CHAPTER 7

Revisiting Indian Ocean History in the Face of the Digital Turn

Mrinalini Luthra, Matthias van Rossum, and Manjusha Kuruppath

Abstract

This chapter focuses on the GLOBALISE project's digitization efforts to innovate how the Dutch East India Company (VOC) archives are accessed, read and researched. GLOBALISE is creating an infrastructure that will allow individuals to read these colonial archives in machine readable text, search the archives for entities and events supplemented with additional historical context, and query relationships between them. To reconceptualize the practice of history in the context of the Indian Ocean world and in face of the digital turn as embodied by GLOBALISE, a key challenge is to decolonize the archive. This needs to be done by critically reviewing the use of the archives, addressing the biases and relational aspects of such digital archival infrastructures, the data derived from it, and the technology used to unlock them. This paper reflects on how we develop critical and inclusive practices of (digital) history writing that mobilizes colonial archives and new techniques, while accounting for the plurality of non-European histories and perspectives? It discusses how we should address colonial and technological biases and how we can spur the writing of Indian Ocean histories in new and desirable directions.

Keywords: VOC, GLOBALISE, early modern, Indian Ocean, digitization, archival projects, bias, decolonization

1. Introduction

This chapter focuses on the GLOBALISE project, which aims to innovate historical research practices by rethinking the ways in which the Dutch East India Company (VOC) archives are accessed, read and researched. The VOC, as an aggressive colonial power in the Indian Ocean and East Asian Seas region in the 17th and 18th centuries, created archives which are a resource of irrefutable value, offering profound insights into the history of colonial expansion and the societies and polities that both endured and resisted the VOC's dominance. GLOBALISE will create an infrastructure that will enable researchers and the wider public to navigate these archives in machine readable text, allowing them to explore and search, and uncover entities, events and

their interconnections, all enriched by additional layers of historical context. Despite its promises, these possibilities for new research practices come with the caveat that neither sources, technology nor history writing/making are 'neutral'. The VOC archives serve as poignant reflections of the colonial agenda, granting only limited visibility and agency to non-European actors. Moreover, existing historiographies, source publications, and data collections that draw from these archives often perpetuate the very epistemologies rooted in colonialism. Furthermore, technologies such as language models, intended for automated information extraction, bear the risk of encoding biases. To reconceptualise the practice of history in the context of the wider Indian Ocean world and in the face of the digital turn, a key challenge is to decolonise archival and research practices, envisioning new frameworks that transcend the confines of colonial legacies.

This touches upon several pressing questions for writing history in the 21st century. Most importantly, how do we develop critical and inclusive practices of (digital) history writing about this wider Indian Ocean world that mobilises the potential of colonial archives and new techniques, while accounting for the plurality of non-European histories and perspectives? How can we navigate the complexities of both colonial and technological biases, all while avoiding the pitfalls of earlier strands of historiographical approaches such as macro-comparative methods? Furthermore, how can we spur the writing of Indian Ocean histories in new and desirable directions?

Using the GLOBALISE project as a case in point, we will reflect upon these issues by critically examining archival sources and inventories, archive derived datasets, the technologies used to unlock these archives, the biases and relational aspects inherent in such digital archival infrastructures. Our exploration will focus on the presence of biases specifically related to historical persons, particularly those outside the European context, within the archive. We will analyse concrete instances of these biases in three domains crucial for the GLOBALISE infrastructure: archives and inventories, datasets, and digital infrastructure. Beyond merely acknowledging these problems, we will delve into their implications for knowledge production and the writing of history, while also exploring avenues to counterbalance and resolve these biases.

2. GLOBALISE: enabling in-depth exploration of the VOC Archives

The GLOBALISE project aims to create an innovative research infrastructure, granting unprecedented access to facilitate research into VOC archives.[1] These archives, one of the most complete and extensive bodies of sources on early modern world

[1] Petram and Van Rossum, 'Transforming Historical Research Practices'.

history, offer a distinctive lens through which to view the histories and cultures of Europe, Africa, Asia, and Australia. From the early 17th century onwards, the VOC established an empire that transcended borders, merging the establishment of influential trading posts across a vast expanse from Persia to Japan. Their strategic conquests extended to pivotal ports such as Malacca and Macassar, while exerting extensive territorial control over Taiwan (Formosa), the Moluccas, Banda, Java, Sri Lanka (Ceylon), Malabar, and South Africa. The Company archives that were created in the course of its activities are a source for the study of not only the VOC world itself, but crucially also for exploring the all too often under-documented societies of early modern Africa, Asia and Australia, and the many local, regional and global interactions between them.

Because of their huge potential, the VOC archives have captured sustained international scholarly interest since the 19th century. Recent strides in digitisation, aided by the publication of related source editions, research aids, datasets, and the National Archives' digitisation efforts, and even first efforts in text recognition, have enhanced accessibility of the VOC archives.[2] However, the path to conducting large-scale and comprehensive research within these archives remains a daunting endeavour. Fragmented accessibility and the absence of transcriptions for most scans impede the exploitation of the vast archival corpus. Keyword searches remain elusive, confining navigation through digitised archival material to traditional tools like inventories and indexes. Extracting the contents of scanned materials still necessitates the meticulous perusal of each page by human eyes.

GLOBALISE aims to tackle some of these key constraints to enhance researchability of the digitised VOC Archives. While large-scale digitisation has revolutionised *accessibility* of source materials, further steps are necessary to improve its searchability, let alone its *researchability*. Through a combination of advanced technologies and contextualisation methods, GLOBALISE aims to enable advanced querying capabilities and foster the writing of diverse and representative histories based on the VOC archives.

At the heart of the project lies the endeavour to make the contents of the archival source material, particularly the 'Overgekomen Brieven en Papieren' ('Letters and Papers received', OBP) series, available in a complete handwritten text recognition (HTR) transcription. Subsequently, a dynamic interplay of historical and semantic contextualisation methods is employed. Entities, events, and their contexts are recognised, identified and semantically modelled, enabling complex querying and in-depth analysis of the archival corpus through a user-friendly

[2] The first million scans of the Overgekomen Brieven and Papieren corpus in the VOC archive were text recognised by the Dutch National Archives in the project *IJsberg Zichtbaar Maken* [Making the Iceberg Visible].

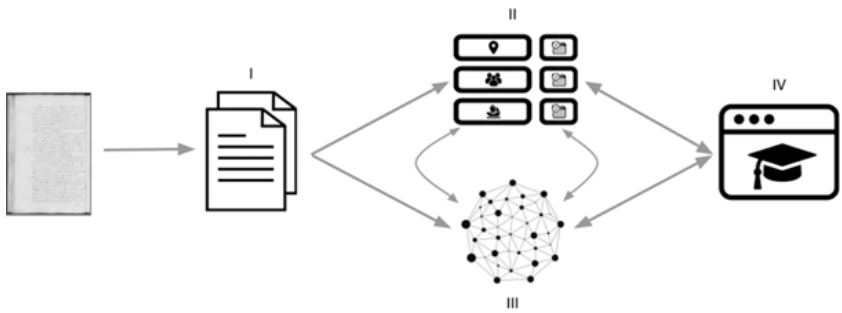


Figure 7.1. The GLOBALISE process from scan to handwritten text recognition (I), and onwards through a process of historical contextualisation (II) interacting with semantic contextualisation (III) to a user-friendly infrastructure for complex querying (IV).

research environment. Distinct yet interconnected, the historical and semantic contextualisation layers play pivotal roles in the research infrastructure. The historical contextualisation layer functions as a 'digital historical encyclopaedia', linking entities mentioned in the source texts to a comprehensive collection of curated reference data sourced from existing VOC data and research aids. On the other hand, the project develops a semantic contextualisation layer, training natural language processing (NLP) models to identify entities, events, and dates in the text while extracting the associated contexts. This includes not only recognising a person's mention but also extracting their actions, attributes, and temporal details. The data extracted from the documents is structured using a dedicated formal ontology, enabling both human and computational reasoning over the vast complexity of information.

