The Sir Robert Hart collection: some experiences of manuscript digitization

In the autumn of 2007 I was a research assistant on a British Academy-funded digitization pilot project. I should note that I use the term 'digitization' to describe the entire process through which material is made available in a digitized format: this process can include preparing meta data (a catalogue, pagination, index, etc), transcribing, creating and maintaining a website, and finally scanning material to create digital images. Funding bodies may use the term digitization more strictly to mean creating digital images. Our British Academy grant funded the preparation of the manuscript for scanning, but it did not fund the scanning itself.

The manuscript collection funded for pilot digitization was the Sir Robert Hart Collection, held in the Special Collections at the Queen's University Belfast Library. The project was supervised Deirdre Wildy, Senior Humanities Librarian at Queen's. My responsibility over three months was to transcribe and prepare one diary in the collection for scanning and to create a new catalogue of the correspondence in the collection. Some aspects of our experience might be useful to historians and librarians considering digitization projects.

The Collection

Sir Robert Hart was born near Belfast in 1835. Following his graduation from Queen's he entered diplomatic service in China. In 1863 Hart became Inspector General of the Chinese Imperial Maritime Customs Service, making him an employee of the Chinese government and the head of a large international workforce. On his retirement Hart became Chancellor of Queen's. Hart's family donated a large collection of his papers to the university in 1970. The Hart collection at Queen's contains all of Hart's seventy-seven diaries, 7,800 items of his correspondence, 700 photographs (c.1900-1913), and a wide range of ephemera from Hart's life in China. The collection is an excellent resource for the social, economic and diplomatic history of late-imperial China, as well as the history of European expatriate communities in Asia.

Reasons for digitizing

The primary reason for digitizing the collection is to make it more accessible. Only the first eight volumes of Hart's diaries have been published, although all the Hart correspondence held at Harvard University has been published.[1] Furthermore, Hart's penmanship is a major obstacle for researchers. Fairbanks, Bruner and Matheson noted that the letters they published had been previously typed up by a secretary familiar with the "complexities" of Hart's handwriting.[2] Edward Lefevour remarked, in an article describing the Queen's collection, that "Hart's

[1] The Queen's diaries were published in two volumes as Katherine F. Bruner, John K. Fairbank, and Richard J. Smith (eds), *Entering China's service: Robert Hart's journals, 1854-1863* (Cambridge, Mass, London: Harvard University Press, 1986) and *Robert Hart and China's early modernization: his journals, 1863-1866* (Cambridge, Mass, London: Harvard University Press, 1991). The Harvard correspondence was published as John King Fairbank, Katherine Frost Bruner, and Elizabeth MacLeod Matheson (eds), *The IG in Peking: letters of Robert Hart, Chinese Maritime Customs, 1868-1907.* Cambridge, Mass., London: Belknap Press of Harvard University Press, 1975.

[2] J. K. Fairbank et. al., *The IG in Peking,* vol. 1, p. xvi

handwriting presents a problem. Volumes 13 through 35 … can be read with difficulty, though a few passages defy the most patient study. From about 1885 the hand changes and becomes much more difficult to read."[3] I had used some of the Hart collection as a master's student and had developed familiarity with his hand and with the issues discussed in the diary.

The digitization process was also facilitated by the Centre for Data Digitization and Analysis (CDDA), a technical and research unit within the School of Geography at Queen's. CDDA scanned all the diary pages, allowing me to work off of digital images. The availability of on-campus scanning was convenient and eliminated worries and costs of moving manuscript materials.

Approach

The first steps involved in preparing the manuscript were to create the meta data. One misconception of digitization is that the bulk of the work goes into scanning the images, but in fact simply putting images on a website does not create a useful and trustworthy research resource. Each image has to be reliably identified with a part of the original manuscript. In this case, I spent the first week of the pilot project paginating the diary and assigning each page a unique manuscript number, then creating a form in which the catalogue information was matched up with the transcription and the image number.

