An Invitation to a Trans-Caribbean Data Project: A Pipeline for Historical Caribbean Population Databases

Björn Quanjer
Thunnis van Oort
Matthias Rosenbaum-Feldbrügge
Cornelis W. (Coen) van Galen
Lisa Hoek
Erik Tjong Kim Sang

The Historical Database of Suriname and the Caribbean (HDSC) is a network of historians and archivists in the Dutch Caribbean, Suriname and the Netherlands who work together with the aim to create open access databases of the population of Suriname and the (Dutch) Caribbean between 1830 and 1950, both for public and scholarly use. The so-called slave registers (1830s-1863) and the civil registry of the free population (1830s-1950) of the former Dutch colonies in the Caribbean comprise the backbone of the project, because they allow both scholars and the public to reconstruct life courses. In this article we discuss the project and invite the reader to collaborate on a trans-Caribbean project. It outlines the possibilities of an interoperable pipeline for historical Caribbean population databases, emphasizing the data construction and linking process. Since 2017, the HDSC developed multiple digital workflows that can function as a blueprint for similar efforts to reconstruct historical populations. Among the project's core principles is a commitment to make all datasets and workflows available open access to any and all researchers. Sharing the developed workflows and discussing the challenges and pitfalls that affect a successful outcome provides a building block towards this long-term goal. Although the project's current focus is on the effects of colonialism, slavery, and indentured labour in the Dutch empire within the Caribbean, the HDSC hopes to work with other researchers to connect all available data to non-Dutch territories and to create an interconnected database of people and migration patterns for the Caribbean region as a whole.

archipelagos • 2025-03-01 • 10.7916/ARCHIPELAGOS-WKX1-PD50 • CC-BY 4.0 archipelagos journal • ISSN: 2473-2206

An invitation to a trans-Caribbean data project: A pipeline for historical Caribbean population databases

Introduction

In 2023, UNESCO recognized the *Documentary heritage of the enslaved people of the Dutch Caribbean and their descendants (1816-1969)* as part of the Memory of the World collection. The collection includes two databases of the slave registers of Suriname and Curaçao, created by the Historical Database of Suriname and the Caribbean (HDSC). For the HDSC the UNESCO designation was an official recognition that the creation of datasets is a crucial element to making the history of the Caribbean more accessible. Moreover, creating fine-grained databases which contain detailed information makes possible the deconstruction of colonial archives and reconstruction of individual lives by extracting personal information from the archival records and linking it to other archival sources in one large data infrastructure. In this way, it is possible to reconstruct people's *life courses* through these sources, not only for specific individuals but for virtually all persons living in a given territory. This so-called life course approach makes it possible to better understand groups as well as individuals.[1] After all, searching for links among multiple archival records can lead to unexpected new knowledge about the behavior and agency of groups and individuals in the Caribbean.[2]

The Historical Database of Suriname and the Caribbean is a network of historians and archivists in the Dutch Caribbean, Suriname and the Netherlands, affiliated with the National Archives of Suriname and Curaçao, the Anton de Kom University of Suriname, and Radboud University in the Netherlands. The overarching aim of the HDSC network is to create databases of the population of Suriname and the (Dutch) Caribbean, both for public and scholarly use. Scholars in the HDSC work together to (1) describe, restore and digitize archival records; (2) transcribe the digitized records; (3) create linked datasets and publish them online; and (4) train citizen scholars, students, and researchers to work with the digitized materials and databases. The National Archives execute the first step; the second one happens online, mainly through the work of citizen scientists in Suriname, Curaçao, and the Netherlands; and linking specialists at Radboud University in the Netherlands in collaboration with the National Archives execute the third step. It is the hope that the fourth step will encourage more people to build their own datasets that might then be linked to the open access HDSC data infrastructure to ultimately create a trans-Caribbean data-platform. This would potentially make it possible to trace the lives of individuals who migrated throughout the Caribbean in the 19th and early 20th centuries.

[1] J. Kok, "Principles and prospects of the life course paradigm," *Annales de demographie historique* 113, no. 1 (2007): 203-230, https://doi.org/10.3917/adh.113.0203.

[2] C.W. van Galen et al., "Endless Digging and Endless Picking. Sex Ratios and Gendered Labour in Surinamese Plantations, 1830–1863," *Historical Life Course Studies* 10 (2021): 46-52, https://doi.org/10.51964/hlcs9566; C.W. van Galen and B. Quanjer, "The Wolf, the island and the sea: Truancy and escaping slavery in Curaçao (1837-1863)," *The History of the Family* 29, no. 2 (2024): 262-279, https://doi.org/10.1080/1081602X.2024.2340542.

The present article invites the reader to engage with a trans-Caribbean project. It has been written with the fourth step in mind and outlines the possibilities of an interoperable pipeline for historical Caribbean population databases. Emphasizing the data construction and linking process, scholars at Radboud University wrote the article with support from two external data specialists who helped develop project workflows.

The work of the HDSC is based on the life course approach. According to Kok, this "approach is essentially a heuristic device to study the interaction between individual lives and social change. It is a way of conceptualizing lives within the contexts of families, society and historical time. The life course can be defined as the sequence of positions of a particular person in the course of time. A position can be either marital status, parenthood, employment, residence in a particular location etcetera."[3] Changes in these positions are called life events. These changes are often registered in archival records and are used to reconstruct the life courses of individuals.

The emphasis of the HDSC is reconstructing full life courses of all enslaved and free individuals who lived in Suriname and the former Netherlands Antillean islands of Aruba, Bonaire, Curaçao, Saba, St. Eustatius, and Dutch St. Martin, between 1830 and 1950. Full life courses mean the reconstruction of the information of people's life events, including birth, death, migration and change of ownership (for enslaved individuals), based on a range of archival sources. Because all types of information are copied from the archival source, the open access data infrastructure allows the public to carry out family history research. It can also be used for educational and cultural projects, and it responds to the high demand for scholarly research concerning the people living in colonial (post-) slavery societies. The data infrastructure can be adapted for all types of qualitative and quantitative research questions, from individual biographies to historical demographic cohort research.

Since 2017, the HDSC has developed multiple digital workflows that can function as a blueprint for similar efforts to reconstruct historical populations. Among the project's core principles is a commitment to make all datasets and workflows currently available open access to researchers and the general public. Sharing the developed workflows provides a first building block towards this long-term goal. The so-called slave registers (1830s-1863) and the civil registry of the free population (1830s-1950) comprise the backbone of the project, because they allow both scholars and the public to reconstruct life courses. They can be complemented by other archival sources. The temporal scope starting from 1830 was chosen given that slave registers and civil registry were introduced in the Dutch colonies around that time. The project concludes around 1950, as it works exclusively with archival sources that are publicly accessible under Surinamese and Curaçaoan law to avoid privacy issues related to the publication of information about living individuals.