By interweaving these historical and semantic contextualisation layers, the GLOBALISE project constructs a research infrastructure that facilitates sophisticated and nuanced querying of the VOC material. Through the semantic contextualisation layer, researchers can effortlessly discover references to specific events such as commodity production or trade at particular times and places with a single query (Figure 7.1, III). These query results are complemented by validated reference data from the historical contextualisation layer (Figure 7.1, II), elevating the significance of textual mentions to relevant 'historical observations'. A user-friendly interface atop this research infrastructure empowers researchers to explore, query, visualise, and analyse the data, opening new avenues for historical research and understanding (Figure 7.1, IV).

GLOBALISE aspires not only to intervene in the realm of infrastructure and archival practices but also to make significant contributions to digital humanities, historiography, and research perspectives. By advancing researchability and embedding historical contextualisation within semantic frameworks, the project

endeavours to contribute to reshaping the landscape of historical inquiry and unlock the hidden narratives within the VOC Archives.

3. Bias in the archives and inventories

The project focuses on the OBP series which is one of the most interesting and important document series in the VOC archives. The selection that the project works with comprises approximately 6900 inventories of more than 5 million handwritten pages. The OBP contains correspondence and copied local administrations received from Company settlements in the wider Indian Ocean region. The collections of documents were copied and sent to the Company Chambers in the Dutch Republic to provide information to the board of directors (bewindhebbers) and serve as institutional memory. The series of 'copied letters' (Overgekomen Brieven and Papieren) that are best preserved are those that were dispatched to the Company chambers of Amsterdam and Zeeland.[3] The archives bears a multitude of genres of reporting, such as the General Letters which were summaries of developments across the VOC's region of operations drafted in Batavia, reports exchanged between Batavia and those settlements which fell under its jurisdiction and copies of civil and criminal proceedings of the Council of Justice in Batavia. Some of these genres of reports are high-level correspondence and provide highly aggregated, sometimes banal, yet important summaries of happenings. Other genres such as letters of local merchants and rulers, or Company court records, can often provide voices and information that are alternative or even conflicting to what can be regarded as the official Company narratives. Additionally, selections of subaltern correspondence such as daily journals which bristle in the details of the everyday also feature in this archive thereby providing a rich medley of reports offering various kinds of information about interactions involving European and non-European actors throughout the Indian Ocean world in the 17th and 18th centuries.

While the GLOBALISE project offers us the possibility of restructuring access to the archive and how the corpus is searched by historians, it forces us to ask epistemological questions to reflect on the existing archive, its offerings, and drawbacks when accessing the archive. As a historical source, the OBP is a paradoxical complex of problems and possibilities. It holds the promise of presenting a deep technicolour view of colonial and other actors in the Indian Ocean World that is potentially like no other. Its pages are populated with mentions of sailors

[3] The VOC archive also features correspondence received by other Company chambers, but these are smaller collections. The OBP series of the Chambers Amsterdam and Zeeland complement each other and partly overlap.

and fishing communities and speak of merchant groups, agriculturalists, forced labourers and enslaved persons. This archive also tells of the diplomatic acts and antagonisms of its geopolitical actors. It holds information for a wide array of societies and regions for which in many cases little other written source material has been preserved. Just as this archive overflows with the potential of being a rich source for the writing of early modern Indian Ocean histories of non-European agents and societies, there is no circumventing the fact that this archive is first and foremost a record of Company ambitions, anxieties, prejudices and perspectives. Several decades of scholarship, particularly from the legions of post-colonial theory have reflected on the relations of power that dictate the creation of colonial texts and their interpretations and instruct on how to critically interrogate the agendas, gaps and silences in these works.[4] Evidently constructed with the intention of furthering the Company's interests, the archive casts the Company and its employees as the principal protagonists in their record of events. Although non-European subjects occasionally come to the fore, they mostly lurk in the background of these narratives. The archive is thereby ridden by the curious paradox of the non-European subjects being present in the archive but not being seen.

These are examples of the perspectival distortion of colonial archives, but what truly is the consequence of this epistemological ordering of the archive that privileges specific narratives and sections of people over others? Where alternative non-European sources are lacking that can either provide the large volume of information of the VOC archives or provide the detailed narratives about the themes such as trade, work and everyday life that are reflected in the VOC archives, there is the danger that the Company archives provide a virtually hegemonic construction of the past that is difficult to critically appraise and see beyond. The effects of this dominance have been acutely felt in the realm of historiography. Historiography has cast the same European actors as their protagonists as foregrounded by the colonial archives and Company histories and the histories of its European employees have mostly dominated the historiographical scene.[5] This has clouded the perspective on the VOC, who for too long have been seen as merely 'merchants' instead of colonisers, and on those who were affected by its activities, as these were not only European sailors, soldiers, artisans and officials, but also the Asian and African employees (sailors, soldiers, artisans) and enslaved, local colonised populations and other non-European societies. Only in the last decades have historians exploited the potential of the archives in unearthing biographies

[4] Landmark works include Said, *Orientalism*; Stoler, *Along the Archival Grain*.

[5] Some examples of the sustained and unwavering interest in the history of the Company's mostly European servants are Nierstrasz, *In the Shadow of the Company*; and Bruijn, *Ship's Surgeons of the Dutch East India Company*.

of non-Europeans who feature in the archive, and recovering histories of enslavement, gender and labour relations.[6] This historiographical failing is mostly to be attributed to the nature of the archive itself.[7] However, this shortcoming is magnified by a traditional access point to the archive: archival inventories.[8]

Inventories and Inadvertent Histories

Why is it important that we reflect on the inventories – the seemingly inane, harmless signposts to a sea of records of archives which would otherwise be virtually inaccessible? The archives possess tables of contents that were either drawn up in the 17th and 18th centuries when the Company ledgers were compiled, or written out by archivists in the 19th century. These tables of contents that feature in the leather bound company records were first compiled into a typescript inventory from the 1970s onwards.[9] This endeavour was subsequently expanded and taken over by the TANAP project which released their index (a compilation of the table of contents of a large section of the Dutch East India Company archives) at the turn of the 21st century.[10] The tables of contents were, like the archive itself, drawn up in the interest of the archive creators, and not for the scrutiny of historians and the public. This impacts the information it bears.

In the context of inventorisation, three principal problems come to the fore:

a) **Double invisibilisation:** The kind of information that tables of contents provide and how informative they are about the contents of the reports they serve to introduce varies wildly in the archive. Admittedly, there are titles in the archive that are extremely revealing of reports' contents like the 'Authentic translation of a Caul in Malay granted by the free 'Boschheer' Nella Maijga Calaka Torin in

[6] In the context of subaltern gender, see for instance Jones, *Wives, Slaves, and Concubines*; Andaya, *The Flaming Womb*; and Taylor, *The Social World of Batavia*.

[7] The historiographical blind spots caused by the nature of the archive can be remedied by the innovative strategies recommended by post-colonial historians about interrogating the agendas and reading through the silences and gaps in the archive.

[8] Yet another portal into the archive that has long determined the nature of historical research are source publications. Source publications such as the published volumes of the Generale Missiven (General Letters) have constituted the gateway to the archive and for scores of historians, the General Letters was the oracle that determined whether or not a certain theme was researchable in the VOC archives. We are unable to discuss source publications in this article owing to word limit constraints.

[9] We thank Lennart Bes for this information.

[10] The TANAP website is no longer online but the inventories and other useful information that featured in the TANAP website can be viewed and downloaded from https://docs.globalise.huygens.knaw.nl/tanap/. The TANAP index can also be found at https://www.nationaalarchief.nl/onderzoeken/index/nt00348.

