Collaboration Across the Archival and Computational Sciences to Address Legacies of Gender Bias in Descriptive Metadata

Havens, Lucy

lucy.havens@ed.ac.uk
School of Informatics; University of Edinburgh, United Kingdom

Hosker, Rachel

rachel.hosker@ed.ac.uk
Heritage Collections; University of Edinburgh, United Kingdom

Alex, Beatrice

balex@ed.ac.uk
School of Informatics; Edinburgh Futures Institute; School of Literatures, Languages and Cultures; University of Edinburgh, United Kingdom

Bach, Benjamin

bbach@inf.ed.ac.uk
School of Informatics; University of Edinburgh, United Kingdom

Terras, Melissa

mterras@ed.ac.uk
College of Arts, Humanities and Social Sciences; University of Edinburgh, United Kingdom

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, 2022), 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.

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 (Greenburg, Spurgin, and Crystal, 2005; Harper, 2016; Padilla, 2019; Cordell, 2020).

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 (Sweeney, 2013; Blodgett et al., 2020; Stańczak and Augenstein, 2021). Recently, more critical approa-ches to dataset and model creation encourage interdisciplinary col-laboration and greater transparency in documentation practices to address the harmful biases of ML models (Crawford, 2017; Mit-chell et al., 2018; Havens et al., 2020; Bender et al., 2021). The longer history of classification in the GLAM sector has much to offer the younger ML disciplines.

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 1) and nontextual 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.

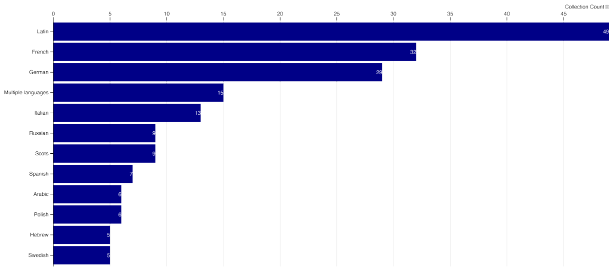
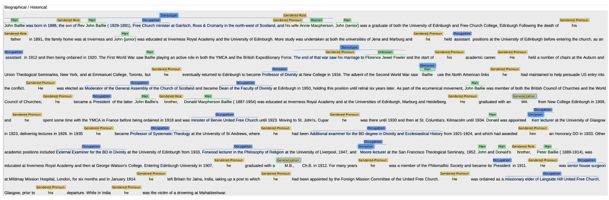


Figure 1. Languages of material documented in the Archives catalog. Most of the HC’s Archives are material written in English (i.e., news articles, manuscripts such as letters, lecture notes, degree awards), however other languages also appear in the Archives (as well as non-textual material such as photographs, sketches, and architectural plans).


Biographical / Historical:

John Baillie was born in 1886, the son of Rev John Baillie (1829-1891), Free Church minister at Gairloch, Ross & Cromarty in the north-west of Scotland, and his wife Annie Macpherson. John (senior) was a graduate of both the University of Edinburgh and Free Church College, Edinburgh Following the death of his father in 1891, the family home was at Inverness and John (junior) was educated at Inverness Royal Academy and the University of Edinburgh. More study was undertaken at both the universities of Jena and Marburg and he held assistant positions at the University of Edinburgh before entering the church, as an assistant in 1912 and then being ordained in 1920. The First World War saw Baillie playing an active role in both the YMCA and the British Expeditionary Force. The end of that war saw his marriage to Florence Jewel Fowler and the start of his academic career. He held a number of chairs at the Auburn and Union Theological Seminaries, New York, and at Emmanuel College, Toronto, but he eventually returned to Edinburgh to become Professor of Divinity at New College in 1934. The advent of the Second World War saw Baillie use the North American links he had maintained to help persuade US entry into the conflict. He was elected as Moderator of the General Assembly of the Church of Scotland and became Dean of the Faculty of Divinity at Edinburgh in 1950, holding this position until retrial six years later. As part of the ecumenical movement, John Baillie was member of both the British Council of Churches and the World Council of Churches; he became a President of the latter. John Baillie's brother, Donald Macpherson Baillie (1887-1954) was educated at Inverness Royal Academy and at the Universities of Edinburgh, Marburg and Heidelberg. He graduated with an MA from New College Edinburgh in 1909, and he spent some time with the YMCA in France before being ordained in 1918 and was minister of Bervie United Free Church until 1923. Moving to St. John's, Cupar he was there until 1930 and then at St. Columba's, Kilmacolm until 1934. Donald was appointed Kerr lecturer at the University of Glasgow in 1923, delivering lectures in 1926. In 1935 he became Professor of Systematic Theology at the University of St Andrews, where he had been Additional examiner for the BD degree in Divinity and Ecclesiastical History from 1921-1924, and which had awarded him an Honorary DD in 1933. Other academic positions included External Examiner for the BD in Divinity at the University of Edinburgh from 1933, Forwood lecturer in the Philosophy of Religion at the University of Liverpool, 1947, and Moore lecturer at the San Francisco Theological Seminary, 1952. John and Donald's brother, Peter Baillie (1889-1914), was educated at Inverness Royal Academy and then at George Watson's College. Entering Edinburgh University in 1907, he graduated with a M.B., Ch.B. in 1912. For many years he was a member of the Philomathic Society and became its President in 1911. He was senior house surgeon at Mildmay Mission Hospital, London, for six months and in January 1914 he left Britain for Jalna, India, taking up a post to which he had been appointed by the Foreign Mission Committee of the United Free Church. He was ordained as a missionary elder of Langside Hill United Free Church, Glasgow, prior to his departure. While in India he was the victim of a drowning at Mahableshwar.

Figure 2. Using the brat rapid annotation tool for manual annotation. Annotators labeled text spans of one or more words with the brat rapid annotation tool (Stenetorp et al., 2011) using eleven labels that were color coded by label category: yellow was Linguistic, blue was Contextual, and green was Person Name.

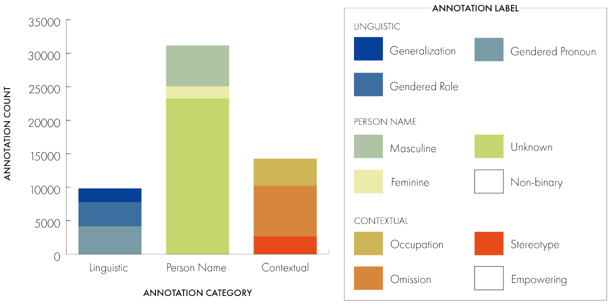


Figure 3. Annotated dataset summary. Five annotators annotated a corpus of 399,957 words across 11,888 descriptions in 245 fonds (collections), resulting in a total of 55,260 annotations. The annotated dataset represents 10% of the entire Archives catalog. Non-binary and Empowering both have a count of zero. (Figure reproduced with author permission from Havens et al., 2022.)

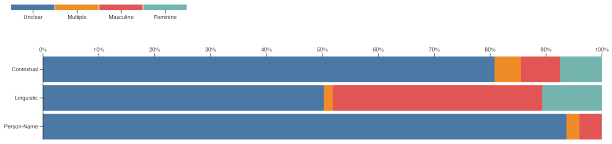


Figure 4. Grammatical gender associations of the Stereotype label. The proportions of each annotator’s labels for the Contextual category that are associated with masculine (blue), feminine (orange), or multiple genders (red), or an unclear association (turquoise). Note: The Person Name annotation category includes the label “Non-binary,” 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.

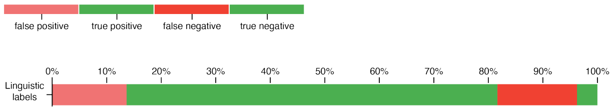


Figure 5. Classification model performance on the Linguistic category of labels. Models’ performance as measured with standard NLP metrics (false positive, true positive, false negative, and true negative) on the Linguistic category, which contains the Gendered Pronoun, Gendered Role, and Generalization labels. Green indicates correctly applied or unapplied labels; red indicates mistakenly applied or missed labels.

