Sussex University contribution to the Botany in British India: enhancement through transcription project: a note

1. Project summary

- To use the Transkribus transcription tool to create transcriptions of the digitised Botany in British India collections, and to make these available for bulk download and reuse alongside catalogue metadata.

2. The collection

- Botany in British India (1780-1890) http://www.bl.uk/reshelp/findhelpregion/asia/india/indiaofficerecords/botanymat.html

- The records of the East India Company and India Office are a rich source of information about surgeon-naturalists, economic and medicinal botany, and the establishment of botanic gardens in the UK and India.

2.1. Institutions:

- Botanic and spice gardens at Bangalore, Calcutta, Dapuri, Marmalong, Ootacamund, Pune, Saharanpur, Samalkot, Tinnevelly to name a few.
- The Royal Asiatic Society, the Agricultural and Horticultural Society of Calcutta.

2.2. Individuals (a selection):

- Benjamin Heyne, 1770-1819, botanist | Whitelaw Ainslie, 1794-1837, surgeon | Charles Lush, 1797-1845, surgeon botanist | Robert Wight, botanist | Hugh Falconer, 1808-1865 | William O'Shaughnessy, 1808-1899, doctor | Johan Eliza de Vrij, 1813-1898, botanist | Clements Markham, 1830-1916, geographer | Robert Kyd, 1746-1793, botanist | James Anderson, 1738-1809, botanist | Andrew Berry, 1764-1833 | Nathaniel Wallich, 1785-1854, botanist | William Roxburgh, 1751-1815, botanist | Hugh Cleghorn | William Augustus Gott | Dr John Forbes Royle, 1798-1858, naturalist

2.3. Plants and commodities:

- Including cannabis, cassia, cinchona, cinnamon, cochineal, cotton, hemp, indigo, nutmeg, rubber, senna, silk, sugar, tobacco.

3. The project

- India Office staff will set up a private collection on the Transkribus server and provide access to an initial 200 images drawn from the BIBI collection. We will provide you with a copy of the catalogue metadata, and you can select items of particular interest to yourself.

- Training in the tool will be provided by Alex Hailey and Antonia Moon during a visit to Sussex.

- Once 100 / 200 pages have been transcribed, the Transcriptorium project will train their handwriting recognition tool on the "ground truth" transcriptions and images**.**

- From this point only the baselines of images need to be marked up: the automatic transcription can then be run. Researchers would be invited to select materials to be transcribed.

- The tool is currently creating transcriptions with 70% accuracy on the Bentham project material, and we hope to see similar results. The results of transcription – including errors – themselves are of interest to the Library in the context of developing HR and OCR technology.

- It takes roughly 5 minutes (5-10 before you are familiar with the tool) to mark-up the necessary regions in the image, and then normally a further 10 minutes to produce a full transcription. Say **25-30 hours** to produce the first 100 transcriptions, **50-60 hours** to produce 200.

4. Benefits to the researcher and to University of Sussex

- Experience of collaborative transcription and a more hands-on approach to collections through the use of a cutting-edge tool, potentially very useful to a postgraduate student / early career academic.

- Contribute in a small way to the development of handwriting recognition technology.

- The chance to help develop a community of interest around these materials and regarding botanical transcription projects, of which there are several recent examples: the Wallace and Joseph Hooker Correspondence projects at Royal Botanic Gardens Kew and the Natural History Museum being best known.

- Improve own palaeography skills, a core historical research skill.

5. Benefits to the British Library and to wider research community

- Full transcriptions open the material up to a range of enhanced search and analysis methods, increasing its attractiveness to researchers.

- From 100 or 200 manually transcribed images, and baselines identified on the rest of the material, the project would generate c8,600 pages of transcription, hopefully at 70% accuracy (before further manual correction).

- The botany material represents a tiny sample of the Board's collections and the Government Proceedings, which contain many hundreds of thousands of manuscript pages. The small manual input has the potential to produce a significant output in terms of automated transcribed pages in the long term.

- The Library would demonstrate its willingness to engage with emerging trends in digital research, strengthen existing relationships with research institutions working in this area, and show a commitment to enhancing its existing digital resources.

Alex Hailey
14/7/15