Odiaarpalium to the Company in the name of Jan Koort, merchant and head in Tegenepatnam…dated 23 October 1688'[11] However, such detailed descriptions are hard to come by and the archive mostly teems with descriptions of another kind. One is confronted with titles like the twenty-page report titled 'Missive van den Gouverneur (Ternate) van den 15 September 1767' which is so opaque that it only reveals details about the nature of correspondence, date of composition and authorship.[12] If only the historian had the resoluteness to look beyond the description and read the pages of the report would they realise that it makes mention of the enslaved woman Soengiraja who was gifted to the Company by Prince Terlucco of Gorontalo in Sulawesi.[13] If the archive therefore invisibilises non-Europeans owing to a bias in subject matter, inventories subject them to a second invisibilisation. Another example of invisibilisation is evident in the report plainly labelled 'Some extracts from the daily journal narrating the most notable news from the court etc from 27th January to 1st May 1740.'[14] Such a description belies the fact that this report bears valuable information about the correspondents of the Dutch factory in Surat like Sampatram who shared relevant information about earth-shattering events that transpired in the Mughal capital of Delhi in the wake of an invasion.[15]

[11] 'Authenticque translaat van een caul int Maleijts verleent door den vrij boschheer Nella Maijga Calaka Torin in Odiaarpalium aan de Comp. op de naam van Jan Korte koopman en opperhooft tot Tegenepatnam [en Porto Novo, gedateerd 23 October 1688]'. N.A. 1.04.02, Inv. No. 1477, Nationaal Archief, Den Haag, Verenigde Oostindische Compagnie (VOC). For the actual report, https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/1477/file/NL-HaNA_1.04.02_1477_1429; For the title of the report as it features in the table of contents, https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/1477/file/NL-HaNA_1.04.02_1477_0013?page=701. Also see TANAP inventory on the Coromandel, p. 426. https://web.archive.org/web/20180130120355fw_/http://databases.tanap.net/vocrecords/pdf/files/C7B6570AE6B11A4E541912B3507DCFE1.pdf. Accessed on 16th July, 2023. We reckon that the document was originally written in Malabaarse (mostly likely Tamil) and not Malaeijts as written by the scribe. We thank Tom Hoogervorst for his input on the issue.

[12] 'Missive van den Gouverneur (Ternate)' (15 September 1767), N.A. 1.04.02, Inv. No. 8135, Nationaal Archief, Den Haag, Verenigde Oostindische Compagnie (VOC), For the table of contents see, https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/8135/file/NL-HaNA_1.04.02_8135_0005. For citation, see https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/8135/file/NL-HaNA_1.04.02_8135_0188.

[13] 'Missive van den Gouverneur (Ternate)'.

[14] 'Eenige extracten uijt het lopend dagregister verhalende het notabelste hofnieuws etcetera van den 27 Januarij tot den primo Maij 1740'. Inv. No. 2509, Nationaal Archief, Den Haag, Verenigde Oostindische Compagnie (VOC), For the table of contents, see https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2509/file/NL-HaNA_1.04.02_2509_0008; For the narrative, see https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2509/file/NL-HaNA_1.04.02_2509_1468.

[15] 'Eenige extracten uijt het lopend dagregister verhalende het notabelste hofnieuws etcetera van den 27 Januarij tot den primo Maij 1740'.

b) **Deceptive categorisation**: The inventory is also ridden by instances of miscategorisation. While it is indeed possible to speak about a typology of reports within the archive which can have uniform structure and content, categories like the daily journals which are narratives about day-to-day trading, political, weather events and other occurrences in company settlements or aboard naval vessels are in fact sometimes amalgamations of a variety of reports. The four hundred page 'Copy of the Daily journal' of Cochin in 1756–1757 for instance carries more than a hundred letters of Company correspondence with nobility and merchants in Malabar.[16] This renders these letters unlocatable through the use of the inventory alone.
c) **Titular discrepancies**: There are often glaring inconsistencies in the titles of documents as they feature in the table of contents and in the actual documents themselves. The aforementioned daily journal from Cochin is one such instance. Interestingly, the book bearing this report carries three descriptions of the report of which the introductory table of contents to the book merely lists the report to be a daily journal. While the second title is more illuminating of the contents of the report, it is the title in the actual report itself that is most representative of the contents of the document indicating that it bears 'not only all incoming and outgoing but also other papers and court affairs between the kings and rulers inside and outside the city of Cochin with the departing and new Company Commanders [paraphrased].'[17]

These shortcomings evidence the fact that archival inventories do not facilitate the historical scrutiny and the recovery of non-European subjects that it has the potential to allow. Owing to the pervasive impact that the accessibility of the archive

[16] 'Copia Dagverhaal Beginnende Met Primo 8ber 1756 en Eijndigende met 23 Augustus 1757', N.A. 1.04.02, Inv. no. 2906, Nationaal Archief, Den Haag, Verenigde Oostindische Compagnie (VOC). For the table of contents, see https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2906/file/NL-HaNA_1.04.02_2906_0005. For the title of the actual report, see https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2906/file/NL-HaNA_1.04.02_2906_0133.

[17] For entry on the report in the TANAP Inventory, see page 1 of the Malabar inventory. https://web.archive.org/web/20180130120436fw_/http://databases.tanap.net/vocrecords/pdf/files/7D56909EE586A-5B792E2F38DA875C1BA.pdf. Accessed on 16th July 2023. The complete title of the report reads 'Inlands dagregister waar in, op hun tijd aangetekent ende g' extendeert staan, niet alleen alle afgaand en aankomende olaas maar ook andere papieren, en hofsche verhandelingen tusschen de koningen, en vorsten binnen als buiten de stad Cochin met de E.E. aftb heeren Fredrik Cunes afgaand Casparus de Jong aankomend Commandeurs en oppergebieders gegouden, aanvank neemende met' N.A. 1.04.02, Inv. no. 2906, p. 1. https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2906/file/NL-HaNA_1.04.02_2906_0133. For the second title of the report, see N.A. 1.04.02, Inv. no. 2906, folio unnumbered. https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/2906/file/NL-HaNA_1.04.02_2906_0122.

has had in determining the nature of historical research that is undertaken and the resulting imagination of the historical past, a radical shake-up involving the reimagination and reorientation of access to the archives is necessary. What can the GLOBALISE project do to remedy the situation?

a) **Free Keyword Search**: One of the aims of the project is to make the entire archive accessible through a free keyword search. Furthermore, as part of our work in historical and semantic contextualisation, we will record all possible name and spelling variations for a clearly defined set of entities which will include persons names. Consequently, the search results for 'Bindrabon' who served as a Company broker in Bengal in the mid-18th century will not only yield the results for this name in the archive. The search results will also scour the archives and generate results for all other recorded variations of the same name which will include 'Brindhabon, Brindabonnaij and Brindabon.[18] The project will transform searchability in such a way that inventories no longer need to be relied on to comb the archive for content. It is however important to note that our own efforts too cannot be free from bias. Firstly, searchability within the archive will hinge on the quality of the machine-readable transcripts. Secondly, the references to persons in the archive are so numerous that our own efforts to document name and spelling variations of person names will most certainly be incomplete. The efficacy of this task will additionally depend on the expertise of our annotators to be able to link the name variations to the correct referent.

b) Enabling Complex Search: GLOBALISE aims go beyond the simple search and facilitate complex querying. Users can not merely recover mentions of entities such as persons but can also find mentions of persons who feature in the context of a range of events like transportation, revolt, or trade. Enabling such precision in querying should allow for qualitative improvements in the research undertaken into and the recovery of information about individual and group action in the archive, be it the transportation of enslaved persons, or social upheaval in the form of war and rebellion. This will, we reckon, be hugely transformative in facilitating the writing of individual biographies and social histories.

In view of the searchability that GLOBALISE intends to facilitate, will indexes and inventories become absolutely redundant and lose their relevance as a gateway to the archive? The answer is an emphatic no. Inventories are still irreplaceable

[18] These search results were procured using the first version of HTR of the Overgekomen Brieven en Papieren generated by the project. GLOBALISE project, 'VOC Transcriptions – GLOBALISE' (IISH Data Collection, May 1, 2023), https://hdl.handle.net/10622/JCTCJ2.

cogs in the architecture of the archive and provide necessary context for the information procured. They indicate the type of report, provide knowledge about correspondence practices, facilitate trackability of the development of events and event-related information within the archive and reveal crucial links between correspondence and decision making. All of this is vital context for the user of the archive for whom, without these inventories, the archives are merely a non-navigable sea of words. If we agree that archival inventories remain valuable, the project can consider enriching these inventories in two possible ways.

a) **Recovery and enrichment of report titles (and other document metadata)**: As previously mentioned, a noticeable shortcoming of the archival inventories is the occasional lack of correspondence between the titles of reports as they feature in the table of contents and the titles in the report themselves. The project can redress this drawback by providing the original title of the report alongside the title as it features in the table of contents in the archive. This redress will also have the interesting consequence of facilitating research into the compilation of the table of contents, and thereby exposing the chasm that exists between what was recorded in the actual titles of reports and what was interpreted in the table of contents.
b) **Enrichment of inventories**: Together with the title of the report, the GLOBALISE project could provide a summary of the report, and more importantly make available the entity mentions in the report, such as a list of referents. This will ensure that references to individuals that were previously subject to the whim of the compilers of the table of contents and subsequently omitted are rightfully recovered.