After preparing the meta data I began transcribing the diary. Given how difficult it is for experienced scholars to read Hart's handwriting, OCR (optical character recognition, or the computer converting the photograph of text into actual text) was not an option. As Denault and Watson explain, even a printed text needs post-processing work to correct OCR; at this stage, there is no widely-available OCR technology to convert manuscripts into texts, let alone a manuscript that a well-trained scholar can barely read.[4] This means that whilst digital images of Hart's writing would be read on the web, scholars would not be able to search the text of the images. We decided that creating a 'tagged' transcription which will be shown side by side with the original diary image would therefore provide the best research tool. That way the transcription could be searched and indexed and the original could be viewed at the same time.

CDDA supplied me with digital images of the diary Volume 31 on compact discs. I transcribed from these images, not from the original diary, on my personal computer. This had three advantages: I could work outside the Special Collections opening hours, I could magnify the image to read more easily, and there was less wear and tear to the original diary. After transcribing the text I wrote a form of XML code ('tagging') into the transcription. Scholars will be able to explore the transcription with free-text searching on the website but we were advised to write the code in order to create indices and to facilitate searching when an "Advanced Search" option is developed. I had no previous experience in XML but I found the tags easy to insert into the text.

---

[3] Edward Lefevour, "A report on the Robert Hart papers at Queen's University Belfast, N.I.," *The Journal of Asian Studies*, Vol. 33, No. 3 (March 1974), p. 437

[4] See Denault and Watson's article on this site, "From page to screen: a case study in the digitization of historical resources."

Conclusions and Future Plans

The largest difficulty we encountered with this collection had nothing to do with technology: it was the very problem that led us towards transcription and digitization in the first place, that of Hart's handwriting. There is no available technology that can substitute for a skilled researcher at this stage. To be able to make sense of the handwritten text, the researcher needs both experience in reading Victorian handwriting and a strong understanding of late-Victorian international politics – probably, at a minimum, a master's degree in history. Researchers would therefore have to be paid at a postgraduate rate. However, the kinds of people who are qualified to do this type of work often find it monotonous. I was pleased to take part in the project and I learned quite a bit, both about digitization and about late-Imperial China –I did a significant amount of research, for example, to ensure that personal names were correctly identified and spelled in my transcription. However, as a new PhD I knew that transcribing and tagging were not fully utilizing my research skills.

We consider the pilot to have been successful: we achieved our aims of digitizing and transcribing one diary and evaluating the amount of time and money needed to digitize and transcribe the rest of the collection. The pilot has revealed the staggering amount of time it would take to transcribe all of the diaries (at the rate I worked, perhaps fifteen years' worth of work), and the corresponding salary costs. The priority now will be to add most of the other seventy-six diaries to the website, with several fully-transcribed diaries from different years to 'train' scholars in interpreting Hart's hand.

We would highly recommend that anyone considering running a major digitization project complete a short pilot first. We can now make well-informed decisions (and write well-informed grants) about proceeding with a full-scale project on this collection. As a nice side benefit, my transcription revealed material that makes us even more confident in the scholarly importance of the Hart collection.

The other piece of advice we can provide is that in the creation of a research resource through digitization, projects need skilled, experienced researchers, not simply scanners and computers. Even without full transcription, most of the work involved in such a project goes into preparing material for scanning, not in scanning it. Digitization is a process through which researchers can make archival material available on the internet; it is not a process through which computers iron out the quirks in archival material. Documents that are difficult to read or understand need to be interpreted by historians before they can be fed to computers. This is an important note of advice to those who think that digitization takes the burden off historians and librarians, and a note of reassurance to those who think that historians and librarians are somehow threatened by digitization. If anything, digitization affirms the importance of traditional historical and archival skills.

Jennifer M. Regan
Belfast, April 2008
jregan01@qub.ac.uk

The new Sir Robert Hart Collection website will be going live in the spring of 2008. Digital images from Hart's photography collection are already available at
http://www.qub.ac.uk/directorates/InformationServices/TheLibrary/BranchesandCollections/SpecialCollections/DigitalImageGallery/