This article discusses these workflows, their challenges and pitfalls, and affects a successful outcome. Although the project's current focus is on the effects of colonialism, slavery, and indentured labor in the Dutch empire within the Caribbean, the HDSC hopes to work with other researchers to connect all available data to non-Dutch territories and to create an interconnected database of people and migration patterns for the Caribbean region as a whole. The complete HDSC workflow is illustrated in Figure 1 (see below).

[3] Kok, "Life course paradigm," 204.

Why use a data infrastructure?

The use of databases for Caribbean research is not new but like in the United States, it has fallen out of favor since the 1990s, after a boost of quantitative history in the 1970s and 1980s.[4] This earlier wave of quantitative history resulted in various quantitative publications on slavery in the Caribbean.[5] Researchers collected statistics often aggregated on a national level, Higman's *Slave Populations of the British Caribbean* being a striking example of such a project.[6] Similarly, they collected sample statistics for Suriname and the Dutch Antilles.[7] Such studies were demographic in nature, and tended to focus on economic factors with limited attention to the lived experiences of the enslaved people. A notorious example is Fogel and Engerman's *Time on the Cross*, whose revision of the American slave economy ignited a major historical controversy.[8] The cultural turn in historical research during the 1990s steered the field away from quantitative data and resulted in much more narrative based approaches.[9]

The economic focus of this earlier quantitative phase in historiography has cast a shadow on the historical research that uses these large databases. This background raises questions regarding the HDSC's data-driven approach and its possibilities for quantitative research. To the distinction of the quantitative historiographies of the past, however, database projects like the HDSC support a variety of methodological approaches and, crucially, can also be used to deconstruct the colonial sources from which the database has emerged. Breaking down the original source into elements which can be connected to and cross-referenced with elements from numerous other sources makes it possible to reconstruct and develop richer understanding of the individual and group experiences of the people mentioned in the sources.

This life course approach enables research on an individual and group level.[10] For example, the HDSC datasets offer possibilities for research on mobility, childbirth and the effects of mother and child separation. Often, the experiences of individuals are only visible when multiple sources are combined. For example, a recent study on escapes from slavery in Curaçao, reconstructed the escape of a group of enslaved and freed

---

[4] S. Ruggles, "The revival of quantification: Reflections on old new histories," *Social Science History* 45, no. 1 (2021): 1-25, https://doi.org/doi:10.1017/ssh.2020.44.

[5] H. S. Klein and S. L. Engerman, "Fertility differentials between slaves in the United States and the British West Indies: a note on lactation practices and their possible implications," *The William and Mary Quarterly: A Magazine of Early American History* 35, no. 2 (1978): 357-374, https://doi.org/10.2307/1921839; J. P. Koplan, "Slave Mortality in Nineteenth-Century Grenada," *Social Science History* 7, no. 3 (1983): 311-320, https://doi.org/10.2307/1171190; H. C. Johansen, "Slave demography of the Danish West Indian islands," *Scandinavian Economic History Review* 29, no. 1 (1981): 1-20, https://doi.org/10.1080/03585522.1981.10407935.

[6] B. W. Higman, *Slave Populations of the British Caribbean, 1807-1834* (Johns Hopkins University Press, 1984).

[7] H. E. Lamur, "Demography of Surinam plantation slaves in the last decade before emancipation: The Case of Catharina Sophia," in *Annals of the New York Academy of Sciences, Volume 192, Comparative Perspectives on Slavery in New World Plantation Societies*, ed. V. Rubin and A. Tuden (The New York Academy of Sciences, 1977): 161-173; W. E. Renkema, *Het Curaçaose plantagebedrijf in de negentiende eeuw* (Walburg Pers, 1981); H. E. Lamur, "Demographic performance of two slave populations of the Dutch speaking Caribbean," *Boletin de estudios latinoamericanos y del Caribe* 30 (1981): 87-102, https://www.jstor.org/stable/25675094; A. van Stipriaan, *Surinaams contrast: roofbouw en overleven in een Caraïbische plantagekolonie, 1750-1863* (Brill, 1993).

[8] R. W. Fogel and S. L. Engerman, *Time on the cross: The economics of American negro slavery* (Boston: Little Brow, 1974). For critique: H. G. Gutman, *Slavery and the Numbers Game: A Critique of Time on the Cross* (University of Illinois Press, 1975); L. Stone, "The revival of narrative: reflections on a new old history," *Past and Present* 85, no. 1 (1979): 3-24, https://doi.org/10.1093/past/85.1.3.

[9] Ruggles, "Revival of quantification," 13-14.

[10] Kok, "Life course paradigm," 204.

persons from Curaçao to the city of Santo Domingo, at that time under Haitian occupation (1822-1844), in 1841. Combining sources shows that a girl, Flora, was essential for the well-organized escape. She was hardly mentioned in the archives, but a reconstruction of Flora's life course showed that she was the only one with connections to all the participants in the escape.[11] In this way, the HDSC datasets help enrich the narratives that have formed over the past decades and how they come to create new narratives.

Furthermore, the HDSC project also has a public goal. The population records contain names and life events that are relevant, not only for scholarly research, but also, and above all, for the people of the Dutch Caribbean islands and Suriname as well as their diasporic communities. By making this information available and searchable online for personal, cultural and educational purposes, the project enables those who have their family history rooted in the Caribbean to connect to their ancestors. For this reason, participating archives in the HDSC project publish the public version of the datasets first, before a research-oriented version is developed. The societal effect of this approach is evidenced, for example, by the fact that the publication of the Surinamese slave registers in 2018-2019 resulted in the writing of multiple family histories using this data set.[12]

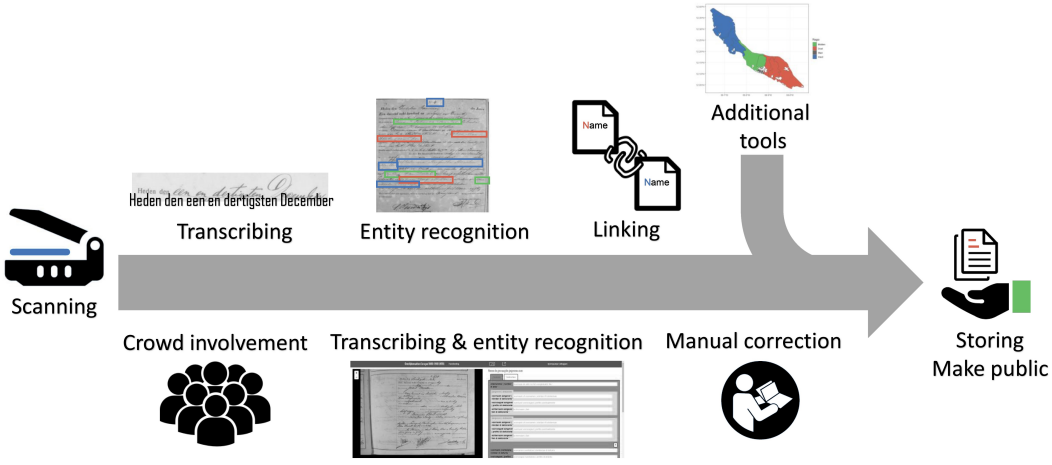


Figure 1 The HDSC workflow

Collaboration, scanning and storing

A project like the Historical Database of Suriname and the Caribbean only works if the partners: archives, researches, users, and funders, organize lasting cooperation. The HDSC started in 2017 as the combined crowdfunding and crowdsourcing campaign *Make*