4. Data and datasets

The VOC archives has lent itself not only to inventories, indexes and a vast historiography, but also for the creation of a wide range of datasets. Often developed as part of economic, social or to a lesser extent political and cultural historical research, these datasets have tried to use the detailed and oftentimes structured character of the VOC administration to shed light on different aspects of the histories of the VOC and to some extent also the wider colonial and non-Western world. These have, for example, been developed to reconstruct the history of shipping and trade,[19] the flow

[19] For datasets on the Dutch East India Company, see Bruijn, Gaastra, and Schöffer, 'Dutch-Asiatic Shipping in the 17th and 18th Centuries'; Schooneveld-Oosterling and Knaap, 'Boekhouder Generaal; Van Rossum, *Generale Zeemonsterollen VOC, 1691–1791*, https://dutchshipsandsailors.nl/?page_id=11; Menno Leenstra, 'Dutch Ships in Asian Waters in the 17th Century'; Knaap and Sutherland, 'Monsoon

of financial transactions and bullion,[20] labour and labour relations,[21] wages and prices, demographic developments, and more recently slavery and slave trade.[22]

It is crucial to note that all of these datasets were created and shaped by a balancing act between the availability of specific information in the source material and the research interests that produced the need for information. Whereas the information in the archival source material was largely shaped by the contemporary activities of the VOC as a colonial, administrative and mercantile organisation, the research questions were mostly dictated by 20th and 21st century academic and societal concerns. The datasets that have been created based on VOC archives are thus not simply a reflection of what information is available in the source material of the archives, but is just as much determined by what types of information and questions have been deemed relevant, what topics or source material have drawn the attention of scholars, how data was structured and processed, etc.

The influence of the state of historiography on the datasets that have been produced over the last decades, and are now available for the history of the VOC and Indian Ocean, is reflected in their focus and coverage. The lingering historiographic notion that the VOC should primarily be understood as a European 'merchant' company active in Asian trade (as discussed in the previous section) is visible in

Traders: 18th Century Makassar Shipping Data'; Brouwer, 'Non-Western Shipping Movements in the Red Sea and Gulf of Aden during the 2nd and 3rd Decades of the 17th Century, According to the Records of the Dutch East India Company: (Part 2)', Arasaratnam, 'The Dutch East India Company and Its Coromandel Trade 1700–1740'; VOC: Bruijn, Gaastra and Schöffer, *Dutch-Asiatic Shipping in the 17th and 18th Centuries*; Boekhouder General Batavia, Huygens ING, https://bgb.huygens.knaw.nl/; Heerma van Vos et al., *Dutch Ships and Sailors: Generale Zeemonsterrollen VOC as RDF*, https://research.vu.nl/en/datasets/dutch-ships-and-sailors-generale-zeemonsterrollen-voc-as-rdf; Leenstra and Parthesius, *Dutch Ships in Asian Waters in the 17th Century*, https://hdl.handle.net/10622/YP2COP; local shipping: Knaap and Sutherland, *Monsoon Traders*; Knaap, *Shallow Waters, Rising Tide*; Blussé, 'Junks to Java'; Brouwer, 'Non-Western Shipping Movements in the Red Sea and Gulf of Aden During the 2nd and 3rd decades of the 17th Century, According to the Records of the Dutch East India Company'; Chakraborty, *Empire of Labor*; also the works of Arasaratnam, 'The Dutch East India Company and its Coromandel Trade 1700–1740'; Prakash, *Dutch East India Company and the Economy of Bengal, 1630–1720*.

[20] See Glamann, *Dutch-Asiatic Trade 1620–1740*; Jacobs, *Merchant in Asia*.

[21] See for instance Velzen and Gaastra, 'VOC Opvarenden, versie 6 – januari 2012.' For an instance of a dataset on ethnic/multicultural composition of ship crews, see L. Heerma van Voss et al., 'Dutch Ships and Sailors: Generale Zeemonsterrollen VOC as RDF'; and earlier studies by De Hullu, *Op de schepen der Oost-Indische Compagnie*; global collaboratory of labour relations, Heerma van Vos et al., *Dutch Ships and Sailors: Generale Zeemonsterrollen VOC as RDF*, https://research.vu.nl/en/datasets/dutch-ships-and-sailors-generale-zeemonsterrollen-voc-as-rdf, as dataset on ethnic/multicultural composition of ship crews.

[22] E.g. Van Rossum and Tosun, 'Acten van Transport Batavia (1664–1780)'; Rossum et al., 'Slave Transport Permissions – VOC Cochin, 1770–1795'. See also the project *Exploring Slave Trade in Asia*, www.exploringslavetradeinasia.com.

three specific and interrelated biases that characterise the earliest and still most well-known large datasets. These could be described as:

a) **The 'maritime' bias:** there is an abundance of datasets that document shipping movements, maritime workers and maritime trade, shedding light on the functioning of the VOC as a 'merchant', but leaving underrepresented the other key roles of the Company, namely those of 'ruler', 'soldier', and 'producer'.[23] Even though historiography has been slowly shifting its attention to a wider array of topics, this is not yet reflected in the data production. In contrast to maritime trade, shipping and movement of people, information on developments related to war and conquest, colonial rule, slavery and slave trade, labour regimes, production frontiers, and the mostly non-European populations inhabiting the colonised societies by the VOC is almost absent in the available structured data.

b) **The 'intercontinental' bias:** within the vast majority of data on shipping and personnel, there is a heavy bias towards the intercontinental connections and movements. The VOC shipping between Asia and Europe, for example, is completely reconstructed in the *Dutch Asiatic Shipping* (*DAS*).[24] For intra-Asiatic shipping of the VOC there is an abundance of datasets produced to complement *DAS*, but these datasets provide only a partial picture, either through the movement of goods to and from Batavia (*Bookkeeper General Batavia*, *BGB*),[25] through the yearly recordings of VOC personnel at sea (*General Sea Muster Rolls*, *GZM*)[26] or because these only cover a specific period (*Dutch Ships in Asian Waters in the 17th Century*).[27] For financial history, the dataset on bills of exchange (wisselbrieven) reconstructed who was sending back money from Asia to the Dutch Republic.[28] Similarly, for the people working under the VOC, the large *VOC Opvarenden* dataset recorded those (mostly) Europeans that were employed on board ships departing from the Dutch Republic, but there is less data on the experiences of these workers in Asia, and on the contract and enslaved workers that came into the service of the VOC in Africa and Asia (*GZM*).

c) **The 'European' bias**: this is connected to an important focus on the 'European' that has continued to prevail in historiographic and dataset production. With a few exceptions, almost all datasets are concerned with tracing histories of

[23] Van Rossum, 'Labouring Transformations of Amphibious Monsters'.

[24] Bruijn, Gaastra, and Schöffer, 'DAS.'

[25] Schooneveld-Oosterling and Knaap, 'Boekhouder Generaal Batavia: The Circulation of Commodities of the Dutch East India Company in the Seventeenth and Eighteenth Centuries'.

[26] See Voss et al., 'Dutch Ships and Sailors: Generale Zeemonsterrollen VOC as RDF'.

[27] See Leenstra, 'Dutch Ships in Asian Waters in the 17th Century'.

[28] See Department of History European Social Fund/DABURH Leiden University, 'Bills of Exchange of the Dutch East India Company (VOC), 1690–1740' (DANS Data Station Social Sciences and Humanities, March 1, 1995), https://doi.org/10.17026/dans-znr-2nk4.

European personnel, European shipping, and European trade. There have been counter-attempts to produce datasets on e.g., Chinese shipping to Batavia (Blussé), European and non-European shipping movements as administered in specific VOC ports (Java, Coromandel, Yemen, Cochin), and transactions of enslaved (especially South Africa, Cochin), but this has far from balanced the Eurocentric bias in the production of structured data.