The case study consists of four steps: define types of gender bias; create a dataset annotated for gender biases (see Figures 2 and 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 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 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 (Scikit learn, 2023) due to documented biases in Deep Learning models (Tan and Celis, 2019; Sharma et al. 2020). 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 5) and members of HC.

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.

Bibliography

Bender, Emily M. / Friedman, Batya / McMillan-Major, Angelina (2021): “A Guide for Writing Data Statements in Natural Language Processing” http://techpolicylab.uw.edu/data-statements/ [28.04.2023].

Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina (2023): “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. arXiv.org. DOI: 10.48550/arxiv.1810.04805.

Blodgett, Su Lin / Barocas, Solon / Daumé III, Hal / Wallach, Hanna (2020): “Language (Technology) Is Power: A Critical Survey of ‘Bias’ in NLP”, in: *Association for Computational Linguistics (ed.): Proceedings of the 58th Annual Meeting of* *the Association for Computational Linguistics*, Online, July 2020: 5454–76. DOI: 10.18653/v1/2020.acl-main.485.

Caswell, Michelle (2019): “Dusting for Fingerprints: Introducing Feminist Standpoint Appraisal”, in: *Journal of Critical Library and Information Studies* 3, 2, October 2020. DOI: 10.24242/jclis.v3i2.113.

Cordell, Ryan (2020): “Machine Learning + Libraries: A Report on the State of the Field”, Library of Congress https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf?loclr=blogsig [28.04.2023].

Devinney, Hannah / Björklund, Jenny / Björklund, Henrik (2022): “Theories of ‘Gender’ in NLP Bias Research”, in: *Association for Computing Machinery (ed.): FAccT ’22: ACM Conference on Fairness, Accountability, and Transparency*, June 2022: 2083–2101. DOI: 10.1145/3531146.3534627.

Duff, Wendy M. / Harris, Verne (2002): “Stories and Names: Archival Description as Narrating Records and Constructing Meanings”, in: *Archival Science* 2, September 2002: 263–85. DOI: 10.1007/BF02435625.

Greenburg, Jane / Spurgin, Kristina / Crystal, Abe / Cronquist, Michelle / Wilson, Amanda (2005): “Final Report for the AMeGA (Automatic Metadata Generation Applications) Project”. Library of Congress https://www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf [28.04.2023].

Harper, Corey A (2016): “Metadata Analytics, Visualization, and Optimization: Experiments in Statistical Analysis of the Digital Public Library of America (DPLA)”, in: *Code{4}Lib Journal* 33, July 2016 http://journal.code4lib.org/articles/11752 [28.04.2023].

Havens, Lucy / Terras, Melissa / Bach, Benjamin / Alex, Beatrice (2020): “Situated Data, Situated Systems: A Methodology to Engage with Power Relations in Natural Language Processing Research”, in: *Association for Computational Linguistics (ed.): Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, Barcelona, Spain (Online), 2020: 107–24. https://aclanthology.org/2020.gebnlp-1.10 [28.04.2023].

Havens, Lucy / Terras, Melissa / Bach, Benjamin / Alex, Beatrice (2022): “Uncertainty and Inclusivity in Gender Bias Annotation: An Annotation Taxonomy and Annotated Datasets of British English Text”, in: *Association for Computational Linguistics (ed.): Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, Seattle, WA, July 2022: 30–57. DOI: 10.18653/v1/2022.gebnlp-1.4.

Heritage Collections, University of Edinburgh. *Archives Online*. [28.04.2023].

Jo, Eun Seo / Gebru, Timnit (2020): “Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning”, in: *Association for Computing Machinery (ed.): FAT* ’20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*: New York, NY, January 2020: 306–16. DOI: 10.1145/3351095.3372829.