[11] Van Galen and Quanjer, "The Wolf, the island and the sea," 266-267.
[12] An example is: M. Karg, *Stil Verleden: de ontdekking van mijn slavernijgeschiedenis* (Walburg Pers, 2023).

the Suriname Slave Registers Public.[13] After the success of this campaign, the different partners formalized the collaboration with the establishment of the Historical Database of Suriname Foundation in 2018, in which the National Archives of Suriname worked with the Anton de Kom University of Suriname and Radboud University in the Netherlands, along with specialists in database construction. The Foundation later expanded to include the National Archives of Curaçao. The board of the Foundation ultimately decides on the vision and planning of the HDSC project.

In addition to the official members, the HDSC has worked with a number of informal partners, such as citizen scientists, individual researchers, and civil society organizations. Some are temporary partners, others support the project long term, as is the case with the National Archives of the Netherlands and the Foundation for Surinamese Genealogy (SSG). A particularly important group among the project's partners are the citizen scientists, the participants who work voluntarily for the project transcribing the archival sources. Over the years, some fifteen hundred persons in Suriname, Curaçao, the Netherlands and elsewhere participated in the HDSC project. Thanks to the contributions of these citizen scientists, it is possible to process large amounts of data at relatively low costs.

In addition to their crucial support to the online transcriptions (more on this in the next section), citizen scientists are essential to the project for two key reasons. The first is that they are a vital part of the HDSC's network within the communities it aims to serve. They act as ambassadors and supporters who significantly amplify the project's visibility. For example, they share information on their social media platforms, initiate their own side projects, and discuss the project within their networks. Citizen scientists represent a highly diverse group, including individuals from Suriname, the Dutch Caribbean, diaspora communities, and the Netherlands, further legitimizing the HDSC project as an international endeavor. The second reason is the ongoing interaction between the project and the citizen scientists. The project organizes online transcription training sessions and live meetings in various countries to disseminate knowledge. This engagement fosters a stronger connection between citizen scientists, archival materials, and the scientific approach. However, this is not a one-way exchange. Many participants are themselves historians, educators, or genealogists with extensive knowledge of the Surinamese and Caribbean societies and their history. Their expertise and insights enrich the project.[14] They actively participate in the project and share their own knowledge as well as the ownership of the datasets.

Working with citizen scientists does involve certain risks. Previous scholarship particularly highlights challenges related to recruiting citizen scientists and, more importantly, keeping them engaged in long-term projects.[15] A critical starting point is recognizing that there is no such thing as *the* citizen scientist. Each participant has their own interests, motivations, and needs for joining the project.[16] By being sensitive to these varying

[13] C.W. van Galen, "Creating an audience: Experiences from the Surinamese slave registers crowdsourcing project," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52, no. 3 (2019): 178-194, https://doi.org/10.1080/01615440.2019.1590268.

[14] Van Galen, "Creating an audience," 186.

[15] S. West and R. Pateman, "Recruiting and retaining participants in citizen science: What can be learned from the volunteering literature?," *Citizen Science: Theory and Practice* 1, no. 2 (2016): 15, 1-10, https://doi.org/10.5334/cstp.8; A. Crall et al., "Volunteer recruitment and retention in online citizen science projects using marketing strategies: Lessons from Season Spotter," *Journal of Science Communication* 16, no. 1 (2017): A01, https://doi.org/10.22323/2.16010201.

[16] O. Nov et al., "Technology-Mediated Citizen Science Participation: A Motivational Model," *Proceedings of the International AAAI Conference on Weblogs and Social Media* 5, no. 1 (2011): 249-256, https://doi.org/10.1609/icwsm.v5i1.14113;

interests and needs, it is possible to maintain their engagement. In this project, regular reflection is conducted on the behavior and motivations of participating citizen scientists.[17] These reflections highlight the differences among participants but also reveal some commonalities. Almost all citizen scientists value recognition for their contribution to the project. For example, this can be done by promptly responding to their questions and sharing weekly updates about the project. Additionally, a significant portion of citizen scientists is motivated by their desire to use the data for their own (family) research. Therefore, it is advisable for citizen science projects to make it clear in advance how and when the produced databases will be made publicly accessible online.

As with many long-lasting projects, funding for the transcription project and the construction of datasets, ongoing infrastructural needs, inventorying and restoring archival material, for posting and hosting datasets on the archives' websites, and for publicity and educational projects, is an ongoing concern for the HDSC. In 2017, a crowdfunding campaign funded the first phase of the project since, at that time, funding organizations were not willing to fund a Surinamese-Dutch collaboration. When this proved successful, the HDSC received additional funding, enabling the project to be scaled up and split into sub-projects: an inventory, restoration, and digitization of records at the National Archives of Suriname (funded through the Gerda Henkel Stiftung); a combined citizen science and HTR project; and a project to link transcribed data (the latter two are funded through the Dutch PDI-SSH digital infrastructure platform). In addition, several smaller projects for education, communication, and the development of new technical applications have also materialized. The variety and number of these ventures enables the archives to raise their profile and strengthen research capacity among the partners, keeping the project relevant for all participants.

It is clear that securing financial resources for the project is significantly easier in Europe than in the Caribbean. However, reliance on European funding often comes with the drawback that donors expect the grant to be managed by a European organization. Combined with the technical expertise predominantly available in the Netherlands, this could easily result in the project's focal point shifting outside the Caribbean. This issue is difficult to address, not only due to economic disparities, but also because the populations of Suriname and Curaçao are much smaller than that of the Netherlands making it more difficult to find enough specialists within the Caribbean.

This project seeks to mitigate this imbalance in several ways. As noted, decision-making authority over the HDSC project resides with the board of the foundation. To ensure adequate Caribbean representation, the majority of board members are Surinamese and Curaçaoan historians and archivists. Working with board members from different nationalities has the additional benefit of making it easier to find funding, either by applying for support from international organizations which have specific programs for the Caribbean or from organizations which fund projects in the countries the HDSC board members come from.

Furthermore, one of the aims of the foundation is to transfer expertise on database construction and data linking. This includes training Caribbean students and archivists

---

D. Fraisl et al., "Citizen science in environmental and ecological sciences," *Nature Reviews Methods Primers* 2, 64 (2022), https://doi.org/10.1038/s43586-022-00144-4.