This deeply influenced the information that is available on historic persons, both 'who' appear in datasets (and who do not) and 'how' they appear (and how not). Let us consider the example of information on people in the datasets on shipping. In one of the first large and perhaps most influential datasets created by (Dutch) maritime historians, the so-called Dutch-Asiatic shipping database (1968–1978), for example, only information on ships' captains (name and origin) and the size and composition of the crew (aggregate numbers of occupational categories, being divided between sailors, soldiers and artisans) were included.[29] The DAS would thus record a voyage of the VOC ship *Gansenhoef* departing from the Dutch Republic in May 1707 and arriving in Batavia in June 1708. And it would provide the name of the captain (Daniel Engels), but not of the 70 sailors, 58 soldiers and 1 passenger. The DAS would also record the return voyage of the *Gansenhoef* in December 1711 to Rammekens (Zeeland) in August 1712. Recording all these voyages of the VOC between Europe and Asia, and back, the *DAS* created a breakthrough in quantitative economic and social historical scholarship.

This in turn influenced initiatives such as the creation of the VOC Opvarenden database (2000–2010) that recorded information from the Scheepssoldijboeken on personnel recruited in Chamber Cities for outward bound voyages and in Asia. This dataset recorded the names, places of origin, occupations or rank at the time of enrolment, the ship of departure, and information on the return in the Republic (year and ship) or conditions for other ways of contract termination (year, reason and place). The dataset includes 774,574 entries for enlisted personnel, and 75,876 entries for creditors (wives and others) who would receive advances on payment.[30] For most of the outward bound voyages between 1680 and 1795, this dataset would give the entire crew. The dataset is not entirely complete. For the case of the *Gansenhoef*, it records the names of the crews in 1705 and 1721, but it does not provide a voyage in 1707. Most notably, it only records who were supposed to be on board, and their formal or ascribed identities. This could lead to distortions especially in the case of people who were hiding and gave out false identities, but

[29] Sources for this dataset were the 'uitloopboeken', ship registers and documents in the Overgekomen Brieven en Papieren (OBP).

[30] Velzen and Gaastra, 'VOC Opvarenden, versie 6 – januari 2012'.

also people from outside the Dutch (and wider Baltic Dutch) language area whose names and places of origins were more often written down phonetic or simply wrong, and even female travellers and crew members who turned out to be men.[31]

More importantly perhaps are two major dimensions that are left out of the perspective almost entirely in these datasets: i) the extensive involvement of the VOC in shipping and other activities in Asia; and ii) the work and life of all these other categories of workers that did not enter into the orbit of the VOC through recruitment in the Dutch Republic (and thus stayed out of the administrative lens of the *Scheepssoldijboeken*). For this first concern, it is interesting that the Scheepssoldijboeken provided important information that has not been included in the VOC Opvarenden database. The Scheepssoldijboeken were used to track the 'overseas' careers of VOC workers that would be paid out (upon return) in the Dutch Republic. This was done by extracting information from the *General Land and Sea Muster Rolls* (*Generale Land- en Zeemonsterollen*), which administered the location, rank and monthly salary of everyone employed by the VOC throughout Asia and Africa, and were copied and sent over every year from Batavia to the Chamber Cities in the Dutch Republic. The connection between the *Scheepssoldijboeken* and the *Generale Land- en Zeemonsterrollen* was made based on the name of the worker, the ship and year of the (most recent) outward bound voyage, and the rank during that voyage. It shows that amongst those who survived the initial hardship of the outward-bound voyage (with scurvy being responsible for large losses of life) and of the arrival in Batavia (with malaria being a threat from the early 18th century onwards) a significant portion remained in VOC service in often different parts of Asia for extensive periods and even retired there to become a 'freeburgher' (*vrijburger*). This non-intercontinental, but more 'overseas' and local colonial side of VOC employment remains largely invisible in the data (except for a mentioning of the 'reason of end of service').

Furthermore, it is crucial to note that only information for those who were recruited and paid out in the Dutch Republic was copied into the *Scheepssoldijboeken*. As the *Land- en Zeemonsterrollen* indicate, however, the workforce of the VOC in Asia did not only consist of Europeans recruited in the Dutch Republic, but also of workers recruited outside Europe; sometimes these were Europeans, but more often Asian, African and Eurasian sailors, soldiers and artisans. These were administered in local and general land and sea muster rolls, which have not yet been systematically analysed or processed into structured data, except for the *Generale Zeemonsterrollen* database (but this only contains metadata on the crew members, and mention of some key not individual entries) and a study of the muster rolls of

[31] There is a whole literature on this on Germans in the service of the VOC; VOC recruitment and employment patterns; women in the service of the VOC.

the VOC settlement in Bengal in the 18th century (but this did not result in a digital structured data collection).[32] The *Gansenhoef*, for example, occurs in the *GZM* database during its time in Asia, when it is engaged in intra-Asiatic voyages and not in intercontinental shipping (and is thus not registered in the *DAS*). For June 1709, we find recorded in the *GZM* database that the *Gansenhoef* is in Palembang under the command of captain Jan van Diemen van Rotterdam, with a crew of 68 sailors. The records note that 26 of them are 'inlandse' (inland/native) sailors, of which only the 'Moorse Zarang' (Moor Sarang) Calee and the Moor Sarangsmate Fakier Mameth are named.[33] In other instances, registrations of Asian crews on board VOC ships were sometimes entirely anonymous, but oftentimes with explicit lists of names and origins, and these were not yet entered into datasets.

To give an impression of the wider scope of workers that fall outside the spotlights: at the height of the power of the VOC mid-18th century, at any given moment some 13,000 European workers were estimated to be at sea, either on intercontinental or intra-Asiatic voyages, and another 20,000 European workers worked ashore in Asia and Africa, of which an unknown part was probably recruited 'in Asia'. Importantly, an approximate 12,000 Asian soldiers and 1,000 Asian sailors were employed throughout the VOC empire. These were not recorded in the Scheepssoldijboeken but should mostly be recorded in the *Land- en Zeemonsterrollen* (and thus do not occur in any available datasets). Thousands of shoreworkers, often hired on a daily basis locally, and some 7,000 enslaved workers were employed by the VOC directly, but not registered in either of these sources. In addition, thousands of corvée workers, drafted from local colonised populations, were used by the Company, but these workers also do not recur in the administrations of personnel.[34]

These problems can be, and are currently being, counterbalanced, but it should not be underestimated what the magnitude of the challenges is that is being imposed by the source material (the perspective of the Company), the historiography (the pervasive and lasting Eurocentric focus) and even by the possibilities to create new datasets (the complexities of representations of people in the archive). In the last decades, new datasets have been produced that reflect renewed interest in non-European actors, histories of slavery, and more. The dataset by Gerrit Knaap on shipping and trade on Java (1774–1777), for example, covered regional and local shipping to Javanese ports, and included the names of captains and owners of

[32] Lequin, H*et personeel van de Verenigde Oost-Indische Compagnie in Azie in de achttiende eeuw, meer in het bijzonder in de vestiging Bengalen*; Rossum, 'Database Generale Zeemonsterrollen', https://doi.org/10.17026/dans-zeu-be9b.

[33] See: https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/11722/file/NL-HaNA_1.04.02_11722_0244

[34] Van Rossum, 'Working for the Devil'.

non-European ships, including references to (normalised) ethnicities that the VOC administered. This leads to references such as 'Abdul Sale', being a 'Banjarees' on a 'pancalang'.[35] Similar sources on non-European shipping and the people involved are available for regions throughout the VOC empire, as work on Cochin shows.[36] In a similar fashion, the Sejarah Nusantara project has produced several datasets that were intended as indexes on the digitised material of the VOC archives in Jakarta, that feature a wider variety of non-European people, mostly rulers, merchants and VOC subjects.[37] Historical research into slavery, slave trade and judicial records provides another body of datasets that are either already available, can be recovered as 'grey' literature, or will be made available. The production of historical work and data has been most impressive for South Africa and specifically the Cape of Good Hope, with data on slave transactions, lists of convicts and slaves, etc.[38] Data on the criminal court record cases preserved for various legal courts of VOC settlements (esp. Batavia, Colombo, Cochin), created by the *Resilient Diversity* project (2017–2022), provide the names, status, the criminal charges, and sometimes additional information such as informal name (*bijnaam*), origin, residency and occupation.[39] The datasets on slave transactions and slave export permissions in Cochin, created by the *Between local debts and global markets* project, provides another body of data on non-Europeans, giving names, status, occupations and residency of local buyers and sellers, as well as old names, new names, status (often indications that hint at caste or community), sometimes age, and sometimes even the monetary valuation of the transaction.[40]

The development of such datasets provides one way forward. The counterbalancing of the heavily Eurocentric (and then even biased by the additional 'intercontinental' and 'maritime') perspectives in the production of data is necessary. And it seems hard to see how this can be done in any other way then producing alternative bodies of data that foreground especially hitherto underrepresented categories of people, such as women, enslaved, soldiers, local populations inside

[35] See Knaap, 'Shipping and Trade in Indonesia'.

