: I wonder if what you could do is just have some kind of, you know, if you have the, the way that the person described themselves at that time, and then in brackets, or as a sort of note, somehow saying, People of this community now self refer, as you know, something as little clunky as possible. We deal with it all the time in the sense of referring to place names that were part of the British Empire that are now called by different names by the people who live there now. So we put, ehm, Kolkata, and then in brackets, Calcutta, you know, when referring to that place, because otherwise we would be applying the colonial term. So I guess it's something that happens in different contexts.

P4: That's kind of what I've been trying to do is include both historical terms that were used, and the records were used by people who themselves identified, especially when it comes to things like the trans community, words that I would never use today,[...] but then also include preferred terms that will also then be used as search terms, because, although the people that I'm describing didn't refer to themselves as transgender, they would use something that I would now consider offensive, it's still something that's an example of, you know, trans history, so, or gender non-conformity. So, it's trying to reflect that. But you're right. It just becomes really clunky, and it looks unpleasant to the eye when you see it.

P6: But if it's for the reason, if you're trying to search it, you know, if you're trying avoid people having to search for well, offensive terms, for one, as well as terms that you'd use now, then, well [...]

P7: And what do you do when whatever it is you do today, in 20 years' time, something you created is deemed offensive?

P4: I know. Well that's, that's, really, that's just what is generally going to happen though, like, and that's when I think making everyone aware of when

something was cataloged is extremely important. Because someone probably will turn around in 20 years, and think that the terms I included, and are- and I try and use other sources and not just my own decisions to back why I've chosen something that is a preferred term it's just it's going to be a continuously evolving-

P: [...] use the catalog to search for a term to see how many times it comes up, which is, you know, really quite interesting in itself. So you know. So it helps, you see, exactly where REDACTED has used that term, how many times he's used it. So use, you know, having your out of date terms is, I would say, really important, because it shows it will be able to show in the future just how prevalent that term was, and within a certain time frame, and how that is changed just even just with a search. So, so I would say use as many words as you can, use the right words, and then have the explanation on it. I think for me here with these ones, example one, it's not clear where the literal transcription stops, and where the descriptive part starts, and that's maybe something that we can think about. So that would easily be solved there by for not saying "he" but "Baillie gave a biographical talk on Duncan's life and work," so we're you're not using- you're making a conscious decision not to use "he" or "her," just by thinking carefully about what you're trying to say. But the other two, it's not clear if there's any description work in there, or whether it's just a literal transcription of a title.

P3: I was going to ask, is it possible to distinguish where exactly in the catalog the language came from. Because I think, like you're saying, that will be very helpful to know, because if it's in titles, well we probably inherit or we certainly inherit that, it's not been created, so it'd be good to be able to isolate language that's been created recently by an archivist. And dated! Having versions of metadata would be really useful for lots of reasons. Yeah, that was most of my commentary is that these would be very helpful at scale. So like this, summarizing the interests of Florence Jewel Bailey who's Unknown to the model, but I'm pretty sure that's a woman, if it happens repeatedly that the interests assigned to a woman are summarized as general interest or whether that's just a one-off. Or if that happens, a lot, regardless of the gender. But that would only be useful like at scale over the entire catalog, as opposed to-

00:34:46.889

LH: I do remember this quite a bit because it was an example where, like here, all these letters were all kind of grouped together, and the description was kind of vague, and then later on there was a collection between 2 men who are academics, and every letter had a separate title entry, and I don't know, like I mean, we've talked about project funding, and I, I'm not saying that that is like inherently wrong, like someone was like trying to just skip over the woman, but it's this sort of- at scale when you're comparing different things, does this reinforce the kind of like, meh, that's the women's world.

P3: Yeah, one key thing I'm thinking is that when we're [...]

LH: [...] not that there was any sort of malicious intent, which I'm sure there wasn't in this case. Um, it- yeah, it's more of like, could there be, I don't know, can there be something more interesting that could be described there? If it, if the material was revisited [...]?

P1: We will always get back to having an army of catalogers that we need.

[laughter, multiple participants speaking]

P1: What's occurring to me as well, when you're talking about standards. I remember back in 2000, REDACTED, we were talking about authorities and authority entries and looking at what we're using for like gap analysis. I'm thinking about thesauri and which thesauri we use. You know- and, immediately I remember there being problems with some of the terms in the thesauri, and I, I as an archivist really pulled back away from using some of some of those things which was part of what we were looking at in the project, because I, I felt that they didn't fit. They just didn't fit, it was, you know, trying to shoehorn things into a standard way of doing things where it was all- the messy archival world is not like that. It, it doesn't- life does not conform, at all. So I, I have real issue around some of those thesauri that were really championed through some of those early projects. And my brain also went to

the Dewey Decimal System, as well.

P7: Well that's a classification scheme so that's something slightly different.

P1: Right, I know it's different but the kind of debates around how it's used, and then the thesauri, were in my head.

LH: Yeah, trying to think about like making sure that there is a standard so people can- collections can be interoperable and systems can be interoperable and stuff but, yeah-

P7: But Dewey is updated.

00:37:30.090

P1: Yes, there was a whole of debate there wasn't there? A few years ago, whether to go with updates, and then they had to-

P7: No no, they update Dewey regularly.

P1: Yeah. I'll, I'll find-

P7: Are you talking about like a call number? Or subject headings?

P1: No, no, definitely Dewey.

P7: Yeah. But they, they, Dewey is being regularly updated ever since it was invented. And they, there's actually a scheme built into it [...] for not reassigning numbers within a certain number of certain editions so that you don't finish up with different things at the same number. But they are relying on it being applied to the sort of libraries that turn stock over, you know, because it's really intended for public library collections which do turn things over.

LH: So, just-

P6: But can I ask one more question?

LH: Oh yeah, yeah.

P6: In example 2, it looked like Dietrich Bonhoeffer's name wasn't picked up-

[multiple participants talking]

LH: That's a mistake.

P7: It's also been misspelled.

P5: I have a question. So like for me, Stereotype, there is, how do you define Stereotype? How do you know something is Stereotype or not, and because it's changed by people, by culture, by things like this so it seems really hard to define what it is.

LH: Yeah. So with, with the way I've been talking about this in this research, I've been situating the work in the UK, saying the dataset I'm working with is from archives in Scotland, at the University of Edinburgh, so trying to kind of- and, primarily in in English- so trying to kind of narrow that down, and then with stereotype, the way that we defined it for this taxonomy is about, kind of a, a restricted, or kind of limiting description of someone's identity that's kind of overly simplified, and puts them in a, a more, narrow category, than is potentially possible for that particular person. Um. But it's difficult to define.

00:39:52.540

P: That is a tricky one then, because, I totally get that, but see example three, correspondence and related items, that's probably what it is. But I mean as a as a person who is cataloging, I would try and not do that. Like, correspondence, I really, so that's, I would say that that's less. Hmm. That's less stereotyping and more, slightly more lazy cataloging, um, because you know, "mixture of press cuttings covering many subjects," well, it's press cuttings of- include, or

including articles. So that's just making that awareness of being tighter in your description. I'm not sure-

LH: So I think in this case, at least something about how, how these words, or how these labels were defined, I think what this stereotype is more about here was that the woman was referred to as "his wife," rather than named. And yeah, it's, it's kind of, it's both an Omission because-

[multiple participants talking]

P: So the Stereotype is "his wife." Okay.

LH: And Omission because there is no name, and then Stereotype, because a woman is being defined in relationship to a man.

P: Okay, and so right. And so the Stereotype for example 2, is the mixture of press cuttings, or that they've, they've picked up house housekeeping? What's the-

P3: I thought it wasn't just the housekeeping but that a woman's interests just get summarized into general interests.

P: Right right right, okay.

P: So what's the Stereotype in the first [...] ?

P6: There isn't one, it's just not as but as it's just that housekeeping and texts of general interest.

P: Okay.

P3: It's very self-revealing to say what you think.

[laughter, multiple participants talking]

P7: These are from the same collection, are they? For each of these examples?

LH: Yeah, so [...]

[multiple participants talking]

P: Mrs. and wives.

P: Ms.

LH: So in the interest of time I just wanted to get to the second question, REDACTED you started to get at with what you were talking about, how-

[multiple participants talking]

42:24.850

LH: What would you do with this information, or is there something additional that you would need in order to do something with this information? Or could it be presented in a way that would be more useful?

P5: [...] a way to flag up something that might be problematic. But then you've got to have an archivist or cataloger go back and decide, actually, no that's fine, or actually, I wish I could do something with it but I don't know her name, so I think that's it, because sometimes it just seems it's not that they just omitted her name because he didn't want to or they think that she wasn't important it's just that most of the time, you don't know the name of the person.

P: Would this have picked it up if it had said Florence Jewel Baillie and her husband?

LH: It would have- it should have picked up an Omission in that case, but it wouldn't, I don't think it would pick up a Stereotype, because that's kind of an anti-stereotype.

[laughter, multiple participants talking]

LH: Yeah at least the way the model is taught it would have picked up on an omission, so there not being a name, and a person being defined in in terms of their relationship with someone else.

P5: I think it would be useful. And then it's also having the resources, or always, if it to go after that and change, but yeah, I think it'd be a useful thing to have just to flag it up. And then you're aware of it.

P1: I would agree in that sense, and I'm going to say, that my brain's working really hard, which means that [...] and I think that I love being in the space where we can sort of unpick things even if we disagree or we have different ideas, it's really, really healthy. But I think the, the flagging of it, gives us an evidence base of, you know, I'm thinking, going back to resource, it gives us an evidence base to say, look, we've got a need for cataloging in this space and this kind of particular skill, and it's not just about, you know, having anyone do it, can we get an intern in to do it or can we get a student in to do it. Actually there is a skill to this and these skills of archivists, librarians, curators are needed in this space and, as I said, we need an army of them.