[17] Van Galen, "Creating an audience; M. Prats López et al.,"Understanding Patterns of Engagement in the Citizen Humanities: The Civil Records of Suriname," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* (2024): 1-16, https://doi.org/10.1080/01615440.2024.2414925.

and educating Surinamese doctoral candidates to work with the datasets. The aim is to end the technical support role of the Dutch Radboud University by 2029 and fully transfer the project to partners in Suriname and the Dutch Caribbean. The data the HDSC works with is the property of the Surinamese and Dutch Caribbean societies, and it is for archivists and researchers in these societies to decide how to expand the project with new types of archival sources or how to connect with similar projects elsewhere in the Caribbean.

Working together with multiple project partners and several hundreds of citizen scientists across different time zones and with different aims and needs is no small feat. The success of the collaboration requires that all project partners consider the colonial history of the Dutch Caribbean to be a shared history, and that they understand that transnational cooperation is essential to make this information fully accessible online. For this reason, the HDSC adopted a strict open access policy. All generated datasets are open access under a cc-license. This means that researchers and other audiences have the same access everywhere, but users are not allowed to redistribute it behind a paywall. The datasets are published online with the archives that hold the original documents, although sometimes the website of the Dutch National Archives is used as a backup. The public can use an interface to search for individuals, see for instance the website of the National Archives of Suriname, in which the results are also linked to the digitized scans. The scientific datasets are stored in comma separated files (.csv) to facilitate their use in most software packages.[18] The scientific datasets are published open access on the Dataverse of the International Institute of Social History, and all have a data description in English that functions as a manual for the user. The scientific datasets offer extra variables and individuals have unique identifying numbers that link them among and within sources. This data will ultimately be converted to Linked Open Data. With the history of intense Caribbean mobility in mind, the optimal follow-up would be to cooperate with similar Caribbean projects and connect their data so that persons can be linked across islands and regions.

Sources, transcription and recognition

Sources

The core of the project is the transformation of the information about historical populations from entries in historical documents into searchable entities. Two types of sources are central to the HDSC citizen science project: the slave registers and the civil records (the birth, marriage and death certificates of free persons). The slave registers were a hybrid between population and property records that listed all enslaved people within a colony, organized by enslaver or plantation. These registers were introduced in Suriname in 1826 and in Curaçao in 1839 and were kept up to date until the abolition of slavery in the Dutch Caribbean on July 1, 1863. Other Dutch Caribbean islands began establishing slave records later. The aim of the slave registers was to combat the illegal slave trade that continued after the ban on the Transatlantic slave trade and to levy taxes on the ownership

[18] See for example: M. Rosenbaum-Feldbrügge et al*., Suriname slave and emancipation registers dataset version 1.1* (IISH Dataverse, 2023) Dataset, `https://hdl.handle.net/10622/CSPBHO`; C.W. van Galen et al., *Curaçao slave register dataset version 1.0* (IISH Dataverse, 2023) Dataset, `https://hdl.handle.net/10622/1OVORM`.

of enslaved persons. The slave registers were structured according to ownership; for each enslaver a list of owned individuals was documented. The entry for each person included at a minimum the first name (enslaved persons were not allowed to carry a surname in the Dutch colonies) and gender. Often these entries included an age or year of birth, and the name of the mother. The registers also documented the birth, death, transfer of ownership, manumission, or escape of enslaved persons. Unlike the slave registers in the British colonies, the Dutch slave registers were kept up to date continuously, which makes it easier to follow individuals through time.[19]

Civil records of the free population were introduced between 1828 and 1832 for each administrative unit in the Dutch Caribbean, either island or district in the case of Suriname and Curaçao. The civil records recorded life events: mainly birth, death and marriage. The records not only recorded the event and the person(s) affected, but also location, the names of the parents, and names of the declarant and witnesses. Enslaved people did not appear in these records until the abolition of slavery in the Dutch colonies in 1863. Surinamese records also excluded the indigenous inhabitants and maroons, descendants of escaped enslaved people, because they did not live under colonial rule. Like the slave registers, the civil records were all kept by civil servants in accordance with a rigidly enforced format. This makes it easier to transcribe the information and to connect the data between different records.

The records can be found in the collections of the National Archives of Curaçao and Suriname where the registers were restored and scanned, facilitating transcription from digital copies. The scans, and the databases created based on their transcription, are available on the websites of these archives as well as that of the National Archives of the Netherlands. The slave registers of Suriname and Curaçao combined consist of 53 volumes, and the databases based on the transcription contain over 180,000 records. The civil records are much larger collections. From the period dating to the start of the civil registration in 1828 until 1950, over half a million certificates of birth, death and marriage are available for transcription for Suriname alone.

Citizen science

To realize the transcription of hundreds of thousands of historical records, the HDSC invited – and received – the help of hundreds of citizen scientists. As mentioned above, in January 2017 a crowdfunding campaign and a consecutive online citizen science initiative took place, which resulted in the publication of online searchable databases of the slave registers of Suriname and Curaçao.[20] In September 2021, citizen scientists started to transcribe the vast collection of civil records from Suriname and Curaçao from 1828 into the first half of the 20th century, as part of a follow-up project. A first batch of which has already been published.[21]

[19] C.W. van Galen et al., "Slavery in Suriname. A reconstruction of life courses, 1830-1863"' *Historical Life Course Studies* 13 (2023): 191-211, `https://doi.org/10.51964/hlcs15619`.

[20] C.W. van Galen et al., *Suriname: Slavenregisters Dataset 1830-1863* (National Archives of Suriname, 2019) Dataset, `https://nationaalarchief.sr/onderzoeken/alle-genealogie/genealogie-slavenregister/persons`; E. Langenfeld et al., *Curaçao: Slavenregisters en emancipatieregisters 1839-1863* (National Archives of Curaçao, 2020) Dataset, `https://www.nationaalarchief.cw/api/picturae/slavenregister/persons`.

[21] M. Rosenbaum-Feldbrügge et al., Suriname Burgerlijke Stand: Geboorte- en overlijdensakten Paramaribo (1828–1921) (2024) Database, https://nationaalarchief.sr/onderzoeken/alle-genealogie/genealogie-burgerlijke-stand/persons.

Participants were invited to transcribe textual data from the scanned civil records into an online form. No financial rewards were offered for participation. Three tasks needed to be performed to transcribe each record, because each certificate was transcribed by two different citizen scientists and checked by a third more experienced citizen scientist to ensure high-quality transcriptions. At the time of this writing, over the course of three years, almost 700,000 tasks have been performed and over 230,000 civil records had been transcribed.