[36] Arasaratnam, 'The Dutch East India Company and Its Coromandel Trade 1700–1740', pp. 325–346; Brouwer, 'Non-Western Shipping Movements in the Red Sea and Gulf of Aden during the 2nd and 3rd Decades of the 17th Century, According to the Records of the Dutch East India Company: (Part 2)'; and more recently, Odegard, 'Maritime India in Dutch Sources', pp. 340–348.

[37] See: https://sejarah-nusantara.anri.go.id/, especially the indexes on the Corpus Diplomaticum, the Asian Letters, the Supplements to the General Resolutions and the Marginalia of the Daily Registers.

[38] The data collections are too extensive to give a comprehensive overview here. Although it does seem that not all data is easily accessible, preserved in durable formats, or collected in a central location.

[39] See Canfijn et al., 'Early Modern Dutch Colonial Court Records'.

[40] See Van Rossum and Tosun, 'Acten van Transport Batavia (1664–1780)'; Van Rossum et al., 'Slave Transport Permissions – VOC Cochin, 1770–1795'.

and outside VOC colonies, as well as even people of non-European elites, such as merchants, translators, captains, foremen and rulers.

Of course, such data, and the creation of it, comes with its own difficulties. The court records dataset of *Resilient Diversity* figure people such as Sonna, Abdol, Jallal, Dolla, Mannak, Sambang, Tekaija, Juda, Gebrien, Labie, Souwer, Daging and Souasa before the Court of Justice of Batavia in 1732. They are all referred to as 'Vrije Maleijer' (Free Malay), and as male (deduced from the context); but apart from the fact that they were all trialled for 'Ongepermitteerd kruijssen en suspect swerven ter zee' (unpermitted sailing and suspicious roaming at sea), there is little that we know about them. In other instances, a bit more detail is provided, such as in the case of Caba Caba, the slave of the Balinese vaandrig (military officer) named Pinattij, who served as a convict labourer (kettingganger) in the artisan quarter of Batavia. The enslaved Caba Caba was before the court in Batavia in 1734, because he was accused of injuring the mandoor (overseer) Bappa Ramboe and murdering the soldier Matthijs Dirksz van den Hoek van Utrecht. It shows both the richness and potential, as well as the gaps, of datasets that try to reach beyond the ease of the conventional realm of data on European Company servants.

But it is not only the availability, and presence or absence of information, that is the problem. The observations that can be retrieved from the datasets, and in fact from the archive itself, are often also difficult to interpret and contextualise. In the Cochin slave trade database, for example, we encounter people like 'Oloenda' (according to the VOC records her 'previous' slave name) who was renamed in a transaction in 1791 as 'Klarinda'. The records state she was sold by Kanatoe Midien Markaat, a 'Moor' residing in Trikame, to Joseph Perera, a 'Toepas' residing in Naipin (Vaijpeen). Oloenda/Klarinda was mentioned to be 15 years old and to be a 'Poelitjij'. This was one of the many ways in which the VOC records seem to refer to people of the Pulaya caste. This, of course, leads to many questions on what is registered, who we see here, how to make sense of them, and how to capture this in appropriate data structures. Are, for example, the many variants of Poelia, Poelitjij, Poeletje and even Poeletjie actually referring to the same social category, or at these nuanced subdivisions that would be erased by a too simple normalisation? It is important to emphasise that this is of course not unique for data on non-European persons, groups and communities, and is also the case for Europeans. What is at stake here is not only issues of translation and the complexity of trying to read and use the colonial archive against the grain. To some extent there is additionally also the difficulty of trying to understand the complexities of local-colonial historical situations in Dutch periods that have later been overwritten by more dominant English colonial frameworks.

Alongside counterbalancing the Eurocentric data-production, these are challenges that will need to be worked on.

5. Technological bias

Technological bias in digital archival infrastructures refers to biases and limitations that can be encoded or reinforced by technologies used to create, access, and analyse historical archives. While AI and digital technologies provide new opportunities for historical research, they also pose risks of perpetuating or amplifying biases that exist in the original sources and introducing new biases.[41] These biases emerge not only from technical factors like algorithm design and training data, but also from the intersection of historical power structures, institutional practices, and modern computational methods.[42] When applied to historical archives, these biases can reinforce power dynamics, privilege certain narratives, and marginalise or exclude certain perspectives, such as those of historically marginalised groups. This challenge is especially critical as unexamined biases in institutions, metadata, and computational tools can transform and amplify historical exclusions.[43] Addressing technological bias is crucial for developing more inclusive and accurate historical research practices. This section discusses common technological biases in digital archival infrastructures and provides examples and solutions from the GLOBALISE project.

Ground truth bias

When deploying machine learning models to analyse historical archives – whether for automated transcription or content extraction – one must carefully consider the quality and composition of their training data, known as 'ground truth.' Ground truth bias occurs when this training data is itself biased or incomplete, leading to skewed outcomes in the model's predictions or performance.[44]

1) Selection bias

Selection bias arises when training data lacks representation of certain populations, resulting in skewed predictions.[45] Consider the VOC archives, where European

[41] Noble, 'Algorithms of Oppression: How Search Engines Reinforce Racism'; Thylstrup, *The Politics of Mass Digitization*; Hicks, *Programmed Inequality*.

[42] Luthra and Jeurgens, 'Humanising Digital Archival Practice. Access to Archives Guided by Social Justice'.

[43] Fyfe, 'An Archaeology of Victorian Newspapers', pp. 546–577; Gabriele, 'Transfiguring the Newspaper'.

[44] Crawford, *The Atlas of AI*; Søgaard, Plank, and Hovy, 'Selection Bias, Label Bias, and Bias in Ground Truth', pp. 11–13.

[45] Suresh and Guttag, 'A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle'.

perspectives dominate while other voices fade into the margins. Modern AI systems, trained on such skewed (historical) records, inherit the same blind spots. Buolamwini and Gebru (2018) demonstrate the real-world impact of selection bias: commercial gender classification systems, trained on imbalanced datasets, showed significantly lower accuracy for darker-skinned and female faces, effectively encoding racial and gender biases into technology.[46]

To address selection bias, the GLOBALISE project takes a systematic approach to creating representative training data for tasks such as entity recognition. The project samples across a broad temporal range (17th and 18th centuries) and diverse document types, from General Missives and daily registers to criminal proceedings and commodity lists. This sampling strategy captures varied perspectives across political events, shipping activities, and geographical regions. By thoroughly documenting these sampling decisions, GLOBALISE supports transparency in their decision making by allowing other researchers to critically examine the training data's composition.

2) Metadata and indexing biases

Metadata and indexing biases emerge from critical choices made during digitisation: what to digitise, how to categorise it, and which metadata to assign.[47] These decisions fundamentally shape how historical records can be accessed and discovered in digital archives. The GLOBALISE corpus illustrates this challenge through its treatment of named and unnamed individuals. Historical documents often name those in positions of power while leaving others anonymous, reflecting and reinforcing colonial power dynamics. GLOBALISE's infrastructure challenges these traditional power structures by recognising both named and unnamed individuals (see Figure 7.2, where all person mentions are highlighted in green).[48] This approach moves beyond standard named entity recognition frameworks like CoNLL or ACE, which typically limit themselves to simple categories of Person, Organisation, and Location.[49]

Colonial archives contain rich contextual information about historical actors – their roles, gender, legal status, and social positions. Instead of reproducing colonial-era classifications, GLOBALISE's metadata framework employs intersectional categories that highlight individual agency, skills, and contributions. By capturing these nuanced entity attributes, the project recontextualises historical data to foreground human stories rather than merely institutional records. This approach

[46] Buolamwini and Gebru, 'Gender Shades'.