P2: It seems like there's two things. There's- one is the existing, is the analysis of the existing, picking up where he's and she's- so you can kind of get a picture of how things are. But yeah, I'm just wondering in terms of my is this intended feed into a kind of "how to" going forward, or what would be useful, you know, because it it seems to me like, of course it's not useful to take out Mrs. MacDonald or, you know, those kind of things but it's like, that, that's kind of interesting for me. It's more like. Yeah. And for people who do, who do metadata work, which I don't really do much of. But it's yeah, is that the intention to kind of go towards something like that, because obviously that will change over time as well. But it's more like, it's, it's a set of rules, or best practice is not to put a gender, or would that actually be the opposite to be like where it's apparent, or where, where you can, can research it or I don't know.

00:46:01.230

LH: I think, I think you brought up some of I was thinking about description practices. And so that's where I was- kind of, I, like, that was sort of a question I had. I was interested in sort of hearing people's thoughts about would this sort of information be helpful in setting up guidelines, or-

[multiple participants talking]

P1: [...] be useful for particular collections, where, we're, we're taking a particular collection, where we have some knowledge about it and we're going to start to catalog it, you know, in a particular way, and we're aware of having language sensitivity around them, whether it's gender or something else. But it helps us- I wonder whether it, it helps us just to check in with ourselves as we do the description, or um, or something about the collection that we haven't really thought of. I, I mean- I've got- I need to think about it a little bit more, and I all of you will have opinions-

P8: I could definitely add it to like some of the volunteer, you know, yeah, I'm on getting the volunteer to look at the, the REDACTED stuff. So it's very, very male and, and, but also race and different controversial- difficult topics and challenges, and I'm quite good at flagging them to the students and saying, you know it's like a name to students saying, you know, come across what you you know. You need to be careful, but you know both in how you how you are with it, and also in recording it. But I have not talked about bias when it comes to and gender at all. So I need to go- it'd be interesting for me to go back and just see, how many "hes" they've put in there, you know. Probably quite a lot. So that they're starting to think about their descriptive skills on a range of different fronts. I think that would be quite useful to have.

P1: It'd be an interesting exercise in archive, library, and curatorial courses when you do those [...]on metadata and, and getting students to think about their cataloging practice. I remember when I was training- I always call it "archive school" like it's a [...]

[laughter]

P1: You know, but thinking about- have that reflective practice as part of our cataloging [...] it's so valuable, because you kind of wind up [...]

P: I think we never got that.

P3: I'm looking at the standards and I'm like, we have structure.

[laughter]

P3: REDACTED and I did our Archives course at the same time at the University of Glasgow and there's a lot of emphasis on, like, standards, and part of that is because job applications for archivists always want you to be familiar with these standards.

P1: And we recruited to that.

P: And interview to that effect as well.

P3: Yeah so it's not just the catalog, it's like the whole profession, and and like, when you're talking about things like the taxonomy and how that gets decided, um, who sits on those board. We're all women in this room, but it'd be interesting to see the gender balances with people who decide which lexicon we're using to describe-

P: Wives.

[laughter]

P3: Yeah.

P8: That's why I would go back and have a look at the REDACTED stuff. And I'm I'm thinking I'm going to answer your question. But is it relevant with them? So I, I want to know how many wives have come across, and then maybe done a lot, Googling around the guys, but have they googled the wives?

Probably there's not findable. That's, that's REDACTED's point as well.

P3: Mmhmm.

P: And that's really, really difficult. But unless we start, you know, working in that way, it's never going to change.

00:50:10.650

P2: Yeah, no, I was kind of wondering about the, the idea of the frequency of the use of the word like "he." If that's a problem, you know. You know, like what I mean, it's more- unless it's assumed that it's "he" based on the name, maybe that's the problem. I know, rather than having "he," I don't know.

P3: We hold the papers of a lot more men than we do women so-

P: And ours are more professionals.

P: Historically, so it might be sort of proportionate.

P9: I'm wondering about - and this is a really willy question – about the use of, or the non-use of pronouns like, for example, if we, if we're saying that we're not going to assume, I mean, I may be picking this up wrong, but we're saying we're not going to assume that Henry Duncan was a "he," is that actually helpful? Because we can probably assume that he was treated as male. Whether he actually, in our- identified as male or not, we probably don't know; but the fact that he presented as male, therefore, was referred to as Henry, and you know, and therefore held that space in society at that time, it doesn't help us at all to unpick that, you know. Is it useful to unpick that? Or-

P1: But I think that's where context is really important

[multiple participants talking]

P1: When you deploy this and when you don't, depending on the collection

knowledge and cataloging required, it's having that flexibility and any kind of tool, or, or model that's being trained. Like we choose to Transkribus on some things, and we choose not to use it on others. It's it's having that flexibility for where we want to analyze it a little bit.

P9: Cause the problem there also would be, if we then also got, many, I mean, and I'm maybe just drawing parallels where there's none. I'm just sort of speaking my thoughts that then we've it sort of creates a problem, because then we've got Elizabeth. Who else? Elizabeth the Second, who also comes up as unknown when in this case it's a woman in power. But who then we're also not recognizing that there were some women in power. If everybody's not got a pronoun, you know, I mean not identified by gender.

00:52:24.870

LH: And then you have the Stereotype label sort of assuming that Florence Jewel Baillie is a woman even though she's labeled as Unknown. So yeah.

P9: I I'm not sure. It's only that thing that I'm not sure about, you know, with the pronouns, and the.

P2: And yeah, I think going back to REDACTED's point about stereotypes. It's so true, because it could be like, well, she's also interested in the Second World War, and house- oh no housekeeping is there as well, I mean-

[multiple participants talking]

P5: It's also a Stereotype, it's true.

P: But also like, yeah.

P2: And also, yeah, like I suppose, yeah some of those spaces are being, reclaiming some sense of, some of those spaces are being reclaimed as being, I guess, good?

P7: Well actually it's a bit of an assumption to suppose that somebody called Florence was female, because there were, that can be a male name as well.

P3: Yeah, I had that thought too.

P7: You could be doing her a huge injustice if her life's mission was housekeeping.

Multiple participants: Mhmm. Yeah. Exactly.

P7: And I also have to put in a check for, ah, Primrose, which is also a male name. And they are shocked. You'd be surprised how many Primroses there are. But I think from me, going back to your question, Does it really matter? I'm not sure that much we can do about it other than maybe being slightly aware, and make our cataloging tighter.

P1: There's a big thing around awareness with this, isn't it? Actually, as REDACTED was saying, that reflectiveness and being aware, it- it isn't necessarily- Lucy might get to the point where, she goes, there is no answer to this.

[multiple participants talking, laughter]

LH: Yeah!

P1: She might! But, exploring it and actually being aware of our practice, or aware of, where we've taken old catalogs and perpetuated things. I really like REDACTED's point just now, with Occupation, "housekeeping" could be an Occupation but it's not, necessarily [...]

P10: And it's the same thing, you know, with "coronation," that's hinting at her occupation. And then I mean, looking at the other occupations, the "preacher" that is highlighted in it at the top. Are you assuming that preachers have to be men, depending on- I mean they don't, depending on your denomination.

LH: With the annotators it was debated, I think it, it was, it might have been "Pope," or something, that someone was noting that as a Gendered Role, and I was like, oh, I guess they've always been men, but I wouldn't have thought to put it-

P: No they haven't!

[multiple participants talking]

P10: [...] depends on what religion you're in.

P: Well that's like "preacher."

P: The Quakers, the Quakers have female preachers in the seventeenth century.

Multiple participants: Yeah.

LH: Well and I- and that's where I was like I wouldn't have noted that as a Gendered Role.

P1: But actually is a preacher an occupation or a calling or what?

[laughter, multiple participants talking]

LH: And I was just going to say, I think as well that I, I think I just really appreciate about the perspective in this group of people is that when you come from like more sort of data science, computer science training, there is no thought going on about, about, what is gender and what is bias. Like, papers are published, and people say that they've debiased something and they don't say how they define bias. And so I think what I- one thing that I'm trying to kind of bring into that world is the complexity of some of these things. So like you said, if there's no answer, that's still important to report. If this is just complicated and it needs to be dealt with on a case by case basis rather than there being a universal solution, that, to me, is really important to have evidence of. Because so often in machine learning and AI, things are

presented as the state of the art, and it works in all domains, and it's just not true, and there are a lot of problems that are starting to arise. But the majority of publications these days are still being presented as a big universal solution for this problem, no matter what. And, and so yeah, just, just to point out that sometimes I think there just isn't a solution, and or it's just going to depend on the specific collection or specific institution or, whatever.

P1: To give you a little bit– and then we'll have our break-

LH: Oh yeah.

00:57:10.250

P1: Being part of this network of Archives and AI, that's sort of run through funded research through universities, um – I think REDACTED you were involved in some of the Archives and AI stuff as well – and what was happening was a real, kind of, archivists need to get on board with AI. And I kept going, oh, okay. Because there's an assumption that the tools would just work that, that we will get to the point where it, it replicates what us as human beings do. And so that's why I really engaged in Lucy's research. Because yeah, we're engaging, and we're, we're exploring it for finding where it does and doesn't work, and allows us to challenge ourselves as archivists, librarians, curators. And question our practice. And question assumptions. And that's not something we were really, taught to do. We were taught to accept them and that's, well, I don't want to.

[laughter]

P1: I want to see them as a framework. I want to see what it, you know, maybe there's something else we could be doing in this space that would be helpful, to get you know, cataloging done, and you know that's not to overhaul cataloging systems of course or projects ... (inaudible) It's just to reflect and think. And have our time and space to do that as well.

P: I think reality is always very different from the standards, and I think it

wasn't fitting it doesn't fit it. It's the story on the on as much as much more complicated than what the framework's about, you know, and we try to make things fit and of- they often don't.

P: [...]

LH: It's like learning a language. I felt like when I was learning French it was learning all these rules and then in the third year it was like, and here are all the times you break those rules.

[laughter]

P9: I was going to say that I think it's sort of socially really important, the work that we're doing as archivists and as librarians, the way that things are being described, because I heard someone use this term recently that the- I think they were using it sort of critically, actually, that when you write a description of something for an archive like sort of an archival entry that that often the sort of the "God" voice is used. So the person writing it hasn't got a position, so there's not a positionality. If I write an academic paper, I'm saying, I'm this person, and this is my thoughts on this. But if I write something to go on to like be searched on us of library resource, I'm trying to use this kind of neutral voice, because I'm just saying this is just the way things are, you know, this is without any opinion. So that kind of default.