Setting up and maintaining a fruitful citizen science project requires a substantial investment in time and effort. Across the consecutive citizen science projects that together form the HDSC, over 1,500 people participated voluntarily in the transcription activities, of whom several hundred became regular, long-term participants. Participants were recruited through publicity in Dutch and Caribbean media and on social media, and via promotion through a series of online and on-site events. In order to stimulate continued participation in and engagement with the project, the project staff communicates regularly with the volunteers, via newsletters, social media, the online forum and online and on-site meetings on both sides of the Atlantic. On average, HDSC citizen scientists are over fifty years of age and a majority of them are women. More than half of the participants have family roots in Suriname or the Dutch Caribbean. Most participants live in the Netherlands. Those in Suriname and the Caribbean islands form a smaller group, but one should keep in mind that Suriname only has about 600,000 inhabitants and Curaçao about 150,000, versus 18 million inhabitants in the Netherlands, of which more than 600,000 are from Surinamese or Dutch Caribbean descent.[22]

Many participants in the Netherlands have ties to Suriname or the Dutch Caribbean and have joined out of interest in exploring their family history. But there is a sizable share of participants who admit to having had little prior knowledge about Suriname or the Caribbean before joining the initiative, let alone any family ties to its contentious history. These participants joined the project out of interest in genealogy, a desire to perform meaningful volunteer work, or the need to repair a historical injustice. The online forum is a key communal space for generating participant engagement. It is a repository of knowledge where participants exchange information and help each other with practical questions, like how to read or interpret certificates, often avoiding the need for project staff to respond. This knowledge exchange creates a sense of community that is reinforced by a second function of the forum: the sharing of stories and personal experiences related to the sources.[23]

Collaborating with citizen scientists has turned out to be a useful approach for the HDSC. Not only does it result in high quality data, but it also enhances the commitment of members of the public to a shared history. Even if the material cost of the approach is very modest, it does require quite an investment in time and care to recruit and retain the citizen scientists' involvement. It is essential to be attentive to the needs and various motivations of participants and to react quickly to any questions, suggestions and complaints. In the communication the project is never politicized but rather presented as a communal effort to foster open and equal access to a shared past.[24] Notwithstanding how sensitive the topic of colonial history can be, the HDSC has thus far remained a constructive and

[22] Van Galen, "Creating an audience," 182; Prats López, "Patterns of Engagement," 5.
[23] Van Galen, "Creating an audience," 193.
[24] Idem, 182-183.

supportive atmosphere without conflicts occurring between participants or with project coordinators.

Computer-aided transcription

Because of the sheer size of the transcription work still ahead, the HDSC began investigating the potential of employing computational methods in order to create a hybrid workflow that combines the effort of citizen scientists with input from the computer, in the form of algorithms and machine learning. The methods, currently applied to the Curaçao death certificates, are called handwritten text recognition (HTR) and entity recognition. The HDSC used the software Transkribus for HTR.

For a long time, archives were only described and accessible through indexes, basically inventory lists. With the 'Googlefication' of information, the last two decades saw a completely new approach to archival sources. Newspapers kept by western archives were the first sources to be digitally transcribed using optical character recognition (OCR). As a result, this type of source was also used first to automatically extract certain entities such as names or places.[25] More recently, archives have been transformed through the development of HTR, which enables the transcription of almost all written documents.[26]

This avenue is promising as it could increase the speed of rendering the documents accessible and searchable. To transpose information from scans into a database involves transcribing the scan via HTR and then using entity recognition to automatically extract dates, names, and places from the transcription and storing them as values within a database. The main challenge, as noted by earlier studies, is that errors made upstream impact all outcomes further down the process.[27]

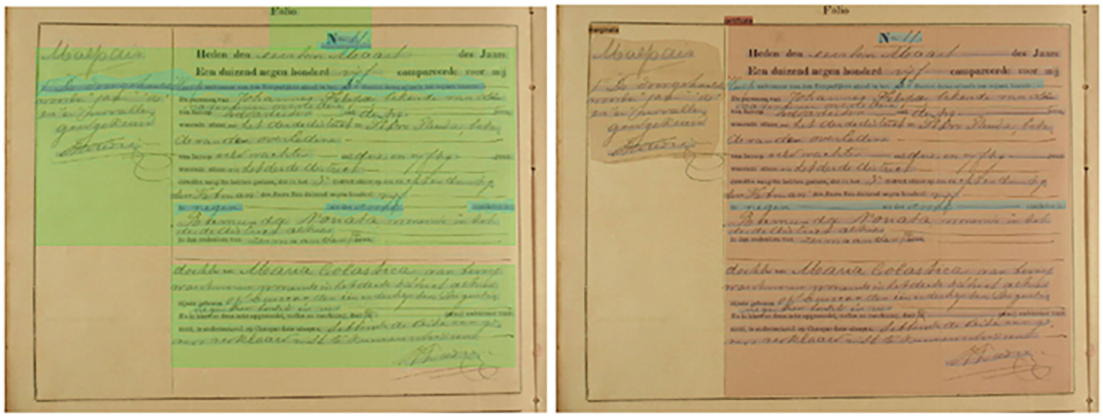
The first step in the HTR process is to correctly detect the layout of the scans. Transkribus offers off-the-shelf detection schemes for text regions, but because the default models suffered various errors a new layout had to be trained using Transkribus. Figure 2, a death certificate from Curaçao, presents an example of the kinds of issues faced. Marginalia on the document risks giving the false impression of one continuous baseline, such that the sentences from the marginalia mix with the text of the certificate. Furthermore, white space between words can split baselines, potentially confusing the word ordering. Both issues are present in (a), whereas (b) shows the same document after the layout detection was improved. As this step is the most upstream, it is crucial to improve layout detection. Training a layout model rather than using the Transkribus

[25] M. Ehrmann et al., "Named entity recognition and classification in historical documents: A survey," *ACM Computing Surveys* 56, no. 2 (2023): 1-47, https://doi.org/10.1145/360493; N. Abadie et al., "A Benchmark of Named Entity Recognition Approaches in Historical Documents Application to 19th Century French Directories," In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, edited by S. Uchida et al. (Springer, 2022): 445-460, https://doi.org/10.1007/978-3-031-06555-2_30.

[26] G. Muehlberger et al., "Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study," *Journal of documentation* 75, no. 5 (2019) 954-976, https://doi.org/10.1108/JD-07-2018-0114.

[27] M. Luthra, et al., "Unsilencing colonial archives via automated entity recognition," *Journal of Documentation* 80, no. 5 (2023), 18=080-1105, `https://doi.org/10.48550/arXiv.2210.02194`; C. B. Monroc et al., "A comprehensive study of open-source libraries for named entity recognition on handwritten historical documents," In *Document Analysis Systems: 15th IAPR International Workshop, DAS 2022, La Rochelle, France, May 22–25, 2022, Proceedings*, edited by S. Uchida et al. (Springer, 2022): 429-444, https://doi.org/10.1007/978-3-031-06555-2_29.

default creates better outcomes.[28] The layout models trained by HTR are freely available for other users.



(a) Certificate with errors. (b) Same certificate with errors solved.

Figure 2 Example of incorrect and correct lay out detection.