McGillivray, Barbara / Beavan, David / Ciula, Arianna / Colavizza, Giovanni / Cummings, James / De Roure, David / Farquhar, Adam / Hengchen, Simon / Lang, Anouk / Loxley, James / Goudarouli, Eirini / Nanni, Federico / Nini, Andrea / Nyhan, Julianne / Osborne, Nicola / Poibeau, Thierry / Ridge, Mia / Ranade, Sonia / Smithies, James / Terras, Melissa / Vlachidis, Andreas / Wilcox, Pip (2020): “The Challenges and Prospects of the Intersection of Humanities and Data Science: A White Paper from The Alan Turing Institute”, *Figshare*. DOI: 10.6084/m9.figshare.12732164.

Padilla, Thomas (2019): “Responsible Operations: Data Science, Machine Learning, and AI in Libraries”, *OCLC Research*. DOI: 10.25333/xk7z-9g97.

Rogers, Anna (2021): “Changing the World by Changing the Data”, in: *Association for Computational Linguistics (ed.).: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* 1, Long Papers, Online, August 2021: 2182–94. DOI: 10.18653/v1/2021.acl-long.170.

Pedregosa, Fabian / Varoquaux, Gaël / Gramfort, Alexandre / Michel, Vincent / Thirion, Bertrand / Grisel, Olivier / Blondel, Mathieu / Prettenhofer, Peter / Weiss, Ron / Dubourg, Vincent / Vanderplas, Jake / Passos, Alexandre / Cournapeau, David / Brucher, Matthieu / Perrot, Matthieu / Duchesnay, Édouard (2011): “Scikit-learn: Machine Learning Research in Python”, in: *Association for Computing Machinery (ed.): The Journal of Machine Learning Research* 12, 85: 2825–2830. DOI: 10.5555/1953048.2078195.

Raji, Deborah / Denton, Emily / Bender, Emily M. / Hanna, Alex / Paullada, Amandalynne (2021): “AI and the Everything in the Whole Wide World Benchmark”, in: *Curran (ed.): Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, Online: 2–20. DOI: 10.48550/arXiv.2111.15366.

Stenetorp, Pontus / Goran, Topić / Pyysalo, Sampo / Ohta, Tomoko / Kim, Jin-Dong / Tsujii, Jun’ichi (2011): “BioNLP Shared Task 2011: Supporting Resources”, in: *Association for Computational Linguistics (ed.): Proceedings of BioNLP Shared Task 2011 Workshop*, Portland, OR, June 2011: 112–120 https://aclanthology.org/W11-1816 [28.04.2023].

Sharma, Shanya / Dey, Manan / Sinha, Koustuv (2021): “Evaluating Gender Bias in Natural Language Inference”, in: *Curran Associates, Inc. (ed.): NeurIPS 2020 Workshop on Dataset Curation and Security*, Online. DOI: 10.48550/arXiv.2105.05541.

Sta#czak, Karolina, and Isabelle Augenstein (2021) “A Survey on Gender Bias in Natural Language Processing”, in: *arXiv.org*. DOI: 10.48550/arXiv.2112.14168.

Sweeney, Latanya (2013): “Discrimination in online ad delivery”, in: *Association for Computing Machinery (ed.): Communications of the ACM* 56, 5: 44–54. DOI: 10.1145/2447976.2447990.

Tai, Jessica (2021): “Cultural Humility as a Framework for Anti-Oppressive Archival Description”, in *Journal of Critical Library and Information Studies* 3, 2, October 2021. DOI: 10.24242/jclis.v3i2.120.

Tan, Yi Chern / Celis, L. Elisa (2019): “Assessing Social and Intersectional Biases in Contextualized Word Representations”, in: *Wallach, H / Larochelle, H. / Beygelzimer A. / d'Alché-Buc, F. / Fox, E. / Garnett, R. (ed.): Advances in Neural Information Processing Systems* 32. DOI: 10.48550/arXiv.1911.01485