[multiple participants talking]

P9: Yeah, that kind of if we're saying that this is just the sort of neutral voice. So we we're sort of creating that, by the way we describe things, you know, that at this point in time this is the sort of assumed just kind of-
P: [...]

P9: Yeah, it sounds like you're trying to present thing as facts and facts are obviously always up for, for questioning. But it's a, it's a different tone in writing, so the way that we do it super important because so many people are going to look it up: students or members of the public, and refer to that

as a kind of non-biased sort of factual resource. So it's different from writing something where you assess your positionality and make that clear to the reader, you know. So it's it's a big responsibility to get right as well.

P: I would say that it, it's what we've seen, at that moment. That's what we have seen on that file in that record. And then you mentioned the processing information as well, I mean, you know where, who I was, and the kind of, date. So that, so that somebody else can come after you and can look at the same thing, and they might see something else, something different, something, that's not right. So then, you do have a conversation about changing it. So this is my, my concern about this first one is. I am not sure where the, eh, where the file title stops, and where, the where the cataloger's observation starts. And so, I've always tried to do that in my language, you know, and I use a lot of inverted, a, a lot of commas to say that this is what I'm looking at and looking at this.

P5: I think that's [...]use inverted commas to signal that you are quoting that directly as opposed to [...] on the file.

01:01:58.420

LH: I'm so sorry I'm going to interrupt. I want to make sure we get a short break. And then there's a different worksheet actually with different information that we'll go through after the break.

BREAK

01:11:08.800

P1: Lucy just said some really interesting things and I said, you have to tell everybody you have to tell everybody.

LH: Yeah. So, so one thing I was saying, that every time- that comes up every time I have conversations with you from Heritage Collections is the sense of responsibility that you feel. And it to me, it's, really, it's fascinating and

kind of reassuring, because in, when you look at the dialogue around AI in both academia and in industry- I was talking to a friend of mine, who was saying, is looking at AI ethics and policy, more that intersection there, and she was talking about how she was getting really interested in this idea of responsibility, and how there's always somebody else that people are pushing the responsibility onto to deal with the like, biases or kind of imbalances in a data set, or the harmful results that are model might, might get back. And, and I was saying to her, it's really fascinating to hear that. And I became really interested in this idea of responsibility, because this group is so different to that. It's always about the sense of responsibility. It comes in literally every single time, I think, like individually, when I've chat with people or in groups that sense of responsibility to the collections, and the way they're You're representing people in collections, but also the, the people who are searching your catalogue and coming to try to look at materials. And I kind of, along with that, I was saying to REDACTED that I think there's often this focus on, like, upskilling people in AI, or programming and things like that, or bringing AI into, more-

[multiple participants talking]

01:12:52.030

LH: I think that needs to get be an exchange. I think there's a lot that is missing, and there's like- there's some people that are starting to recognize it. There's this one paper that was published in a big machine learning conference called Lessons from Archives, and it talks about libraries and archives specifically, and documentation practices, and how they're very different. And there really aren't documentation practices in machine learning that are, that's taught in schools. And so it's very, I think. Yeah, basically, there's just a lot of-

[multiple participants talking]

LH: Yeah.

P: There's a message from REDACTED about joining remotely.

LH: Oh!

[setting up remote participation option for another participant (no remote participants joined, though)]

LH: So, the questions are fairly similar. And what you're seeing here is you the types of information that provide overviews [...] So what I was interested in initially as well is, is for the, the tables, and then the charts on the back. What do you understand from this information, what questions do you have about it? And is there, is there a different way that the data could be grouped together, or kind of summarized, that would be more useful than, than it is now. So REDACTED, as you were asking, can you group things by the collection or back to specific description and things like that.

[a participant returns]

LH: Yeah. So for this sheet at just like 5 to 10 min, going to ask people with this kind of overview version of the information. There are tables here, then a couple of charts from the back. What do you understand from it? Do you have questions about what you're seeing?

P: [...]

01:16:14.860

LH: Yeah, like exclusion.

P: [...]

LH: Yeah. Like a missing name.

P: [...]

LH: Yeah, yeah.

INDIVIDUAL WORK TIME

01:19:32.940

P: Could you remind me of what an Omission label is again?

LH: Yeah, so an Omission is like an exclusion of someone's name. So if you know there's a person being talked about and their identity is being described in terms of another person.

P: Okay. One question: is Marjorie, in this sense, a female, a female name.

LH: In my memory that collection was about a woman.

P: Okay.

INDIVIDUAL WORK TIME

01:21:06.920

P9: [...] on how you're going to identify a Stereotype in this context? Like Omission is quite clear, but it's very to me Stereotype is less clear.

LH: Yeah. I, I don't have a great answer, to be honest, which, is, something I need to think about more.

P9: So like how, you know, if you're saying it was most picked up here, what was it picking up? What type of stuff? Like some examples.

LH: Okay yeah so different examples would be often be, women appeared in descriptions as secondary to a man, which is, which is difficult, because often the collection is highlighting the work of the man. But I think there are some examples where- I remember once he was with Florence Jewel Baillie, where it was, there was a description about her kind of a following in her

husband's footsteps rather than them potentially sharing interest, and this kind of assumption that she was supporting his work. But then the tasks that were gone on to be described were about, like, putting together a book about his work, and doing things that seemed quite, quite substantial, were described as something that she did in support of him rather than allowing for her potentially to have had the same interests and maybe that's why they worked well together. And, and so there were sort of things like that where this sort of, the woman or the wife was often described in a very secondary way, and having less, less of an identity of her own, and more in relation to her father or her husband.

01:23:03.100

P3: It's the way it's phrased as well, it doesn't say, "Florence Jewel Baillie was an avid researcher of Dietrich Bonhoeffer, the Second World War, housekeeping, etc." It just says here are some news cuttings on some things that she kind of read about. Which kind of makes a judgment. That's what I thought the Stereotype was in this instance.

P1: If you actually look at that collection, if you actually look at the life of her, and in comparison to her husband's, she has some really, roles of leadership, and [...], and none of that comes through necessarily. And that's where REDACTED was saying about having that neutral voice is possibly-

P8: Or that more work needs to be done. So REDACTED will go exactly the same with REDACTED and there, there's the temptation to overstate, then her role and purpose, and that's not right, either. And so, and, and it's really hard, because, you know, I'm not a researcher. I'm, you know we've just checked, but kind of like just to catalog this stuff. But the research hasn't been done. But you almost have to do a certain amount of research to actually be correct in your cataloging. And the way, the whole tone of, you know. So all the clarity of actually what you can say is really, really complicated. But you know I can't say that Mary was a woman of science or woman scientist in her own right, the evidence just isn't there. But for his secretary, his female secretary went on to write books. So I'm slightly more confident about saying- I mean, so then

you have to research her relationship, his relationship to get a hint as to what Mary's role was as well, and that takes you into like a 3 year, you know, Ph.D.! And you know you've got 6 months to catalog the collection.

[multiple participants talking]

P8: She's [Mary is] the wife.

LH: One other example that we talked about before was Susan Binnie Anderson. So there is a description that said that she essentially like, she started to have a career as a GP, as a General Practitioner, and she had gone to med school, which like, from the sounds of it she was one of the few women in med school at the time so that was quite an accomplishment. But it eventually, the description about her, or this collection or, or collection item ended by, you know, she had a- she tried to keep her career going as a as a GP but with her two kids it just wasn't meant to be.

[laughter, talking]

LH: That really stood out as a Stereotype, and I'm having trouble kind of clearly defining it, but it's, it's like you said you read that, and you're just like, it just wasn't meant to be?

[multiple participants talking]

01:26:10.780

P1: I'm pretty sure, I'm pretty sure - I need to do the reading - but I'm pretty sure she did continue to be a GP because I think she was married to one of the, one of the artists who's quite well known from ECA. And I think I've got film of her in Murray house being the GP who tends the kids there. And, and so they were- the, the complex layers of these things, and I absolutely love you bringing up that thing about research, because the way cataloging is often is – and I'm going off on a tangent here so reign me in – the way cataloging is often talked about is you just go and catalog. And actually to bring the research that

it takes into the very heart of it, and to acknowledge that that is real, powerful research that is analyzed like we we analyze, and we could over analyze, you know how we describe, and, and, and that is really important. And I think we should talk about cat logging in that way, and the process of cataloging in that way the amount of time it, it takes. So that we, we, we, the articulation of that is much more powerful.

P3: Yeah, we have to get something down though, so, so the researchers that you might have the funding can come, can tell us, can find in.

P: So I think a lot of the Omissions will just be cataloging, and lack of a time, you know, and and lack of other sources to help you kind of build that kind of narrative. It just won't be there.

P7: Can anybody remember how big the Roslin Slide Collection is? Because I thought there'd just be pictures of animals in there.

[multiple participants talking]

P1: I am not surprised the Roslin Slide Collection is, you know, ones with the most Stereotypes, and the Omissions as well.

P5: Who cataloged it?

P1: Well I'm not going to name names.

[laughter, multiple participants talking]

P1: [...] item-level description, actually taking the what you see on, on the slide, and putting it in as the catalog description. Now, the layers of this, you know you have to have the layers of context for what the Rosalin Slide collection was: a teaching collection. The images were collected during the height of empire and colonialism. And they um, there is, there are quite a few gendered- language within it that is gendered. There's also how women are represented and that can be in terms of different parts of the world and

communities.

LH: And groups within that where they, they're showing women of a particular tribe, or particular-, and it is, it is animals as well. And so it will talk about animals in a gendered sense as well.

P9: Which we can't really avoid! Especially in terms of biology.

LH: [... (talking about manual annotation)] not focused on labeling animal things, and that it was just focused on, on humans. I don't know how much a model can pick up on which will be interesting. I should, I should actually investigate that collection, it would be a good-

P8: Yes, I suspect there is some complicated problem which is the data and the and the collection and the way the data is being compiled, which is why it's got such high numbers there.

