Database review: Imperial China and the West part II, (1865–1905)

Joshua Seufert

Follow this and additional works at: https://scholarsarchive.byu.edu/jeal

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/jeal/vol2022/iss174/7

This Review is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Journal of East Asian Libraries by an authorized editor of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.
Database Review

Imperial China and the West part II, (1865–1905)

Joshua Seufert
Princeton University

The publisher Gale recently published the online database Imperial China and the West part II, (1865–1905) (in the following: Part II) and submitted it for review to the Journal of East Asian Libraries. This database is the second part of the earlier Imperial China and the West part I, (1815–1881) (in the following: Part I). Both databases are available as standalone parts of an ongoing series, distributed under the title China and the Modern World. The series consists of mostly English language primary sources on Chinese relations with Western countries from the early 19th to the mid-20th century, with a strong focus on British-Chinese interaction. Previous parts of the series have covered records of the Maritime Customs Service of China, the British Colonial Office (mostly on Hong Kong), the British India Office, as well as documents from various departments of the East India Company. Other databases in the series have focused on early missionary periodicals with a sinological interest.

Source of the included material and scope of the database

The database at hand gives access to 19th-century files of the UK Foreign Office relating to China, in particular the archival series FO 17 (Foreign Office: Political and Other Departments: General Correspondence before 1906, China) held at the UK National Archives at Kew. The files have been digitalized and made full-text searchable using either classic optical character recognition (OCR) or in the case of hand-written documents, handwritten text recognition (HTR).

The FO 17 collection at the UK National Archives consists of 1777 files.1 Of these, 1771 files were digitized. Part I covers the files FO 17/1 to FO 17/872, Part II includes FO17/873 to FO 17/1768. Although the information provided by the publisher gives the impression that the whole FO 17 collection was digitized, six files FO 17/1769-1 to FO 17/1769-6, are not part of the databases.

The publisher has, however, added in Part II the correspondence between British representatives in China and law officers in London dating from 1833–1876 that were originally collected in seven files of FO 83 (Foreign and Commonwealth Office and predecessors: Political and Other Departments: General Correspondence before 1906, Great

---

1 The numbering for FO 17 runs from 1 to 1769. FO 17/545 was never used, FO 17/76 has been split into two files A and B, FO 17/469 was split in four files A, B, C, and D. FO 1769 was split into six files, FO 17/1769-1 to 1769-6. See, UK National Archives: FO 17 Foreign Office: Political and Other Departments: General Correspondence before 1906, https://discovery.nationalarchives.gov.uk/browse/r/r/C7337 (accessed 26.01.2022).
Britain and General. Cases: Legal. Law officers' reports).\textsuperscript{2} These files have, similar to the main corpus, been processed with OCR and HTR technology and are full-text searchable.

Various registers and indices have also been added to allow researchers to explore the archive in a more traditional way. The finding aids that were originally developed and used at the Foreign Office provide an alternative way of browsing the collection. Adding these materials to the database is a very commendable step. In detail they consist of the following:

The main index of the collection for material prior to 1890 is the register FO 605 (Foreign Office: Registers (Library Series) and Indexes of General Correspondence).\textsuperscript{3} The volumes of this register provide detailed itemized information on the contents of the general correspondence found in FO 17 for the years 1815–1890. Each item in the register contains important details such as volume number, despatch number, date, author and/or recipient, as well as descriptive notes. The publisher has manually transcribed the content of FO 605 and used this data to create in-depth descriptions for the individual documents included in each of the FO 17 files. This allows the user to search through the collection on the document level. The main indexes for the FO 17 series past 1890 are individual volumes of the FO 566 register (Foreign Office: Registers of General Correspondence),\textsuperscript{4} together with the relevant volumes of the indices in FO 738 (Foreign Office: Indexes to General Correspondence, Photographic Copies)\textsuperscript{5} and FO 804 (Foreign Office: Indexes to General Correspondence).\textsuperscript{6} None of these indices above were OCR-processed, so they are only available for visual browsing and cannot be searched through in the database.

One problem is that the FO 605 register in its original form is only available to the user if the library subscribes to Part I, whereas all of the other registers and indexes, as well as the material from FO 83, are only available to the user if the library subscribes to Part II. The user will not see these respective sections when browsing the collection and might therefore not even be aware of them when the library only subscribes to one of the parts.

Nature of the included material
For the most part, the files of the FO 17 series run into several hundred pages and are a mixture of printed and handwritten documents. Included are despatches, letters, memorandums, and notes sent between the British Foreign Office and mostly British institutions in China (e.g. British legation in Peking, British consulates). Thus, the documents detail British foreign policy, British economic interests, and British-Chinese diplomatic exchanges. Beyond letters from British officials, correspondence by Qing officials such as Guo Songtao 郭嵩焘, the first Chinese diplomatic envoy to the UK, can also be found.