[47] Søgaard, Plank, and Hovy, 'Selection Bias, Label Bias, and Bias in Ground Truth'.

[48] Luthra et al., 'Unsilencing Colonial Archives via Automated Entity Recognition'.

[49] Doddington, 'The Automatic Content Extraction (ACE) Program: Tasks, Data, and Evaluation'; Sang and De Meulder, 'Introduction to the CoNLL-2003 Shared Task'.

Met dit schip gaen derwaerts Gillis Venant met sijn huijsvrou, item eenen Alberto Gomes met sijn huisvrouw Isabelle de Mato ende haere twee zoonen Johan de Mato ende Thome de Camine haer moeder Lucia Fernandes, mitsgaders ses stucks slaven ende een slavinne, altesamen over haere delicten ende quade handelingen gerelequeert naer d' Eijlanden van Banda Conform de neffens gaande respective sententien soo van Venant als de Alberto Gomes, met sijne familie U.E. gelieve de selve aldaer woonplaetse t'assigneren, ende nae dat sij hen met goede comportementen waerdich maken U.E. faveur laten genieten.

On this ship are sent Gillis Venant and his housewife, together with an Alberto Gomes and his housewife Isabelle de Mato and her two sons Johan de Mato and Thome de Camine, her mother Lucia Fernandes together with six units of male slaves and a slave woman, all of them have been exiled to the Banda Islands for their offences and wrongful actions. We request your honours that they are assigned accommodation there following the accompanying sentence related to Venant and Albert Gomes and his family, and that you show them your benevolence when they have made themselves worthy of that with their good conduct.

Figure 7.2: Excerpt of persons annotation

enables a more complete understanding of historical narratives, particularly those of marginalised individuals and communities whose voices are often absent from or present in traces in traditional archival structures.

3) Annotator Bias

Just as colonial archives reflect European frameworks of categorisation that distorted local identities and social structures, modern digital archives face a similar challenge through annotator bias. This bias emerges when human annotators, tasked with labelling data for natural language processing tasks, bring their own cultural assumptions and perspectives to the classification process. An annotator's subjective judgments – shaped by their cultural background, language knowledge, and personal beliefs – can significantly influence how historical information is categorised and interpreted, affecting model performance and perpetuating misrepresentations.

GLOBALISE addresses this challenge through several strategies. The project develops detailed annotation guidelines enriched with instructive examples to promote consistent interpretation across different annotators.[50] It also tracks annotator information by including annotator IDs as features in the training data and carefully separates training set annotators from test set annotators.[51] This systematic approach helps identify and mitigate individual biases while enabling better evaluation of model performance across different annotator perspectives.

To ensure cultural and historical accuracy, GLOBALISE collaborates with area specialists in early modern Indian Ocean histories, gender studies experts, and data

[50] Parmar et al., 'Don't Blame the Annotator'.

[51] Geva, Goldberg, and Berant, 'Are We Modeling the Task or the Annotator?'.

ethics specialists. This interdisciplinary approach helps develop more nuanced and historically accurate categorisation systems that better reflect local social structures and identities.

Language models and algorithmic bias

GLOBALISE will make use of large language models (LLMs) in its infrastructure to enable various natural language processing tasks. LLMs are computational models that learn patterns and structures in linguistic data, allowing them to generate coherent and contextually relevant text, and be deployed for text and document and text classification tasks.[52] However, LLMs require careful consideration since they can inherit, perpetuate, and even amplify biases through their training data and algorithmic design.[53]

1) LMs and low resource languages

The use of LLMs on historic Dutch – as found in the GLOBALISE corpus – pose challenges due to the differences between 17th and 18th century and modern Dutch. Contemporary language models like spaCy and BERTje, trained on 20th-century Dutch news and 19th century books, struggle with historical variations in vocabulary, grammar, and spelling.[54] However, multilingual models such as mBERT and XLM-R offer promising solutions. Trained on a diverse range of languages, these models exhibit cross-lingual transferability and can capture the language variations found in historical texts. By utilising multilingual models, like mBERT or XLM-R, GLOBALISE can tap into their capacity to learn generic representations across different languages and scripts, making them particularly suitable for analysing historical texts with non-contemporary language and unstandardised spelling.[55] These models, combined with careful fine-tuning on historical Dutch data, offer more robust solutions for analysing non-contemporary texts.[56]

[52] Jiang et al., 'How Can We Know What Language Models Know?'; Vaswani et al., 'Attention Is All You Need'; Devlin et al., 'BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding'.

[53] Mehrabi et al., 'A Survey on Bias and Fairness in Machine Learning', 115:1–115:35; Baeza-Yates, 'Bias on the Web'.

[54] De Vries et al., 'BERTje: A Dutch BERT Model'.

[55] Conneau et al., 'Unsupervised Cross-Lingual Representation Learning at Scale'; Pires, Schlinger, and Garrette, 'How Multilingual is Multilingual BERT?'.

[56] Arnoult, Petram, and Vossen, 'Batavia Asked for Advice', pp. 21–30.; Vaswani et al., 'Attention is All You Need'.

2) Language model biases

It is crucial for GLOBALISE to consider and address biases inherited from LMs such as biased training data, biased training objectives or model limitations.[57] One of the biggest issues is biases inherited from the often unrepresentative, vast, and opaque or 'unfathomable' training data of LLMs.[58] While some biases may originate from the ground truth, the LMs own biases can magnify or reinforce them, such as gender, racial, or cultural biases in the analysis and interpretation of historical documents.

Conducting thorough evaluations of model outputs and analysing the impact of different components (e.g., data sources, training objectives, fine-tuning procedures) can provide insights into the sources of bias. Post-processing techniques, such as debiasing methods or fairness-aware algorithms, can help mitigate biases introduced by the model while being aware of the biases present in the ground truth.[59] Furthermore, transparent documentation is vital in navigating the biases of LLMs inherited from their training data and the biases within the ground truth.[60] This combination of technical intervention and transparent documentation enables informed engagement with historical data and its computational processing in GLOBALISE.[61]

Environmental and labour inequities in language models

While not a bias in the traditional sense, the environmental and labour costs of LLMs reflect broader power imbalances that parallel the colonial histories GLOBALISE studies. Just as colonial trading companies like the VOC extracted resources and labour from across the Global South, today's LLMs rely on unsustainable computational resources and often exploitative labour practices. Training a single BERT model consumes energy equivalent to a trans-American flight, while data labelling frequently depends on underpaid workers in poor conditions.[62]

[57] Navigli, Conia, and Ross, 'Biases in Large Language Models'.

[58] Bender et al., 'On the Dangers of Stochastic Parrots', pp. 610–623.

[59] Bolukbasi et al., 'Man is to Computer Programmer as Woman is to Homemaker?'; Bordia and Bowman, 'Identifying and Reducing Gender Bias in Word-Level Language Models'; Jin, 'On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning'.

[60] Gebru et al., 'Datasheets for Datasets', pp. 86–92; Bender and Friedman, 'Data Statements for Natural Language Processing', pp. 587–604.

[61] For examples of dataset documentation have a look at the data-envelopes and accompanying documentation for GLOBALISE's datasets: https://datasets.iisg.amsterdam/dataverse/globalise. For instance have a look at the dataset and documentation on Ethnicities, Religious Groups and Castes in the Archives of the Dutch East India Company (1602–1799).

[62] Strubell, Ganesh, and McCallum, 'Energy and Policy Considerations for Deep Learning in NLP'; Mohamed, Png, and Isaac, 'Decolonial AI', pp. 659–684; Gray and Suri, *Ghost Work*; Perrigo, 'OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic'.

While GLOBALISE has not yet identified open-source models developed under fair labour conditions and with low environmental footprint, we believe it is crucial to acknowledge these inequities, particularly as we use these tools to study historical colonial exploitation. GLOBALISE addresses these challenges where possible by prioritising energy-efficient approaches and ensuring ethical labour standards in its own data annotation processes. We maintain transparency about these limitations and continue to advocate for more equitable practices in AI development.

