Decolonising the Archives:

'Two Centuries of Indian Print' at the British Library

Dr. Nur Sobers-Khan, Principal Investigator

Dr. Priyanka Basu, Project Curator

Tom Derrick, Digital Curator

Olivia Majumdar, Project Cataloguer

Paramdip Khera, Project Support Officer

'Two Centuries of Indian Print' at the British Library is a research and digitisation project in collaboration with the School of Cultural Texts and Records (SCTR) at Jadavpur University, Srishti Institute of Art, Design and Technology in Bangalore and the SOAS Library in London. It is an extensive and rich repository of rare printed Bengali, Assamese and Sylheti books of which approximately 1800 titles have now been digitised and made available online. This international collaborative project has been running for nearly 5 years now with funding from the AHRC Newton-Bhabha Fund and the Department for Business, Energy and Industrial Strategy, UK. Focussing on early printed books between 1714 and 1914, the project's key aims have been towards complementing digitisation and research with skills-sharing, capacity-building and working with Indian institutions (e.g., the National Library in Kolkata, Asiatic Society in Mumbai, and Jadavpur University). *The Book Unbound*, a book history symposium (2017) and two workshops on 'Print and Islam in South Asia' (2018) have been organised as part of the project, alongside the ongoing British Library South Asia Seminar Series which offers a research dissemination platform to doctoral students, early career academics and senior scholars. In the current climate of continuing uncertainties due to the COVID-19 pandemic, the online digital collection of 'Two Centuries of Indian Print' has been a valuable resource for researchers who are unable to travel and access the physical collections at the British Library owing to intermittent lockdowns and restricted opening of the reading rooms. Through its research, cataloguing and digital humanities aspects, 'Two Centuries of Indian Print' has incorporated some vital aspects of decolonisation with respect to colonial archives.

Research, Dissemination and Access

The practice of legal deposits has existed in English law since 1662. With the establishment of the printing industry in Bengal in the early nineteenth century, 'The (Indian) Press and Registration of Books Act' was passed in 1867 in order to record the number of books being published from the Indian provinces. It, therefore, 'became mandatory for all books published in provinces of British India to be sent to the provincial secretariat library for registration' (Derrick 2016). As a result, both the India Office Library and the British Museum Library (later to come jointly under the British Library) 'were separately given the privilege of requesting books from these lists free of charge in what amounted to a colonial legal deposit arrangement' (Ibid.). While the intent of this measure was primarily archival, it was political too as a close look at the debates around readership, cultural taste and government interventions in nineteenth century Bengal reveals. While cheap printed literature in the form of chapbooks were some of the prolific publications from Calcutta's Battala (similar to London's Grub Street), it eventually lead to their proscription under the 'Obscene Books and Pictures Act' of 1856 (Moitra 2018). Revd. James Long, who brought out extensive catalogues of books published in the Bengal Presidency, was one of the prominent figures demanding the proscription of this low-life literature (Ghosh 2006) owing to their 'obscene' content. Later in the century, the 'Dramatic Performances Control Act' (1872) was a step further in proscribing printed literature and performances under colonial surveillance.

The impact of such colonial scrutiny and containment has been manifold and long-drawn. The questions of sustainability and access have thus been important and challenging to grapple with both for archivists and researchers in South Asia as the place of origin for these publications. The colonial system of legally acquiring the early printed books have ensured in most cases that duplicate and multiple copies of these books are extremely difficult or impossible to locate in other libraries, especially those in South Asia. As Siobhan Senier (2014) has shown in decolonizing the archives in the context of Native American writers in New England, sustainability is closely tied to ecological and economic imbalances. While ecological factors have been instrumental in deterring systematic conservation and preservation of rare printed collections in South Asian archives, economic factors impede many researchers

(largely from the Global South) from accessing resources housed in museums and archives, e.g. as that of the British Library.

The decolonising methodology in relation to archives is incumbent on challenging colonial ideologies/inheritances on the one hand and actively engaging in reuniting archives, at least digitally, in locales where colonialism has dismantled or destroyed resources. Equally important are the responsibilities of researchers who work with such collections since writing on them entails informing the existent archives/archival methodologies as well. The digitisation of early rare printed books within 'Two Centuries of Indian Print' has ensured an increasing engagement with the digitally available resources for researchers based in South Asia and an active dissemination of research in the areas of book history, readership, gender, performance and printing technologies. In its current phase of research and engagement, the project is thus invested in exploring three broad yet related areas of research:

(i) *Visibility and Access*: The project is looking to compile a comprehensive catalogue based on the digitised collections. This stems from queries from a number of South Asian scholars and researchers who have been facing difficulties in accessing the material due to the lack of a comprehensive catalogue dedicated to the 'Two Centuries of Indian Print' collections alone. Visibility and access of the collections will also include highlighting caste hierarchies that occlude the presence of low-caste printers and publishers, and critiquing racialized hierarchies of digitisation project work flows and creation of 'ground truths'.