LH: Well, so the other thing about the photos, another kind of sort of subcategory I guess of the Stereotype label that the annotators and I decided to note was if there was a photo, and it was described as a man and a woman but there was no name or kind of evidence that we really knew the gender of the person. And I- this is where I'm kind of curious of like how useful this sort of labeling is for you all. But basically, where it seems like in the photo a person's or a group of people's gender was being assumed, we would flag like women or woman, man and men, because it was always binary. And so it started to feel like, maybe, maybe that was problematic. But-

01:31:03.970

P5: I feel like with the Gendered Pronouns [...] I think there's a lot of debate at the moment. One of the uses I could see, for example, with the Generalization [...] with the Occupation of someone. So, because, a doctor for example assumed to be a "he" and nurse assumed to be a "she." In this case that would be very useful. But I don't know if [...] . So when you picked Gendered Pronouns did you literally just go through the text and highlight all

the pronouns?

LH: Mmhmm.

P5: Okay. So I think we can assume that there will be pronouns in the text because that's part of the language but for me the situation in which that would be useful is [...] an assumption had been made. But I don't know, yeah-

P2: It'd be hard to even figure out because you're trying to understand if the cataloger assumed gender, so you'd have to, yeah, you'd have to the research would be basically, was there any other indication outside of the fact that it said "nurse" written on a slide-

P5: Yeah, yeah so it's really hard, but uh-

[multiple participants talking]

P1: I'm probably saying what you just said but, it's about combining some of these things together. It's not about just the, the Gendered Pronoun that it's, it's actually when they come together, they produce something that's really useful.

P2: It's about like, the assump-, like, assumption, right? I guess that's maybe what you're saying REDACTED as well. Is like, where can we identify an assumption has been made about gender. And there is implication to that, this- or yeah, how that assumption has been made because if it's a well-known historical figure, you'll have lots of external references. So you know, he, he- or you know, but if there's absolutely no evidence as to whether-

[multiple participants talking]

01:33:08.220

P3: [...] rather than a man and a woman, they're just- It could still be useful if someone is saying i'm looking for a photo of my parents on the day they got married. And like, they were married in the year 1817, and, I guess that would

be like great grandparents.

[laughter]

P3: It could still be useful to have what the assumed, gendered assumptions of the cataloger as long as you know that the cataloger assumed that not that the cataloger knew that.

P1: That, that, that's quite interesting to define in the realm of the cataloger.

P3: Yeah!

P1: This agency of the cataloger.

P3: We do that in digital archives, is we put a name to every action that's taken on that digital data. It's just trickier to do for analog stuff.

P: But do you put pronouns?

P3: Probably! Like in some cases it depends-

P: I think practically we just all assume. So I don't know whether that really is going change because we've got to just get the stuff done. And-

P3: That absolutely at the end of the day is true.

P: Yeah! And so, what's, I don't know how, how useful it, it would be, really, because at the end of the day I probably just hit assume all the time. I mean, even if you just Google somebody and they say they're married, that link doesn't go anywhere you don't, you can't actually put your hands on the woman's information at all. It's not just not, it's just not been done. The records haven't been kept. They're just not, they're just not represented at all.

P9: It's like that if you tried to look into your family's history, you can basically follow the men because the women didn't have professions, or weren't, sort of,

recorded as having professions. It's much more difficult to-

P: Yeah, they're just not there.

[multiple participants talking]

P: [...] notebooks and stuff just haven't been kept. So they're [women are] on certificates, but that's about it.

P1: What pops into my head, was, the, a collection which we don't have not here, so it's probably hard to study, but were the papers of a journalist and travel writer who was born male and then transitioned, in the sixties, I think, to become female. I need, how, I, I want to go and look up how the, the institution that cataloged, the institution that had that collection, how they treated and described that particular collection. I mean just, just, it, you know, you know, do they start and then transition? How does that person talk about themselves in their own collection. That's where it becomes that denser, research-based cataloging again. And we we're trying to balance time, and effort and expectation of getting through catalogs and getting through descriptions. Because I can think of the work that we'd have to do. And so so yeah, it's, it's density of information versus where there needs to be more research.

Just, just going back to REDACTED's point of flagging it can be useful. So if we do get resource, and we get to go back to something, then we've got we've got like a marker down there, and it's easily findable again.

01:36:40.530

LH: So one thing I was wondering about too, because I remember from the Critical Archives Reading Group, that some of you went to, and one thing that came up along my was that someone was saying they wished people would see the catalog as a work in progress, like, kind of, visitors. So not necessarily people in Heritage Collection, but researchers or members of the public coming to the, the catalog and searching through it. And so I was curious from your perspective, is there any information on these, this current worksheet or any

information on the first one that I passed out that you would want to share with visitors to the collection in any way? Or variations of it, not necessarily exactly what's on the worksheets?

P1: I, I think it'd be really good to ask some of our User Services people who get inquiries in [...] for example where they get the questions asked about some collections and then sometimes they come and speak to us but and it it could also be that they feel that there isn't the information there for them to actually provide, in a way. So they, they would appreciate having something as part of the catalog.

P: Not, I, it'd be useful to speak to users generally, anyway. Because I've got a feeling that they come with an, a knowledge that we don't have. So, so, we're, we're kind of cataloging and try to make it as clear and as transparent as possible as to what we've seen. And the users are using that, and coming with their own opinions as to the that particular person or that particular place that they are, you know, probably more knowledgeable on specifically. So are they bothered about, I don't know whether they be happy with us, spending a lot of time on, you know. I feel as though they've been much more getting this stuff out there, and then it's about them coming or correcting, or enhancing or building on what we've done.

So I'm, I'm not sure, it would be really interesting to see what users thought of I I mean, I totally understand that clarity of language has to be in there, and you know it's been really good for that. But I would say that they were more interested in through-put, possibly, but I may be wrong there. I mean, I know that we can catalog something better, and already people are finding it, because it's better cataloged. But that's, that's a step back before, before where you're going, but certainly, certainly helpful.

P9: I mean in this case it's if you're like writing an algorithm or something that will pick up these instances, then it would enable us to correct things more quickly if we wanted, for example, to revisit, ah, just say, for example, the ArchivesSpace platform that REDACTED and I've been working on together that you could sort of run your algorithm through the data and pick up these

instances that we could just go and check rather than us having to quickly to have to go through every entry thinking about it, you could say, well I ran my algorithm on your on your collection and I found sort of 10 cases where you've got an Omission, and it means we could just go to them. And that way, maybe the type of thing you're developing could be useful, because maybe sometimes we say, oh, no well there, that's actually not a real case, because we couldn't do anything else or whatever. But it would maybe enable this kind of revisiting of and correcting to be done more quickly and with less, kind of, labor.

P: Definitely.

01:40:56.430

P: Because we don't want to keep going back over stuff that's kind of been done. But then I mean I remember looking at catalog that was supposed to be done, and it was done in like 2003, so it's not even that long ago. And it was done of like a, a plantation. And it was written in in such a biased way, the, the catalogue was so biased. And it was really, it really needed to be redone. And it was pretty useless at, in its current form. But it did represent that one side of that story, but it didn't represent the other side at all. And again, of course, that research all probably needs to be done, but that catalog should have reflected that. So it was really kind of out of date quite quickly.

P2: It seems like context is so important for all of this as well. Like, you know, like I guess there's like, you know, you know, things like this, and yeah, like, names, and when you go and revisit, or you look at the context, it can completely change these things. So it's how, that is that is added to the, more kind of, algorithm or processes. It just feels like, I don't know, is there an outline of a certain methodology going forward and an acknowledgment that, yeah, I don't know. I'm, I'm just I'm just curious about how how the context is woven into the kind of analysis. You know. Because it it's just more like how will this be used? Or who's going to look at the data? And, like, yeah. What can be drawn from it? What conclusions can be drawn from the, the data? And the research seems [...] kind of a big, big question, you know, like, for example, something like so there's a load of Omissions with very few Stereotypes. But

on the other way round, you're like, okay, well, actually, why was that? Or is there a certain reason? And it doesn't necessarily mean that the cataloger was thinking a certain way, or doing something-

P9: Yeah, I'm thinking of quotation.

P2: Exactly. Exactly. I'm thinking about assessments and saying like which ones are being seen as this is high priority, because it's like full of Stereotypes or it's full of Omissions. But yeah, anyway, yeah. I'm just curious about how it's to be used. But I think you're right in a very practical way to do a search of something that you're actively working on. But that's because you guys are working on it. You've written all the content already. It's kind of a process of, of checking, double checking. So it's useful for assessment here

[multiple participants talking]

01:43:55.380

P2: [...] more quickly than like going back through the historical material, which is so diverse.

LH: For a lot of the Processing Information fields, when I looked at the Catalog initially, only about a third of the descriptions, or the collections, actually have a date. And I think REDACTED, REDACTED I think you were saying that, that that's more regular to record now, and there were different standards, I think you said it was a library standard initially that had been used for the archives, [...] switched to an archival standard, is that correct?

P5: Mmm, oh yes yeah they used to [...] catalog archival material as if they were books. [...] and now we're applying the ISAD(G) archiving standard.

LH: Yeah so I think the challenge with context is sometimes it's just not known. And you guys probably from looking at the collection and the language, you would probably have a better sense of when that was cataloged but as someone just coming to the data, if there's no date, I'm like, well, I have no idea!

P1: In terms of the dashboard, actually having the date of the collections, would be really useful. That starts with the context. The other thing just looking at the collections that come up, I'm absolutely fascinated that Godfrey Thomson has come up in the Stereotype because, knowing the dates of when he worked, knowing that, he did, his work was on, you know, mental assessments, and describing people in his research? They might be described, and either gendered or stereotyped because of the time period he worked in but also the environment. And you know, I think some of this work, early on looked at eugenics, and so that kind of classification of people, and how clever they were, and how- what skills. So I'm like, Wow! That came up. And I don't know we'd have to investigate further. And this is where, you know, another aspect of it to be able to, you know, like you said REDACTED, tag up and go that's why that's there, is it because of what I've just said, or no, that it's not playing out like that. To see if it, it works.

Again with Koestler, ooh! Coming up where he comes up. I'm, I'm thinking of who actually created that catalog of Koestler's. And we took that catalog and converted it and put it online. The time period he lived in, and also his journalism, what that covered, and what we know now about how, how he was misogynist, and probably was with women and, and gender, and so, that, there, there may be bias in what I'm saying with that as well but. Yeah, I, I'm really interested in what came up.