Problem of less frequent types

Machine learning models often struggle with accurately identifying low-represented types or underrepresented groups in data. Due to imbalances in training data, models perform poorly when encountering less common classes or identities, further amplifying historical patterns of marginalisation.[63] For example, when certain social groups or historical actors appear infrequently in colonial records, the models may fail to recognise or properly categorise them in future analyses. GLOBALISE approaches this challenge through targeted data collection strategies and bias-mitigation techniques. This includes enhancing representation in training datasets where possible and developing specialised methods for handling low-frequency cases.[64] These efforts aim to ensure more equitable treatment of historically marginalised groups in our digital analysis.

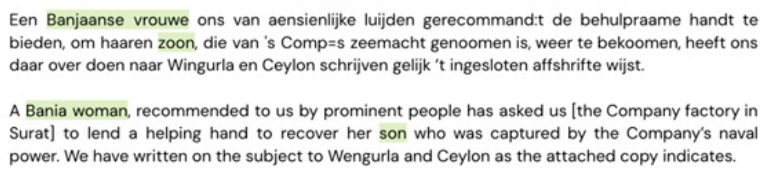
Search term bias

Search bias is a prevalent issue in digital archival infrastructures where search algorithms can exhibit biases in determining the relevance and prominence of information. This bias can arise if the algorithm favours Eurocentric keywords or concepts, thereby making it challenging to discover non-European histories or perspectives. Furthermore, the reliance on anachronistic terminology and colonial sources can reinforce colonial naming conventions and limit the searchability for those unfamiliar with the colonial terms used. To address this bias, one approach is to provide alternative labels for concepts. For example, in the case of 'Myanmar,' attaching alternative labels such as 'Burma' and 'Arakan' can enhance searchability and accommodate different historical perspectives.

[63] Friedman and Nissenbaum, 'Bias in Computer Systems', pp. 330–347.

[64] Chouldechova and Roth, 'The Frontiers of Fairness in Machine Learning'; Buolamwini and Gebru, 'Gender Shades'.

However, this strategy of providing alternative labels for similar concepts also comes with its own dangers of reductionism. For instance, consider the following example in Figure 7.3[65]:


Een Banjaanse vrouwe ons van aensienlijke luijden gerecommand:t de behulpraame handt te bieden, om haaren zoon, die van 's Comp=s zeemacht genoomen is, weer te bekoomen, heeft ons daar over doen naar Wingurla en Ceylon schrijven gelijk 't ingesloten affshrifte wijst.

A Bania woman, recommended to us by prominent people has asked us [the Company factory in Surat] to lend a helping hand to recover her son who was captured by the Company's naval power. We have written on the subject to Wengurla and Ceylon as the attached copy indicates.

Figure 7.3: Excerpt of persons annotation

In this example the first person labelled is a woman belonging to the caste group Bania.[66] Some of the spelling variations that the caste name Bania features in the archive are as follows: 'Benjaan', 'Banjaan' and 'Benjaen'. When historians refer to this caste in the early modern period, they are accustomed to using the standard spelling from the present day, Bania to refer to the caste from the early modern period. This is problematic because it can equate early modern castes with their modern-day namesakes.[67] The other alternative is to use a name and spelling variant of the term as found in the VOC archives. This is also questionable because it renders the caste name less recognisable to the modern reader, and makes us reliant on names as they feature in the Dutch archive. In the case of GLOBALISE, we aim to provide the various spelling variations of the term Bania, accompanied by a disclaimer about the possible conflation with its modern-day namesake. This approach acknowledges the complexities of historical naming conventions while informing users about potential limitations and the need for critical interpretation.

[65] Originele missive dato primo Junij [1658] haare Ed. uyt Suratte per terra toegesonden, N.A. 1.04.02, Inv. No. 1226, f. 745v. https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/1226/file/NL-HaNA_1.04.02_1226_0518. This information was found in the archive using the GLOBALISE project, 'VOC Transcriptions – GLOBALISE.'

[66] Haynes, *Rhetoric and Ritual in Colonial India.*

[67] Dale, *Indian Merchants and Eurasian Trade, 1600–1750*.

Strategies to mitigate technological bias

In order to address technological bias in digital archival infrastructures, it is imperative to adopt a critical and reflective approach. Drawing from the sections above, here are some strategies to consider:

1. **Diversify training data**: Ensure that the training data used for developing language models and other technologies is diverse and representative of the historical context being studied. Incorporate records, narratives, and perspectives from non-European actors to mitigate biases.
2. **Evaluate and validate results**: Critically assess the outputs generated by technology and compare them with existing historiographical knowledge. Scrutinise potential biases and interpret results in light of the historical and cultural context.
3. **Collaborative approaches**: Involve historians, domain experts, and stakeholders from diverse backgrounds in the development and implementation of digital archival infrastructures. Encourage interdisciplinary collaborations to provide a broader range of perspectives and insights.
4. **Transparent and accountable processes**: Make the processes and algorithms used in digital archival infrastructures transparent and subject to scrutiny. Document and disclose the limitations, biases, and uncertainties associated with the technology and the archival data.
5. **Ethical considerations**: Develop and adhere to ethical guidelines when using technology in historical research. Consider the potential impact on marginalised communities, respect privacy concerns, and ensure that the research serves the interests of inclusivity and social justice.

6. Conclusions and future work

This chapter is in many ways an act of introspection. In creating a research infrastructure that will re-envision the manner in which the VOC archives are accessed and researched, the GLOBALISE project will have deep implications on how this past is read, comprehended and written about. This chapter articulates our concerns with shouldering such a responsibility particularly in the context of recovering marginal historical actors from the archive. It is therefore vital that we adopt a four-step process to combat and mitigate bias in the research infrastructure that we create. The first step is *acknowledging the presence of biases* in the archive, existing datasets and digital infrastructures and recognising the fact that they cannot be entirely eliminated. This chapter makes it apparent that we can devise solutions to consciously mitigate these biases in such a way that the archive,

inventories, datasets and digital infrastructures visibilise marginalised historical actors. The second step is *surveying and inventorying the biases* that either plague or could threaten to affect our research infrastructure. It enables us to keenly perceive issues that we need to be aware of but also consider preliminary solutions that the project can consciously adopt at every stage. As many of the solutions proffered by the chapter are suggestions that the project is presently working on or hopes to execute, the third and future step of the project will be to actively *document* the decisions we make to handle bias in our approach to the archives and the creation of datasets and digital infrastructure *and make transparent* the implications that our actions have on our infrastructure and resultant output. The fourth step is to ensure that this *documentation* is not only presented to users of our interface as auxiliary reports, but that it is in fact critically incorporated into the very design of the infrastructure that informs users of its context, potential biases, and subjectivities.[68]

Importantly, this four-step process must also include critical reflection on our own positionality as researchers. Our identities, institutional contexts, and disciplinary training shape every aspect of how we approach these archives – from why we do the work, to whom we choose to collaborate with, to what we choose to digitise, to how we interpret and present historical actors. Acknowledging this positionality is not merely an academic exercise but essential for developing responsible research infrastructures.

In this period in time when numerous archival digitisation projects are conceived and created, we hope that our chapter on biases in the GLOBALISE project serves as a valuable resource and toolkit for addressing bias in historical research. This chapter should serve as a catalyst to such projects to prioritise and reflect on the issue of bias from the early stages of their endeavours, providing them with guidance on navigating similar challenges. This chapter also highlights the fact that while the issue of bias has been extensively interrogated within the discipline of history and separately within computer science, any rigorous and comprehensive effort to combat bias requires us to transcend these narrow disciplinary boundaries, join forces and thereby think of innovative strategies to handle the problem.[69]

[68] Verhoeven, 'As Luck Would Have', pp. 7–28; Whitelaw, 'Generous Interfaces for Digital Cultural Collections'.

[69] The insights and questions from this work and the GLOBALISE project more generally, and our stubborn, necessary commitment to prising open these archives with conscience intact gave rise to the Combatting Bias project, born from the recognition that the careful work of making visible the invisible in historical records offers vital lessons for contemporary data practices. Combatting Bias is now creating systematic guidelines for creating bias-aware data throughout the research lifecycle.

Acknowledgements

This publication is part of the project GLOBALISE (with project number 175.2019.003) of the research programme Research Infrastructure which is financed by the Dutch Research Council (NWO).