As can be expected, the material covers all the major events which have shaped the Sino-British relations up to the early 20\textsuperscript{th} century: The Opium Wars, the Taiping Rebellion, the Unequal Treaties, the Boxer Rebellion of 1900, and the Wuxu Reform. It documents everyday events such as official appointments, reports, title deeds, legal disputes, and information

\textsuperscript{2} In particular the files FO 83/2247 to FO 83/2253.
\textsuperscript{3} In particular the files FO 605/44 to FO 605/51.
\textsuperscript{4} In particular the files FO 566/199, FO 566/580, and FO 566/653 to FO 566/667.
\textsuperscript{5} In particular the file FO 738/11.
\textsuperscript{6} In particular the file FO 804/16.
about births, marriages, and deaths. It is commendable that the whole range of documents of the FO 17 series has been scanned. This makes the collection a treasure-trove not only for political historians but also for researchers in economic, social and cultural studies. Some of the files focus on specific topics, for example:

The Coolie trade and Chinese emigration (FO 17/873-891), Piracy in China Seas (FO 17/802-807, 1110-1113), Anti-Missionary Riots at Foochow (FO 17/853-855), matters of Student Interpreters (FO 17/563, 617-620, 641, 1229-1230, 1301, and 1684), material on Telegraphs in China (FO 17/1007-1011, 1097, and 1189), Anti-Foreign Riots in China (FO 17/1126-1129, 1146-1149, 1171-1172, 1260-1264, 1292, and 1368), the Rendition of Criminals (FO 17/613-616, 799, 1077-1081, and 1591), British and Foreign Concessions in the Treaty Ports (FO 17/1153, 1297-1298, 1461-1465, 1593, and 1712-1713), Affairs of Tibet (FO 17/1745-1756), the railways (FO 17/1757-1758, 1761-1764), and the Chinese Engineering and Mining Company (17 /1759-1760), to just name some examples. Furthermore, the collection also includes 48 files each on Korea (Corea) and Burma (Burmah).

Most of the files however are named after major British office holders or according to the corresponding institutions (e.g., a specific consulate) and do not specify the topics of the material covered within.7

As with the other database titles in the China and the Modern World series, the publisher has focused on English language documents, leaving out six files of Chinese language items, namely the files FO 17/1769-1 to FO 17/1769-6 which have not been included in the database. This omission is rather unfortunate as these files might contain vital historical documents, among others, correspondence received from Chinese officials during the First Opium War. Specifically, the section collects Chinese-language correspondence from the early stages of the First Opium War, when British expeditionary force arrived in China under Commodore Sir James John Gordon Bremer.8

A rare find within the collection is a register of coolie emigration. This register stretches from page 52 to page 211 of the very first volume of FO 605/44 and is entitled Coolie Emigration (1865-1881). It provides much data about Chinese emigrant labor. It is a great pity that this information has not been HTR-processed and thus is neither searchable nor easily exportable from the database. This data could serve as a basis for research into the question of how coolie labor has substituted other forms of forced, unpaid, or underpaid labor in the UK and the British colonies.

---


8 The documents in question date from July 12, 1840 to October 14, 1840. See The National Archives’ catalogue at https://discovery.nationalarchives.gov.uk/browse/r/r/C4236681 (accessed January 26, 2022).
Structure, text corpus and OCR/HTR-processing

The database follows the structure of the FO 17 series. Each archival file has been processed as a single entry within the database. Since each of the FO 17 files generally consist of a large number of physical documents, these documents are not individually cataloged but are marked-up thorough notes in the larger archival files. When reading through a certain file, these notes are a great help. For the database as a whole, however, this approach also creates a lot of problems. Structurally, the database is not able to give access to these notes in the same way as it would for individually indexed documents. For the unaware user, this structure hides much of the content and makes individual documents hard to discover.

The FO 17 and FO 83 files have been digitalized and made full-text via OCR/HTR-processing to extract the text of the files and create a digital text corpus. In the past, many publishers have been actively trying to hide the “messiness” of the text corpus of their databases. Only in recent years have companies started to lay open the underlying data that they have extracted via technologies such as OCR and HTR in order to make these available for further analysis.