LH: And this is yeah, like I said, this is the first in 20% of the catalog that was extracted, so it's not, it's not every collection, but it's interesting to hear that you're kind of not particularly surprised with what's come up.

P1: Especially for some of them. But also I'm thinking of when we have REDACTED in doing the social work records of the Department of Social Work at the University, and she did the mental health surveys, and that- the catalog is really rich. If that was put through this model, the gendered language, the stereotypes of how those people were described in the 1940s and 50s would-

[multiple participants talking]

01:48:13.010

P3: Like the women were described as "promiscuous" but that term wasn't used to describe men. Would your model have flagged that?

LH: Well that's exactly- I don't know because I don't at least I can't remember in kind of looking at the terms, because I looked at the most frequent terms that were coming up with some of these labels, and I don't know that "promiscuous" came up so far, so I wonder it might not be it might not be picked up on, and that's where again, I think it's like these kind of like with the evolution of language it's like. I think these models are things that would just always be partial. You would constantly need to be finding ways to kind of add to them, and it would be great if there was a way to start adding like keywords in or something like that to look for in flag.

P1: So that that comes back, then, to whose language is it? Whose language is it? Plus, who gets to decide, you know, how a term is, I suppose it all comes back to context, doesn't it? But if you use, you know something descriptions like "promiscuous" or something, then is it used in a negative or a positive way? Is it used in a particular construct? And you know it's it, it's what I was saying earlier about thesauri and that problem I had with thesauri is who's on the boards that get to deem that this is the right language?

LH: Yeah. So there was another category that I'd including the taxonomy but it didn't end up being super relevant, and it was Empowering. The idea was, when there's this reclamation of word that had been derogatory. Like the word "queer", now there a lot of people that identify as queer but initially, that was, it was very derogatory, and and it didn't I, I don't know if it's, if it's cataloging in particular, but that language just isn't there as much, or if it's more just the collections that we ended up looking at for the annotation process. But I, I think that kind of, like you were saying, how is the word being used? It's like it's about more than just flagging words sometimes. It's about trying to pick up on what the judgment attached to that word is, if you can.

P1: I wonder, they, maybe I'm wrong, but I wonder whether actually, if there

were collections, we would suggest, putting through the model to test it, to see if it picks up on things that we know are in those collections, whether that would be something useful- I can see nodding heads. I was thinking REDACTED, about, the descriptions of the [...]

P2: Yeah, definitely.

[multiple participants talking]

P8: [...] that's a completely different kind of approach. So we've got a bit of description going on, and then a literal transcription of his historical indexes. And we're saying REDACTED's own indexes, and making it clear that either he or REDACTED or his other secretary at certain points have created these indexes. So there is historical language right there in the catalog.

P1: I just think that, it's, different examples [...]

[multiple participants talking]

P1: [...] language that is more, near history, as it were.

01:52:09.720

P: So actually, yeah a good catalog for me, is, something like that, yeah, something really full of historical language. Because that's kind of what we're seeing. And then it's about content support or, like, trigger warnings, or whatever. At a higher level.

P3: I know it's not directly related to your research [...] with digital collections the use of AI to do the cataloging, to automate the description [...] but are you perpetuating biases from the resource itself?

LH: And from yeah, from datasets that are so large that they haven't been, they're not publicly accessible. So you can't interrogate them, and when bits of them are interrogated they find things like pornography in them, and it's yeah,

it's, it's concerning.

P3: For digital archives, like at the scale some of them are collected, what, there is no, like, the human race will die out before we be able to catalog every website.

LH: Thank you everyone. Yeah, I just quickly, I can send a message afterwards too, but I wanted to order catering just as a thank you, but I was a bit slow and getting it organized. And so I was curious if there is a good day of the week that most of you are in, so that I could order just like coffees and pastries, and we'll provide sort of breakfast-

[multiple participants talking]

P3: Can we get together to talk about this some more?

[multiple participants talking]

END: 01:54:00.460

Appendix I

ACH 2020 Presentation Abstract

I was lead author on the following abstract, with my supervisors, Benjamin Bach, Melissa Terras, and Beatrice Alex, as co-authors. I presented the work described in the abstract at the 2020 Association for Computers and Humanities Conference, which was held virtually.

Documenting Gender Identities: Challenges and Approaches to Records of Gender in Archival Metadata Descriptions

Gender bias has been built into algorithms through data collection practices that privilege a particular perspective, misrepresenting or excluding perspectives of many gender groups. Assumptions these algorithms make about data's representation of universal truths shape the way people find and interpret information for learning and research, rendering so-called unauthorized, minoritized, or perverse perspectives invisible. As digitization of heritage collections progresses, and the online discoverability of heritage data grows, there is a risk that historical perspectives within cultural heritage collections will amplify gender stereotyping and discrimination already well-documented as sources of oppression. Scholars and practitioners have published approaches for the removal of gender bias, attempting to create objective technologies. However, little attention has been given to understanding the origins of gender bias, and how it manifests in the descriptive language of heritage collections' metadata. As records of culture and history, heritage institutions' metadata

descriptions provide repositories of text well-suited to serve as data sources for diachronic, intersectional analyses of gender-biased language. Only once we understand gender bias – where it comes from, how it is communicated, how it varies from one culture to another – can we begin to effectively mitigate its harmful consequences and design technological systems that can navigate it. Through a case study with English-language archives at the University of Edinburgh's Centre for Research Collections, this work outlines the challenges of respecting gender identities in a heritage context and lays out a path to addressing these challenges through natural language processing and participatory research methods.

Appendix J

SRW 2022 Presentation

I authored the following non-archival paper, which I presented at the Student Research Workshop at the 2022 Conference of the North American Chapter of the Association for Computational Linguistics in Seattle, US.

Towards Gender Biased Language Classification: A Case Study with British English Archival Metadata Descriptions

J.1 Introduction and Background

The need to mitigate bias in data has become urgent as evidence of harms from such data grows (Noble, 2018; Perez, 2019). Due to the complexities of bias often overlooked in natural language processing (NLP) bias research (Devinney et al., 2022; Stańczak and Augenstein, 2021), Blodgett et al. (2020) and Crawford (2017) call for greater interdisciplinary collaboration and stakeholder involvement in NLP and machine learning (ML) research. The Gallery, Library, Archives, and Museum (GLAM) sector has made similar calls for interdisciplinary engagement, looking to applications of data science and ML to better understand and mitigate bias in GLAM collections (Geraci, 2019; Padilla, 2017, 2019). Supporting the NLP and GLAM communities' shared aim of of mitigating the minoritization[1] of certain social groups that biased

[1] D'Ignazio and Klein, 2020 propose the term "minoritization" to describe a group of people's *experience* of oppression, in place of "minority" which defines people as oppressed.

language causes, this project aims to develop a classification model that categorizes biased language in GLAM documentation. The project uses the term *biased language* to refer to "written or spoken language that creates or reinforces inequitable power relations among people, harming certain people through simplified, dehumanizing, or judgmental words or phrases that restrict their identity; and privileging other people through words or phrases that favor their identity" (Havens et al., 2020). The project uses the term GLAM *documentation* to refer to the descriptions of cultural heritage collection items written in catalogs of galleries, libraries, archives, and museums.

Studying GLAM documentation provides an opportunity to study the evolution of biased language, because descriptions in contemporary GLAM catalogs contain both historical and contemporary language. To provide a record of the past, GLAM continually acquire and describe heritage items, structuring descriptions of the items according to metadata standards (such as Research Description and Access (RDA Steering Committee, 2022)) and subject authorities (such as Library of Congress Subject Headings (Library of Congress, 2021)). The heritage items included in GLAM, along with the language used to describe them in catalogs, have a continual influence on society (Benjamin, 2019; Cook, 2011; Smith, 2006). The processes of selecting which items to bring into GLAM, and organizing those items according to standards and authorities, privilege particular perspectives (Adler, 2017; Bowker and Star, 1999; de Jong and Koevoets, 2013; Furner, 2007; Olson, 2001; Tanselle, 2002). These processes shape society's understanding of the present and can either reinforce or challenge existing power relationships among people (Benjamin, 2019; Cook, 2011; de Jong and Koevoets, 2013; Noble, 2018; Smith, 2006; Yale, 2015).

Through case studies of free-text descriptions in many GLAM catalogs, variations in biased language over time and across locations could be better understood. Should patterns in the evolution of biased language emerge, language technology could one day be trained to identify newly-emerging types of bias that it has not yet seen. This project takes the first step in that direction, with a case study of biased language classification for GLAM documentation.

To create biased language classifiers, the project defined three objectives:

O1. Define types of bias for GLAM.

O2. Measure the prevalence of biased language in GLAM documentation.
O3. Build and evaluate classifiers to detect bias.
O1 has been achieved and O2 is in progress (§J.4). As the project proceeds, several approaches are under consideration for building and evaluating classifiers (§J.5). Recently passing the halfway point of a three-and-half-year Ph.D., the project would benefit from feedback at the Student Research Workshop to discuss approaches to O3.

J.2 Related Work

Awareness of limitations in approaches to bias mitigation in Natural Language Processing (NLP) and the wider Machine Learning (ML) community is growing. Publications about NLP bias research now include not only efforts to debias datasets and algorithms (Webster et al., 2018; Zhao, Wang, Yatskar, Ordonez, et al., 2018), but also recommendations to address the complexity of bias that debiasing efforts often miss (Bender and Friedman, 2018; Blodgett, Lopez, et al., 2021; Goldfarb-Tarrant et al., 2021; Gonen and Goldberg, 2019; Havens et al., 2020; Jo and Gebru, 2020; Mitchell et al., 2019). Recognizing the harmful impacts of deep learning models trained on datasets too large to be adequately interrogated (Bender et al., 2021; Birhane and Prabhu, 2021; Noble, 2018), this project will train supervised NLP models on a dataset small enough to be interrogated (399,957 words, 24,474 sentences). Moreover, collaborators include archivists who manage the collections described in the project's data and have expert knowledge to inform annotation and analysis processes.