(ii) *De-centring Calcutta*: While a large part of the scholarship in Muslim Bengali literary output has focussed on *punthi* literature in Musalmani Bangla in nineteenth century Bengal, the role of Muslim writers, printers, publishers as well as sellers have largely been marginalised owing to the dominance of Calcutta as an urban centre in nineteenth century literature and cultural discourses. A number of publications within the project were published from Dhaka and the adjoining districts, a close reading of which can be a potential way of looking at connected geographies and bringing east Bengal/Bangladesh back in to the discussion on print histories.

(iii) *Expanding on Languages*: While the focus of the project has been to digitize and publish online Bengali-language material (one of the largest in the South Asian language collections of the British Library), it is intended that in the subsequent phases Hindi, Punjabi and Nepali will be added to the existing collection, primarily through the creation of catalogue records.

Cataloguing Collection Items

In the museum and archive sector, an increasing number of institutions are committed to demonstrating their robust commitment to antiracism and equality. This has led in turn to a greater scrutiny given to previously accepted systems of classifying and cataloguing collections. As a digitisation and cataloguing project, Two Centuries of Indian Print (2CIP) has to carefully consider the language used to label and record the items in the British Library's collections. This particularly relates to the use of Library of Congress Subject Headings (LCSH), which are terms controlled and authorised by the Library of Congress, for use in categorising catalogued records and as a helpful way to filter records by subject and theme. It has been useful for the 2CIP project to refer to work undertaken by other institutions on this topic, for example the work of the Association for Manitoba Archives (AMA) in Canada. A 2016 article by Christine Bone details the work undertaken by the AMA to identify and improve upon inappropriate language used to catalogue their collections. Bone notes that one of the major problems of the existing language was the use of the word "Indian", which is a 'generally outdated' term in Canada (Bone 2016: 2). In this case, 'changing "Indians of North America" to "First Nations" would be a huge improvement and would bring [AMA's] subject headings in line with more current and accurate terminology' (Ibid: 3). This example is instructive as it effectively shows how an institution can carefully assess the existing language used to describe their collections and offer replacements for terms that may be unsuitable and even offensive. In our cataloguing work, 2CIP seeks to sensitively and appropriately describe and classify collection items so that they are both easily accessible to Library users and also reflect the Library's commitment to antiracism and equality. We are also working closely with colleagues at Jadavpur University in Kolkata, India to obtain their feedback and input on our digitisation and cataloguing work.

Digital Humanities

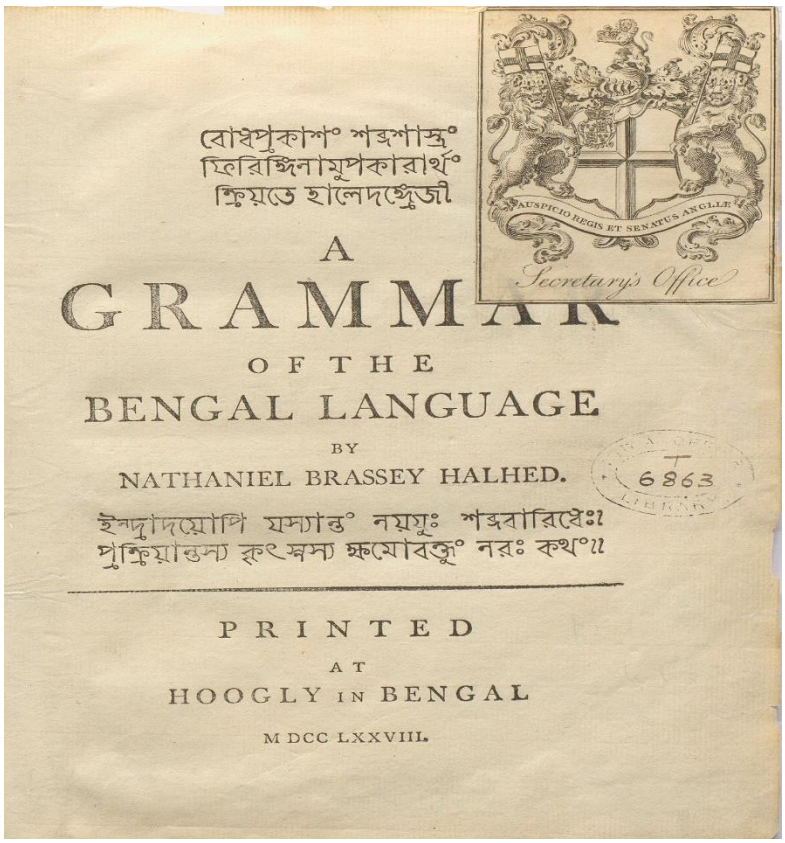
One of our core ambitions with this project has been to obtain machine readable text for the books, enabling key word search of the digital editions that will translate into quicker and more efficient navigation for online users to relevant passages of text. The technology used to achieve this, known as optical character recognition (OCR), has historically either underserved, or completely bypassed Bangla and other South Asian languages. In contrast, OCR for western languages can be very good indeed, leading to reliable search and retrieval experiences, as well as raw OCR datasets that digital humanists can be confident in using for their text analysis and data mining.