The second step in the HTR process is the automatic transcription of the text based on the legible base lines. Whereas OCR is based on a finite number of fonts, HTR is based on learning new characters, in this case those of the civil servants who wrote the certificates. The Transkribus model is trained on the shapes and characteristics of the documents by feeding it more than 100 fully transcribed certificates (ground truth). Furthermore, the model makes use of a language model to make educated guesses for characters that the model does not recognize. Unseen sequences of characters can be replaced by the closest of the known words in the language model. Consequently, each model is limited to a particular style of writing combined with a language, especially if the training sample is relatively small. So-called character error rates (CER), which indicate the share of incorrect characters when compared to the total number of characters, are typically used to evaluate HTR transcriptions. Earlier studies have shown that a character error rate below five percent is needed to obtain good results later in the process.[29] The Curaçao HTR project achieved a CER of 3.9%.

After the documents have been transcribed with HTR using Transkribus, the relevant entities from the transcribed documents are extracted and stored in a database. In other words, from a string of text generated with HTR entities like a name, date, occupation, and place of birth are automatically distilled. For entity extraction the HDSC team explored both the use of regular expressions (relying on a standardized structure of the source document to find the location of an entity) and large language models (ChatGPT version 3.5 and 4.0). After tests, the team chose ChatGPT 3.5 and 4.0 because large language models outperformed regular expressions for extracting entities such as dates and names. While the share of correctly identified dates is acceptable for ChatGPT (>80%), ChatGPT only succeeded in extracting names correctly in roughly one third of the cases. The main

[28] For default approach, see: C. M. Dahl et al., "Applications of machine learning in tabular document digitisation," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 56, no. 1 (2023): 34-48, https://doi.org/10.1080/01615440.2023.2164879.

[29] Monroc, "A comprehensive study."

reason for the low accuracy with names was that HTR produced quite many character errors for names, as the Transkribus language model had difficulties accounting for the various first and family names derived from the multiple languages spoken in the Caribbean context.

The current quality of HTR and name extraction does not match the quality needed for publication of the extracted data for the wider public. However, the HDSC recently started a pilot to integrate automatic transcription into the current citizen science workflow. The citizen scientists are shown automatically generated pre-filled entities that they are then able to verify and correct, like dates, names, occupations, etc. This accelerates the transcription process, allowing for the substitution of one of the two human transcribers with the computer-aided transcription pipeline. Nevertheless, this solution raises new questions. The citizen scientists might miss errors in the prefilled fields, might lose motivation to work with the records or might feel made redundant. Initial feedback from the citizen scientists indicated that speeding up the project by adding pre-filled entities is appreciated as their main motivation to participate is to make the data available. Nevertheless, the participants also mentioned that their participation depended on the quality of the computer-generated output as they refused to check low quality transcriptions. This particular case makes clear the importance of continuous dialogue between the main project team and the community of citizen scientists.

Linking

Linking sources is pivotal to the project as it brings both unique and general but illustrative life courses to the surface. For example, in June 1856, a seventeen-year-old enslaved girl named Roset, was sold in Aruba by J. Lampe to David Gaerste Jz. She was then transported to Curaçao and sold to Nicolaas Curiel upon arrival. Roset escaped her enslaver at least two times in January and June 1859. Both times she temporarily fled from the Otrobanda neighbourhood of Willemstad where many enslaved people lived. Roset was not able to secure her freedom. In May 1860, she was sold to Martis Felix and worked for him as a housemaid. After the sale, Roset made several new attempts to escape. In September 1861 she received ten lashes and was sentenced to wear shackles due to “repeated acts of absconding.”[30] She remained enslaved until emancipation in 1863 after which she received the surname Veldhoop. Tragically, she died after only 53 days of freedom at the age of twenty-six. Her death certificate states that she was born in Curaçao. In the Curaçao escape registers, she was described as having “a small and plump posture with brown eyes and a wart on her right cheek.”[31]

Reconstructing Roset’s life course required establishing links between and within various sources. In the slave registers, each enslaver had his or her own entry. When an enslaved individual was sold, this person was added to the list of enslaved of the new owner. Linking those discrete entries was thus necessary to reconstructing Roset’s life course in slavery. Because her enslaver Nicolaas Curiel reported her escape attempts, a physical description of Roset preserved in the archive emerges from the linking of the slave and the escape registers. Additionally, her conviction in 1861 was officially

[30] National Archive of Curacao (NAC), 005-3-31 Register van door de politie gestrafte slaven, 1857–1862, 20 September 1861.

[31] National Archive of Curacao (NAC), 005-3-08 Register weggelopen slaven, 4 January and 6 June 1859.

recorded by the Attorney General and could be linked, too. In 1863, when slavery was abolished, all formerly enslaved persons were given a last name which was recorded in the emancipation register, requiring yet another link between sources. This last name was needed to establish a link to the death certificate, as this record included Roset's full name as she was a free person at the time.

The linking process was done with linking algorithms because of the large number of names involved. As long as a set of basic rules are followed the process can easily be replicated, because it is done using the programming language R.The core of the linking algorithm can be found on GitHub.[32] It does not matter whether sources are linked within (as was done within the slave registers) or between sources (as was done between the slave registers and the other documents). The first step in the process is to identify all possible links based on the name of the individual. Because the spelling of the name often differs between historical sources, a Levenshtein distance threshold was implemented in the linking process which measures the similarity between two names.[33] For each possible link the number of character insertions, substitutions and deletions needed to match the first name was calculated. For example, in table 1 (below), matching Roset to Rozet required one alteration: replacing the z with an s. Matching Roset to Rosette, required two alterations: deleting the t and the e. Depending on the length of the name, a threshold of three alterations was used. Lowering the threshold reduces the number of successful matches but raising it risks introducing more incorrect matches (so-called false positives). Trial and error is ultimately necessary to establish the ideal number of alterations.

Registration and the use of names in the Dutch Caribbean was quite flexible and thus requires a flexible approach to linking. This is difficult with the available digital humanities software tools for data-linking, which has been developed based on the available digital archives in the United States and Europe. Given the rigid structure of the various European nation states, the current software struggles with the fluid nature of the Caribbean context. For example, when an enslaved individual was sold, but the new enslaver already had an enslaved person with the same name, a nickname, number or mother's name was added to be able to make a distinction. Additionally, names in the free population were often not cast in stone. Those who migrated between islands sometimes changed their last names. And even the names of members of the local population could change overtime. For example, Karel Panneflek Faulborn died at less than two years-old and his first name changed to Carl. The same is true for Carlos Carado Rodriquez whose name at death was recorded as Karel Koronal Rodriquez. In Curaçao, examples were found wherein a matronym was used for record-keeping rather than the given family name. In either case, the names were split prior to the linking process because, as table 1 shows, leaving the additional names results in a higher Levenshtein distance. As a result, the linking then only occurs based on a last name or second name.