Recognizing the subjective nature of certain NLP tasks, such as detecting hate speech and bias, Davani et al., 2022; Sang and Stanton, 2022; and Basile et al., 2021 have questioned annotation approaches that create a single gold standard or ground truth dataset. The "perspectivist" approach to NLP this inspired, which incorporates multiple annotators' perspectives in published datasets (Basile, 2022), aligns with the data feminist approach that (D'Ignazio and Klein, 2020) put forth. Data feminism views data as situated and partial, drawing on intersectional feminism's view of knowledge as particular to a specific time, place, and people (Crenshaw, 1991; Haraway, 1988; Harding, 1995). Feminist theories argue that the standpoint (perspective) of a person

impacts knowledge and understanding, and that a universal standpoint cannot exist (Haraway, 1988; Harding, 1995). Indigenous epistemologies, such as the Lakota concept of *waȟkàŋ*, further the notion of the impossibility of a universal truth (J. E. Lewis et al., 2018). Translated as "that which cannot be understood," waȟkàŋ communicates that knowledge may come from a place beyond what we are capable of imagining (ibid.). To create a dataset of GLAM documentation annotated for gender biased language, this project creates an annotation taxonomy that allows for gender information to be labeled as uncertain or excluded, and incorporates multiple annotators' perspectives in the model training data.

To practically apply theories and approaches from perspectivist NLP, data feminism, and indigenous epistemologies, the project applies the case study method common to social sciences and design research. Case studies use a combination of data and information gathering approaches to study particular phenomena in context, focusing on "consideration of the whole, covering interrelationships," which provides a "depth [that] compensates for any shortcomings in breadth and the ability to generalize" (Martin and Hanington, 2012a, p. 28). Furthermore, case studies report and reflect upon outliers discovered in the research process (ibid.), useful for this project's effort to create space for the perspectives of minoritized people. This project provides a case study for NLP bias research with the long-term aim of building a collection of case studies, which would enable the NLP community to determine the aspects of bias mitigation approaches that can and cannot be generalized across contexts.

J.3 Methodology

The interdisciplinary nature of the Ph.D. project warrants a combination of methods and frameworks from several disciplines. Adopting the bias-aware methodology of Havens et al., 2020, case study and participatory action research methods complement NLP methods for creating the project's annotation taxonomy, annotated datasets, and classification models. Critical discourse analysis, feminist theories, queer theory, and indigenous epistemologies provide frameworks through which to analyze the project's metadata descriptions and annotated datasets. To begin, the author defined

gender bias using the definition of biased language of Havens et al., 2020 (quoted in §J.1) narrowed to *gender* bias. This definition informs the annotation taxonomy, which in turn will influence classifiers created with the annotated data.

Participatory action research methods are used to incorporate stakeholder perspectives, necessary for situating a study of gender bias in a particular time, place, and people. Situated in the United Kingdom, the project works with archival documentation written in British English from Heritage Collections at the University of Edinburgh (HC).[2] Due to the numerous characteristics on which bias may be based, such as racialized ethnicities, economic class, gender, and sexuality, a focus on *gender* bias was chosen. This focus supports the HC's existing effort to mitigate gender bias in its collections. A person's gender is considered to be self-described, changeable, and capable of falling anywhere along a spectrum of femininity to masculinity (Keyes, 2018; Scheuerman, Spiel, et al., 2020). Archivists provided feedback during the development of the project's annotation taxonomy (§E.8), and will provide feedback in future work analyzing the data annotated with the taxonomy.

The annotation taxonomy and instructions for applying the taxonomy focus on documenting information explicit in the text to avoid misgendering (Scheuerman, Spiel, et al., 2020) annotations do not infer a person's gender from the person's name, occupation, or other descriptive information, nor do the annotations assign a particular gender to a person. Rather, the annotation process records whether the terms used to describe a person are "feminine," "non-binary," "masculine," or, if only gender-neutral terms are used, "unknown." Annotators were instructed to read the metadata descriptions from their contemporary perspective. That being said, as the historian Shopland writes, "when writing of historic LGBTQIA+ people, we use a definition which simply did not exist in their lifetimes" (2020, p. 1). Consequently, the project acknowledges that the perspectives documented in the annotation process are situated not only geographically and culturally, but also temporally, in the 21st century.

Following Smith's (2006) approach, the project views heritage as a process. Smith writes, "what makes certain activities 'heritage' are those activities that actively engage with thinking about and acting out not only 'where we have come from' in terms of the past, but also 'where we are

[2] archives.collections.ed.ac.uk

going' in terms of the present and future" (ibid., 84). The annotation process of this project visits, interprets, and negotiates with heritage (ibid.) in the form of archival documentation, directing NLP technology towards trans-inclusive conceptualizations of gender, and making gender biases in archival documentation transparent. Smith's approach to heritage draws on Fairclough's (2003) approach to critical discourse analysis (CDA).

Discourse consists of language and its production, interpretation, and social context (Fairclough, 2003; Marston, 2000). CDA thus provides a valuable lens through which to study the heritage material of this project: descriptions from an archival catalog. Considering language in its context of use, CDA offers an approach to studying how language legitimizes, maintains, and challenges power (Fairclough, 2003; Marston, 2000; Smith, 2006). The project uses CDA to follow the data feminism principles of examining and challenging power (D'Ignazio and Klein, 2020). Through annotations of gender biased language, the project examines and challenges the dominant perspective of men in the archival metadata descriptions, making this perspective explicit and identifying opportunities for perspectives of additional genders to be incorporated into the descriptions.

	Title	Biog. / Hist.	Scope & Contents	Processing Info.	Total
Count	4,834	576	6198	280	11,888
Words	51,904	75,032	269,892	3,129	399,957
Sentences	5,932	3,829	14,412	301	24,474

Table J.1: Total counts, words and sentences in the aggregated dataset. Counts displayed per in the descriptive metadata field and across all fields, namely "Title," "Biographical / Historical" (Biog. / Hist.), "Scope & Contents," and "Processing Information" (Processing Info.). Calculations were made using Punkt tokenizers in the Natural Language Toolkit Python library (Loper and Bird, 2002).

J.4 Work Achieved

The project has accomplished O1, defining and categorizing types of gender biased language for archives, through the creation of an annotation taxonomy. The taxonomy defines types of gender bias to label in a corpus of archival documentation. Currently the project is progressing on O2 and O3, which

are interrelated: the manual annotation process allowed for calculations of the prevalence of gender biased language on a subset of archival documentation, and the classifiers, once built, will enable more complete calculations of gender biased language on the remainder of the descriptions in the archive's catalog. This section summarizes the work achieved on O1, O2, and O3; the next section (§J.5) outlines potential directions for completing O3. Havens et al. (2022) contains a detailed discussion of the annotation taxonomy and its application to create an annotated dataset.

The project's annotation taxonomy builds on literature from ML (Hitti et al., 2019), human-computer interaction Keyes, 2018; Scheuerman, Spiel, et al., 2020, gender studies (Butler, 1990), archival science (Tanselle, 2002), and linguistics (Bucholtz, 1999, 2003; Fairclough, 2003). Group interviews and workshops (participatory action research methods (Moore, 2018; Reid and Frisby, 2008; Swantz, 2008)) further informed the annotation taxonomy. The final annotation taxonomy consists of three categories. Each category contains subcategories with the labels that the annotators applied to archival metadata descriptions. §E.8 contains the complete taxonomy with definitions and examples.

The first two categories of labels, Person Name and Linguistic, annotate vocabulary choices and lexical relations that are explicit in the descriptions, providing a record of the "internal' relations of texts" (Fairclough, 2003, pp. 36–7). The third category of labels, Contextual, annotates according to the descriptions' relationship with a social context (for example, events, behaviors, and power structures), providing a record of the "'external' relations of texts" (Fairclough, 2003, p. 36). Approaching the archival documentation as discourse, the annotations make the connections between the internal and external relations of language transparent.

Annotating heritage in the form of archival metadata descriptions adds to the process that is heritage, evolving the meaning of the descriptions (Smith, 2006). Applying annotations to archival metadata descriptions from a 21st century perspective recontextualizes the descriptions, adding to the genre chain, or network, of archival documentation that begins with the archival items and continued with catalogers' descriptions of the items (Fairclough, 2003). The taxonomy permits annotators to record uncertainty and absence of information (J. E. Lewis et al., 2018; Shopland, 2020), deviating from past

NLP documentation approaches (i.e., Dinan, Fan, Wu, et al., 2020; Garnerin et al., 2020).[3] Participatory action research found that archivists view archival documentation as incomplete. The primary purpose of describing archival items is to enable their discoverability, but this must be balanced with the need to describe a backlog of new archival items perpetually being acquired.

The corpus of archival documentation for annotation were created by harvesting metadata descriptions from an online catalog, reformatting the descriptions for annotation, and manually labeling the descriptions according to the annotation taxonomy. The archival documentation comes from four metadata descriptions in the online archival catalog of the HC: "Title," "Biographical / Historical," "Scope and Contents," and "Processing Information." Though not all descriptions have a date recording when they were written, the earliest recorded date of a description's writing is 1896 and the latest, 2020. The HC Archives include a variety of material, such as photographs, letters, manuscripts (letters, lecture notes, and other handwritten documents), and instruments; and cover a range of topics, including town planning, research and teaching, and Scottish Presbyterianism. The language of the HC Archives' materials are mostly English (1,018 out of 1,315 collections, about 77%), though over 80 languages total are present across the collections. The descriptions that were annotated account for about 20% of the text in the entire online catalog of the HC's Archives. Table 1 provides summary statistics about the data. For further detail on the size, contents, and organization of the annotation corpus, please refer to the paper by Havens et al. (2022) and the data statement (Bender and Friedman, 2018) in Appendix E.

The project received grants to hire four annotators, who were Ph.D. students selected for their experience in gender studies or archives. The total cost of the annotation work amounted to circa 400 hours of work and £5,333.76. The four hired annotators each worked 72 hours over eight weeks, receiving £18.52 per hour. The author spent 86 hours annotating over 16 weeks for her Ph.D. project. Though all annotators identify as women, due to the historical dominance of men's perspective in the English language and the pejoration of terms describing women (Lakoff, 1989; Schulz, 2000; Spencer, 2000),[4] the

[3] Domains beyond GLAM also face the challenge of uncertain and absent information (Andrus et al., 2021).