Whilst it would be naive to think the Two Centuries of Indian Print project could redress the imbalance between east and west when it comes to OCR provision and performance, we hope the initiatives we have undertaken go some way towards helping institutions in India and those with Bengali collections make the full contents of their textual material available for research. Central to these initiatives has been the creation of 'ground truth'; a representative selection of pages from the books manually transcribed by members of Jadavpur University's School of Cultural Texts and Records (SCTR). This ground truth has been used to train Transkribus to automate recognition of the printed Bangla texts with a high degree of accuracy.

So often the success of OCR is credited to the technical capabilities of the software used. Whilst the advancements in artificial intelligence and computer vision continue to benefit those of us digitising South Asian languages, OCR technology is still largely reliant on being trained to recognise the specific language and scripts it is asked to automate. The best training is through manually created, 'word perfect' transcriptions. This is why the contribution from the SCTR and language skills they bring, has been essential in providing a working solution for OCR of printed Bangla. The Bangla trained model in Transkribus is available for other Transkribus users who have printed Bangla collections they would like to automate text recognition for. The ground truth is available as an open dataset on the British Library's Research Repository.

Additionally, the project has run several *digital skills* workshops in India. These events have served as important forums for likeminded GLAM professionals and students to share skills and knowledge for making South Asian cultural heritage more digitally accessible. Bringing together representatives institutions in one place has fostered healthy conversation around the

specific challenges facing digitisation of collections belonging to this region and the strategies, tools and practices to address them.



বোধপ্রকাশং শব্দশাস্ত্রং
ফিরিঙ্গিনামুপকারার্থং
ক্রিয়তে হালেদখ্রৈজী.

A
GRAMMAR
OF THE
BENGAL LANGUAGE
BY
NATHANIEL BRASSEY HALHED.

ইন্দ্রাদয়োপি যস্যান্তং নযযুঃ শব্দবারিধেঃ।
প্রক্রিয়ান্তস্য কৃৎস্নস্য ক্ষমোবক্তুং নরঃ কথং॥

PRINTED
AT
HOOGLY IN BENGAL
M DCC LXXVIII.

AUSPICIO REGIS ET SENATUS ANGLIÆ
Secretary's Office

6863

Works Cited

Bone, Christine. "Modifications to the Library of Congress Subject Headings for use by Manitoba archives." Paper presented at: IFLA WLIC 2016. Web. 14 December 2020.

Derrick, Tom. "Quarterly Lists: Digitally Researching Catalogues of Indian Books." *Early Indian Printed Books*. 2017. Web. 14 December 2020.

Ghosh, Anindita. *Power in Print: Popular Publishing and the Politics of Language and Culture in a Colonial Society, 1778-1905*. New York: Oxford University Press. 2006. Print.

Moitra, Swati. "The World of Battala." *Sahapedia*, 2018. Web. 14 December 2020.

Senier, Siobhan. "Decolonizing the Archive: Digitizing Native Literature with Students and Tribal Communities." *Resilience: A Journal of the Environmental Humanities* 1:3 (2014): n. pag. Web. 14 December. 2020.