[32] R. J. Mourits and M. Rosenbaum-Feldbrügge, *Property to Person* (2022), `https://github.com/RJMourits/P2P`.

[33] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady* 10, no. 8 (1966): 707–710.

Name 1	*Name 2*	*Levenshtein distance*
Roset	Rozet	1
Roset	Rosette	2
Roset	Rozette	3
Roset	Roset Maria	6

Figure 3 Levenshtein distance examples using the name Roset.

To make sure that no links are established between individuals with similar names, additional parameters are used to find the best match. For example, in the case of multiple possible links in the slave registers, a second step was taken in which the Levenshtein distance between the mother's name and the enslaver's name was calculated as added variables. Other steps include age and even order. Age or birth year should be within a reasonable margin if mentioned in a second source. In the case of an event, the end date of the previous entry must match that of the new entry.

Overall, although the exact approach is very much context dependent, following the basic principles of the linking process allows for a flexible approach that is well suited to the Caribbean context. Using this linking protocol, most life courses can be reconstructed. To also link the remainder of the data, researchers might decide to relax thresholds or to check the remaining entries manually. At present, the search interfaces for the general public use the unlinked databases. This presents the most accurate set of data, which is desirable for users that are mostly searching for individual names and families. The linked files, available for open access downloads, are primarily meant for scholarly use. For a large-scale analysis, the number of linking errors that might be disruptive for individual users is small enough to not influence the outcomes.

An example of how the Caribbean context of the early nineteenth century can affect the linking process concerns the civil certificates of manumitted individuals. Upon manumission and emancipation, the enslaved often received a surname that resembled that of their former enslaver. For example, Nicolina Martina Daal was manumitted by the Daalen family. Table 2 provides the Levenshtein distances between the last name of the enslaver and of the enslaved upon manumission for 1,527 manumissions on Curaçao between 1839 and 1863. In more than 300 of those cases the newly received last name of the manumitted individual falls within the default Levenshtein distance of 3 when compared to the last name of the former enslaver. This shows that extra variables such as age and mother's name must be incorporated in the linking process to avoid erroneously linking the descendants of former enslaved to the descendants of former enslavers.

Levenshtein distance	N	%
0	17	1%
1	69	5%
2	137	9%
3	122	8%
4	135	9%
5	219	14%
5+	828	54%
Total	1527	100%

Figure 4 Levenshtein distance between last name enslaver and enslaved upon manumission, Curaçao 1839-1863.

Additional tools and contextual datasets

Beyond creating life course databases, the project facilitates further research applications by providing additional contextual tools and datasets. A striking example is the creation and publication of shapefiles that can be used to analyse the data geographically. Figure 3 plots the administrative districts used in the first half of the 19th century on the island of Curaçao. The figure shows the enslaved population as a share of the total population per administrative district. The map makes apparent that the free population comprised a substantial share of the community in all districts. In the urban district of Willemstad, a minority of 29 percent of the population was enslaved. In the rural districts close to the city, the share of enslaved was around 37 percent. This was relatively low, especially when compared to the more rural districts further away from the city where the majority of the population was enslaved. By plotting the large amount of collected data on a map, insights into how slavery was shaped geographically on an island like Curaçao emerge.

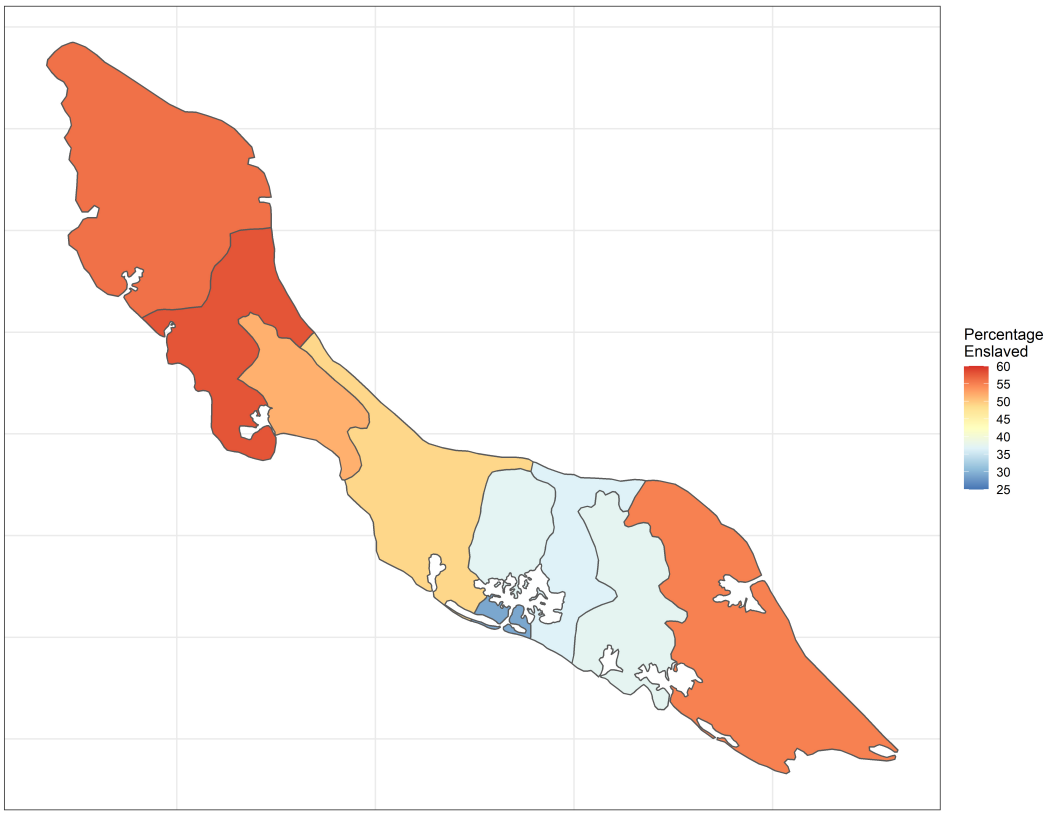


Figure 5 Enslaved population as percentage of the total population per administrative district in Curaçao 1840. Source: National archive of the Netherlands, inventory number: 1.05.12.02 Inventaris van de archieven van Curaçao, Bonaire en Aruba, (1771) 1828-1845 (1914), 172 Bevolkingsstaten van Curaçao alsmede naamlijsten van plantages, tuinen en hun eigenaars, opgemaakt door de wijk- en districtmeesters, 1838-1845.

In mapping the past, the project is hampered by the relatively small number of surviving historical maps. Additionally, the project must grapple with descriptions of changing ownership or borders that cannot be accurately represented on the maps. Although geo-referencing of the maps is helpful, a balance between accuracy and clarity must be found. The developed digital maps should be viewed with caution, as the aim of the visualisation is to provide a general overview, and borders on the maps might remain slightly uncertain. Nonetheless, for the purpose of locating the life courses of enslaved and free individuals on the islands, these maps are sufficient. The databases ensure that the location IDs match those used in the maps.