[4] In the 16th century, grammarians instructed that *man* precede *woman* in writing; in the 18th century, *man* and *he* began to be used in place of *human* and *their* (Spencer, 2000).

project's annotated dataset does challenge dominant perspectives in archival discourse to advance gender equity (D'Ignazio and Klein, 2020; Fairclough, 2003).

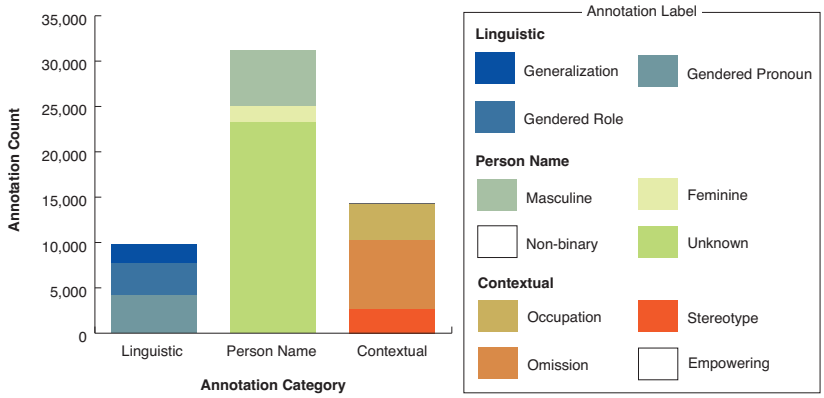


Figure J.1: **Annotations Per Label.** A stacked bar chart of counts of annotations per label across all annotators in the aggregated dataset of 55,260 total annotations, organized into the three categories of labels: ***Linguistic***, ***Person Name***, and ***Contextual***. *Non-binary* (a ***Person Name*** label) and *Empowering* (a ***Contextual*** label) both have a count of zero.

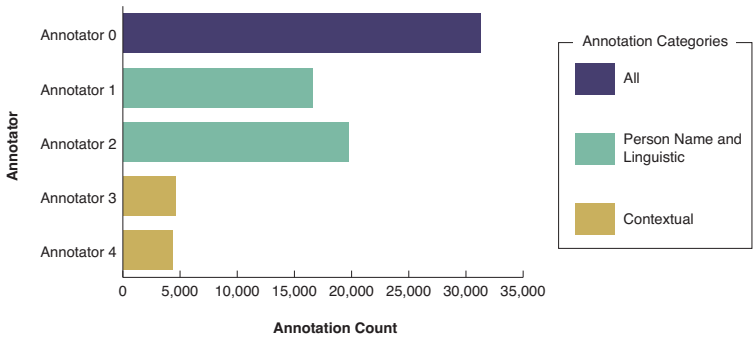


Figure J.2: **Annotations Per Annotator.** A bar chart of the total annotations from each annotator included in the aggregated dataset, with colors indicating the category of labels each annotator used. For annotations that matched or overlapped, only one was added to the aggregated dataset, so the total number of annotations in the aggregated dataset (55,260) is 21,283 less than the sum of the annotators' annotations in this chart (76,543).

Inter-annotator agreement (IAA) calculations reflect the subjectivity of gender bias (Appendix F). Annotating *gendered* language proved to be more straightforward than annotating gender *biased* language. We report IAA

with F_1 as our metric due to the limitations of coefficients' assumptions and interpretability as Artstein and Poesio (2008) discuss. F_1 scores for the gendered language labels "Gendered Role" and "Gendered Pronoun" fall between 0.71 and 0.99. F_1 scores for annotating gender biased language are relatively low, with the greatest agreement on the "Generalization" label at only 0.56, on the "Omission" label at 0.48, and on the "Stereotype" label at 0.57. Manual analysis of disagreements among annotators demonstrated the value of a perspectivist approach to disagreements (Basile et al., 2021; Davani et al., 2022; Sang and Stanton, 2022), as multiple annotators' labels were often deemed correct for the same text span.

The five annotators' datasets were merged into one aggregated dataset, which will be divided into training, development, and test sets for creating classifiers. Aggregation began with a one-hour manual review of each annotator's labels to identify patterns and common mistakes, which informed the subsequent aggregation steps. Disagreeing labels for the same text span were then manually reviewed, with either a combination or an individual label being chosen for each text span to include in the aggregated dataset.

Next, for annotations with overlapping text spans and the same label (considered to be in agreement), the annotation with the largest text span was added to the aggregated dataset. All remaining annotations were then added to the aggregated dataset, with the exception of one annotator's Person Name labels, as these were applied with great inconsistency relative to other annotators. §E.1 details the review and aggregation of the annotated datasets. Figures 2 illustrates the prevalence of each of the taxonomy's labels in the aggregated dataset and Figure 3 illustrates how many annotations from each annotator are in the aggregated dataset. The annotated datasets are a starting point to identify gender bias in GLAM documentation in the UK; they are not intended to comprehensively cover of all gender biases that may come through in GLAM documentation. They will serve as training, development, and test data for developing classifiers, and will be published alongside the classifiers in future work.

J.5 Discussion and Conclusion

Now passing the halfway point of a Ph.D. degree, with a year and six months remaining, the project would benefit from feedback on possible approaches to the project's last objectives. Several approaches are under consideration for building classification models (O3).

Four algorithms are under consideration for building a gender biased document classifier: (1) logistic regression (LR), as a classification baseline (Jurafsky and Martin, 2023); (2) decision tree or (3) random forest (a combination of randomized decision trees), as the decision trees are the most transparent algorithm for classification (S. Bird et al., 2019); and (4) support vector machines, as Adhikari et al. (Adhikari et al., 2019) found this outperformed LR and neural models on document classification for select datasets. The document classifiers could be developed as single task or multitask; the project would like to investigate correlations between labels. As the perspectivist approach to disagreements in NLP encourages (Basile, 2022), classifiers could be trained on individual annotators' datasets in addition to the aggregated dataset. Document classifiers could also be pre-trained on a deep learning model such as DocBERT (Adhikari et al., 2020) to see if pre-training improves their performance.

The focus on document classification comes from the intended use case of the classification models: to support archivists in identifying descriptions with gender biases in their catalogs. Such identification would support the efficient prioritization of reparative description practices that add context to or reword harmful descriptions. That being said, annotators applied labels to text spans, not documents, so sentence classification could also be pursued. Though all approaches have the potential to contribute to NLP and GLAM's efforts to mitigate bias, the 18 months remaining in the Ph.D. provides only enough time for select approaches to be pursued. The project would appreciate feedback at the Student Research Workshop on approaches under consideration to build and evaluate classifiers that detect gender biased documents.

Appendix K

DH 2023 Publication

I was lead author on the following paper, which I presented at the Digital Humanities Conference in Graz, Austria (Havens et al., 2023). My co-authors were Rachel Hosker, University Archivist and Research Collections Manager, and my supervisors, Benjamin Bach, Melissa Terras, and Beatrice Alex.

Collaboration Across the Archival and Computational Sciences to Address Legacies of Gender Bias in Descriptive Metadata

K.1 Introduction

This presentation reports on a case study investigating how Natural Language Processing, a field that applies computational methods such as Machine Learning to human-written texts, can support the measurement and evaluation of gender biased language in archival catalogs. Working with English descriptions from the catalog metadata of the University of Edinburgh's Archives, we created an annotated dataset and classification models that identify gender biases in the descriptions. Conducted with archival data, the case study holds relevance across Galleries, Libraries, Archives, and Museums (GLAM), particularly for institutions with catalog descriptions in English. In addition to bringing Natural Language Processing (NLP) methods to Archives, we identified opportunities to bring Archival Science methods, such as Cultural Humility (Tai, 2021) and Feminist Standpoint Appraisal (Caswell and Cifor,

2019), to NLP. Through this two-way disciplinary exchange, we demonstrate how Humanistic approaches to bias and uncertainty can upend legacies of gender-based oppression that most computational approaches to date uphold when working with data at scale.

K.2 Literature Review

Since the end of the 20th century, GLAM have seen growing resistance to claims of neutrality that characterized previous centuries' collection and documentation practices (Duff and Harris, 2002). Consequently, catalogers, librarians, archivists, and curators have begun to revisit descriptions of heritage items in their institutions' catalogs, looking for instances of omissions and misrepresentations to address through revisions or additions. Revisiting descriptions is a daunting task, however. GLAM catalogs are large and ever-growing: institutions always have a backlog of new items to document so visitors can discover them with catalog search queries. Computational methods, particularly Machine Learning (ML) models, offer ways to lighten the burden of manual labor required to revise and add to catalog descriptions (Cordell, 2020; Greenburg et al., 2005; Harper, 2016; Padilla, 2019).

However, ML disciplines' approach to dataset curation largely reflects pre-20th century GLAM approaches. ML researchers and practitioners create datasets primarily based on which data are readily available in large quantities (Raji et al., 2021; Rogers, 2021). Concepts of bias are overly simplified and uncertainty is largely hidden, leading to biased ML models with harmful consequences, particularly for groups of people who already have a history of experiencing marginalization (Blodgett et al., 2020; Stańczak and Augenstein, 2021; L. Sweeney, 2013). Recently, more critical approaches to dataset and model creation encourage interdisciplinary collaboration and greater transparency in documentation practices to address the harmful biases of ML models (Bender et al., 2021; Crawford, 2017; Havens et al., 2020; Mitchell et al., 2019). The longer history of classification in the GLAM sector has much to offer the younger ML disciplines.