It is important to study individuals within their (spatial) contexts. Accordingly, the HDSC also publishes datasets that can be added to the life course databases to embed individuals geographically and socially. To give an example; Suriname plantation databases that contain detailed information on the number of enslaved persons, plantation product, and location for each plantation present in the period 1830-1863 have been published, as has a database of the Paramaribo Ward Registers that provides detailed information on the residents of Suriname's capital Paramaribo in the period 1828-1847.[34] As in case

[34] M. Rosenbaum-Feldbrügge et al., *Suriname plantation dataset version 1.0* (IISH Dataverse, 2023) Dataset, Retrieved from `https://hdl.handle.net/10622/VTL43W`; T. van Oort et al., *Plantations Surinaamse Almanakken, 1818-1860* (IISH

of the life course databases, these contextual databases come with detailed manuals and variable descriptions.

Conclusion

This article was meant as an invitation to work together on trans-Caribbean population data projects. For this reason, the core building blocks and challenges of the Historical Database of Suriname and the Caribbean (HDSC) project's workflows were presented to provide necessary knowledge for scientific and heritage communities aspiring to pursue similar projects. It describes in detail how the project generates public and scientific databases of enslaved and free historical populations based on archival sources, from the scanning and digitization of the sources to the hybrid transcription process to the linking of individuals within and across databases.

The success of the HDSC can be attributed to its foundation, a network of different types of partners: archives, historical researchers, data specialists, societal partners and citizen scientists. Local knowledge of citizen scientists, professionals, and scholars who are familiar with sources and context can be backed up by funding opportunities and applicable digital humanities tools available in the Caribbean, the United States and Europe. The aim of the project, publishing the generated data online open access, available to everyone, adds to the synergy of the consortium. As obvious as that may sound, as partners stem from different backgrounds, historical sensitivities, unclear definitions, and limited financial capacity might too easily result in misunderstandings.

To make use of everyone's strengths the devil is in the details. Projects members must lay out their expectations clearly and not resort to buzz words that might do well in grant applications. What is open access, what are digital humanities, or how is access to these reconstructed life courses guaranteed? To take the example of open access: the term open access could imply that all data is shared across the internet and stored, kept, and used by whoever is interested. However, for the HDSC open access means that local archives are able to share datasets of their archival material with the world on their own platforms without any limitations, presenting the data in a historical context.

What stands out from the HDSC project is that its scholarly success is inextricably connected to its societal impact. The project enables members of the wider public to retrace and recreate their own history and that of others through their involvement in the data creation. Appreciating the value of discovering and studying one's own history has enabled the project to mobilize a large crowd of citizen scientists in both the Caribbean and Europe. Public events fully focused on the societal impact for individual participants has heightened the motivation of citizen scientists and created the feeling that all are in this together. It has become abundantly clear that the success of ambitious scholarly endeavours like the HDSC lies in the grassroot foundation of such projects. Scholars must think, therefore, beyond research questions and scientific relevance if their aim is to successfully execute similar projects. The HDSC constantly seeks new partners across geographies and languages, understanding that to study Caribbean history and the lives of Caribbean individuals demands pushing against the limits of colonial borders.

Dataverse, 2023) Dataset, `https://hdl.handle.net/10622/4VOJYS`; J. Sang-Ajang et al., *Paramaribo Ward Registers 1828-1847* (IISH Dataverse, 2023) Dataset, `https://hdl.handle.net/10622/VLN8FD`.

As mentioned above, it is crucial to engage new tools from the field of digital humanities, recognizing that the American or European origin of many of these tools can make them challenging to utilize in the Caribbean context. By training supposedly universal software on Caribbean documents, the HDSC helps to ensure that new models are developed for Caribbean contexts. It is the hope that newly established technologies like HTR along with large-scale projects like the HDSC, the Danish *Into The Same Sea* project, and the Danish West Indies (DWI) panel[35] will motivate other researchers to make more Caribbean population data available in the upcoming years. Such linking between and across contexts will enable further research into, among else, Caribbean migration patterns both before and well beyond the abolition of slavery.

[35] S. Galli et al., "Reconstructing a slave society: Building the DWI panel, 1760-1914," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 57, no. 3 (2024): 163–184, https://doi.org/10.1080/01615440.2024.2400188.

Björn Quanjer

Dr. Björn Quanjer is a specialist in historical microdata and researcher for the Historical database of Suriname and the Caribbean (HDSC) project. His research focusses on inequalities in wealth and health in historical populations. Björn constructs and uses large scale population database to explore social, economic and demographic trends in the past. Within the HDSC project Björn focusses on the island of Curaçao. Björn's PhD research entailed a study of early life determinants of adult height within the Netherlands between 1850-1950. In 2024 he defended his dissertation cum laude, using household level data linked to conscription records.

Thunnis van Oort

Dr. Thunnis van Oort is a historian interested in digital methods and techniques, citizen science, cultural heritage, colonial history and media history. He has researched and/or taught at Radboud University, University of Amsterdam, Oxford Brookes University, Antwerp University, University College Roosevelt and Utrecht University.

Matthias Rosenbaum-Feldbrügge

Dr. Matthias Rosenbaum-Feldbrügge is a historical demographer at Radboud University Nijmegen, the Netherlands. As a specialist in historical microdata, he is involved in the construction of the Historical Database of Suriname and the Caribbean (HDSC). Through advanced record linkage methods, Matthias also reconstructs the life courses of the enslaved population and their descendants which enables innovative research into kinship networks, health inequalities, and living standards during and after slavery. Matthias has published in demographic journals such as Demography, European Journal of Population, Population, Space, and Place, The History of the Family, and Demographic Research.

Cornelis W. (Coen) van Galen

Dr. Cornelis W. (Coen) van Galen is an associate professor in social and colonial history at Radboud University Nijmegen. He is the project manager of the Historical Database of Suriname and the Caribbean (HDSC), a combined data and citizen science project aimed at the production of open access datasets of the population of Suriname and the Dutch Caribbean between 1830 and 1950.

Lisa Hoek

Lisa Hoek, MA, is an independent data scientist who completed her master's degree extracting entities from Curaçaoan death certificates. Her latest work involves handwritten text recognition and prompt engineering for large language models. She also specialises

in other machine learning and AI, such as natural language processing and named entity recognition

Erik Tjong Kim Sang

Dr. Erik Tjong Kim Sang is a research software engineer of the Netherlands eScience Center in Amsterdam. After obtaining a PhD in computational linguistics from Groningen (RuG), he held lecturer and postdoc positions in Uppsala (UU), Antwerp (UA), Tilburg (UvT), Amsterdam (UvA and Meertens Institute) and Groningen (RuG). His research focuses on machine learning applications of natural language processing, in particular in named entity recognition and social media analysis.