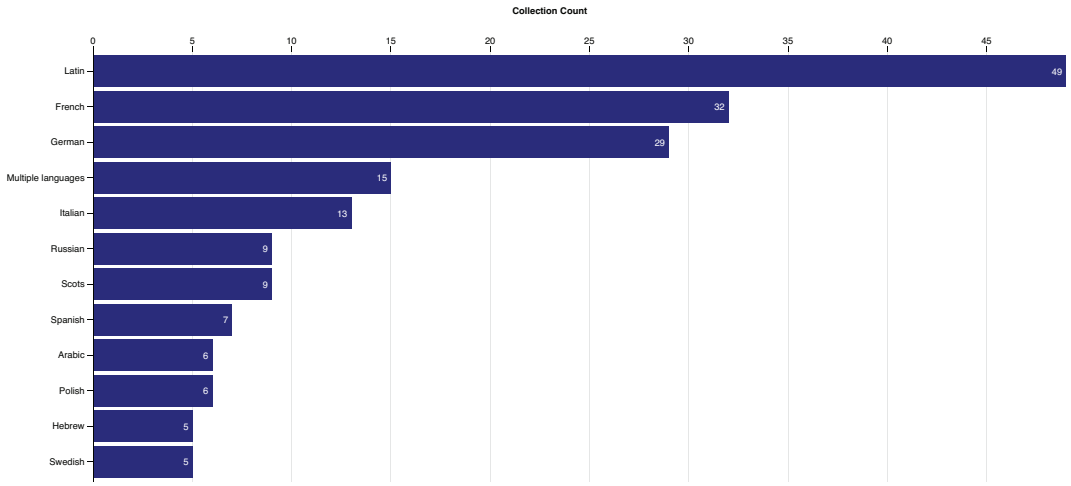


Figure K.1: **Languages of material documented in the Archives catalog.** Most of the HC's Archives are material written in English (e.g., news articles, manuscripts such as letters, lecture notes, degree awards), however other languages also appear in the Archives (as well as non-textual material such as photographs, sketches, and architectural plans).

K.3 Methods

This presentation will report the results of our case study creating classification models that measure gender biases in metadata descriptions, specifically those of the Archives' catalog of Heritage Collections (HC) at the University of Edinburgh (Heritage Collections, n.d.). The Archives mainly contains material written in English, however other languages (see Figure K.1) and non-textual material are also documented in its catalog. The Archives' need to measure and evaluate gender biased language across its entire catalog motivated us to take an atypical approach to bias research in NLP. Rather than trying to remove or fix gender biased language, we aim to identify it, arguing that biases are inherent to all language and should be made more transparent to the reader. This approach aligns with the subjective nature of cataloging that Bowker and Star (1999), Duff and Harris (2002), Cook (2011), and Adler (2016) describe; and implements the interdisciplinary collaboration that Jo and Gebru (2020), McGillivray et al. (2020), and Devinney et al. (2022) call for in computational research.

The case study consists of four steps: define types of gender bias; create a dataset annotated for gender biases (see Figures K.2 and K.3); create NLP models that identify gender biases in language; and analyze the results to study how gender biases manifest in descriptive metadata. An interdisciplinary

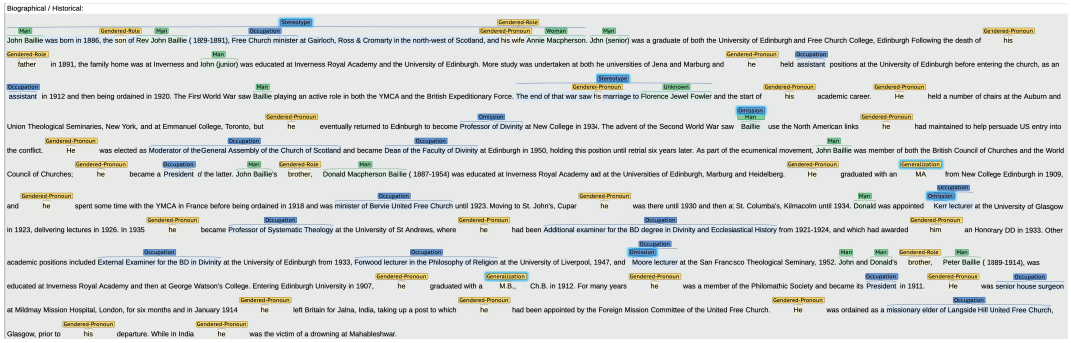


Figure K.2: **Annotations in brat.** Example of metadata descriptions from HC's Archives catalog annotated with the brat rapid annotation tool (Stenetorp et al., 2011). Annotators labeled text spans of one or more words with eleven labels, color coded by label category: green is ***Person Name***, yellow is ***Linguistic***, and blue is ***Contextual***.

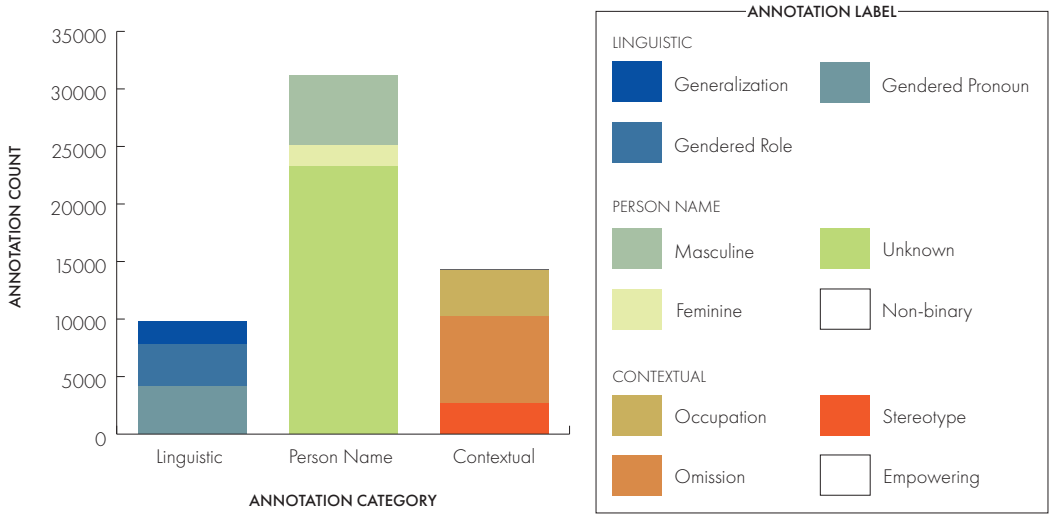


Figure K.3: **The annotated dataset.** Five annotators annotated a corpus of 399,957 words across 11,888 descriptions in 245 fonds (collections), resulting in a total of 55,260 annotations. The annotated dataset represents 10% of the entire Archives catalog. *Non-binary* and *Empowering* both have a count of zero. (Figure reproduced with author permission from Havens et al., 2022.)

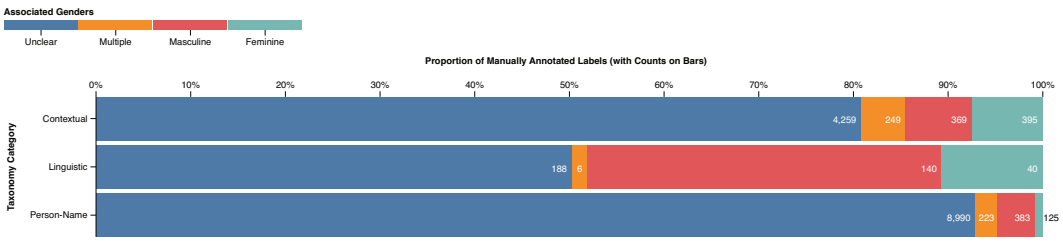


Figure K.4: **Grammatical gender associations of the Stereotype label.** The proportions of each annotator's labels for the Contextual category that are associated with masculine (blue), feminine (orange), or multiple genders (red), or an unclear association (turquoise). Note: The ***Person Name*** annotation category includes the *Non-binary* label, however annotators did not find text in the selection of archival metadata descriptions they read that used explicitly non-binary referents, so no name in our data has a *Non-binary* annotation.

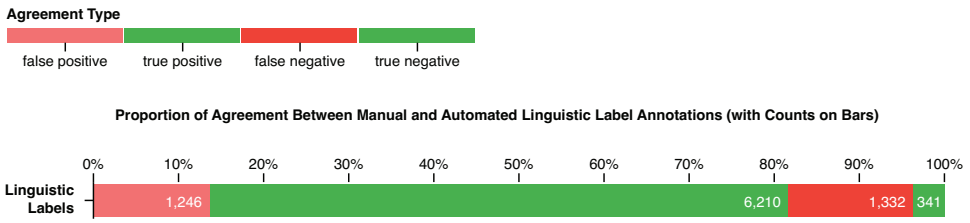


Figure K.5: **Classification model performance on the Linguistic category of labels.** Models' performance as measured with standard NLP metrics (false positive, true positive, false negative, and true negative) on the ***Linguistic*** category, which contains the *Gendered Pronoun*, *Gendered Role*, and *Generalization* labels. Green indicates correctly applied or unapplied labels; red indicates mistakenly applied or missed labels.

literature review and participatory action research informed our definitions of types of gender biased language, which guided five annotators in labeling archival metadata descriptions (Havens, Terras, et al., 2022). Following a supervised approach to training NLP models, we applied several algorithms to the annotated data, training token, sequence, and document classification models to identify the gender biased language that had been annotated manually. We used traditional ML models (Pedregosa et al., 2011) due to documented biases in Deep Learning models (Sharma et al., 2021; Tan and Celis, 2019). The models classify gendered terms (e.g., "she," "Sir") to quantify gender representation across a catalog, as well as gender biased language (e.g., someone referred to only as "his wife") to indicate how descriptive language may misrepresent or exclude people. Figure 4 provides an example of the analysis possible with our models' output. Our presentation will report further detail on the performance of the classification models, including evaluation with NLP metrics (see Figure K.5) and members of HC.

K.4 Discussion

We aim to create NLP models that support HC's effort to mitigate gender bias in its Archives' catalog's descriptive metadata. The process of applying NLP methods to archival descriptions highlighted opportunities for GLAM as well as limitations with NLP methods. For instance, grammatical gender in text does not correspond one-to-one with gender identities, so communications about model findings must clearly explain the uncertainty around gender in language. ML offers promising tools for supporting GLAM documentation practices, and approaches from Archival Science and the Humanities more broadly offer ways to address the complexities of data that are missing from ML. Through the collaborative creation of gender bias classification models, we illustrate the urgency of prioritizing Humanities' ways of thinking in ML research, complementing Digital Humanities with Humanistic Computation.