In this aspect, Gale has to be applauded for making the extracted text visible to the interested researcher. For each of the scanned FO 17 files, the database additionally displays an OCR confidence value to describe the expected validity of the recognized text. This value, however, seems to be only calculated on the level of a whole file rather than on the level of each individual document or even each single page. With each file consisting of hundreds of pages of different printed and handwritten documents the calculated value seems to be of little use as the OCR reliability will vary considerably for each document included in a single file. It seems that the publisher has not tried to standardize or post-process any part of the generated text corpus. Searches for obvious mistakes like “Feking,” “Cninese,” or “opiun” produce a large number of results. Important place names such as Foochow appear in many more or less obvious mistakes in the database (e.g. Foochon, Foomch, Fochon, Fochow, Fooehow, Fouchow). In most cases, these errors are due to the limitations of the text recognition process. They could, at least in part, have been addressed by creating and providing access to a browsable word index. Such an index could help to identify errors and improve the text corpus step-by-step. The publisher seems to have opted to only tinker with the larger problem by providing a fuzzy search option called “Allow variations.”

Because of these problems, it is impossible to claim that something that cannot be retrieved from the database via the full-text search is not actually included in the FO 17 archive. Similarly, one cannot safely assume that a search for a specific term will produce all of the occurrences of that term in the FO 17 files even with the fuzzy search option activated.

Spot checks on various FO 17 files give the impression that the images used for OCR/HTR-processing are identical to the ones available on the website of the UK National Archives (for more on this, see below). They only seem to have been cropped. No other pre-processing

---

9 This might be very different for the other databases of the China and the Modern World series if in those cases individual documents and not whole files are indexed separately.
10 An info button tells the user that the option it is there to “[r]etrieve imperfect matches to accommodate spelling variations or approximate spellings sometimes found in historical documents” and fails to clearly mention the limitations of the OCR/HTR-processing.
such as adjustment of brightness and contrast, de-skewing, segmentation, or zoning seems to have taken place, as the OCR/HTR-processed text is not structured into main text and annotations. The recognized text rather seems to just follow the general reading direction from top-left to bottom-right without taking into consideration the angle of any interjecting annotations.

It would have been very desirable for Gale to employ modern technologies such as Convolutional Neural Networks (CNN) and Convolutional Encoder-Decoder (CED) architectures in order to segment printed and handwritten text as well as handwritten annotations before any OCR/HTR-processing was undertaken. This would have certainly resulted in a better-structured source and therefore a more reliable text corpus. Such improvements might however still be possible in the future.

**Usability and tools**

The database comes with a variety of additional tools and resources to support researchers using the archive. Under “Research Tools” the database gives access to a number of shorter essays on various parts of the collection of the China and the Modern World series, including several articles on the relevance of the FO 17 series for modern Chinese history by scholars such as Hans van der Ven, Stephen Platt, and Eiichi Motono.

The FO 17 file numbers can be browsed and filtered by keywords. Both a simple search and an advanced search option are available. The advanced search fulfills every possible need and allows cross-searching all of the China and the Modern World databases (as far as these have been acquired) and filtering by document type. The date range selector for example provides an option to include or exclude documents with no known publication date, a feature that is missing in many similar databases.

---

The Help overview is not easily findable placed in the footer of the page. The more prominently placed question mark in the main menu only opens a general database description.

The main interface of the database is the document viewer screen (see Figure 1). On the left the user is offered a variety of tabs depending on the specific nature of the opened file. In the Explore tab the user can search within the opened file and access individual documents via a note system. The image viewer and the plain text viewer (OCR/HTR-text) can be switched on and off. The plain text viewer shows the underlying text corpus for the active document page and the OCR’d text can be copied directly from the interface into other applications.

For each file, a citation tab provides a suggested citation format and an individual unique and stable ID (see Figure 2). If the opened file has illustrations or other items of special interest such as charts, tables, drawings, maps, or seals an additional tab for illustrated works gives access to those parts of the file directly (see Figure 3). Regrettably, there seems to be no option to browse all illustrated works throughout the database, a feature that could be of tremendous help. A search for all illustrated works in the collection is possible via the advanced search function, but only produces a list of the document folders in which the illustrated works are to be found and does not show any thumbnails of the illustrations themselves.

![Figure 2. Citation tab.](image)

In the document viewer, the reader can browse the scanned images and can jump between documents (via indexed notes) in the same file. In the image viewer, the user can adjust hue and contrast as well as invert and resize the image in predefined steps. The viewer in its current state does not allow common functions such as stepless zooming, rotating of images, or resizing of the viewer window which often becomes necessary. Using the database on a laptop, therefore, turns out to be a challenge. For proper use of the database, a widescreen monitor is necessary.
The database supports downloading of documents of up to 200 pages as PDF files. Larger documents can be downloaded in several sections. The PDFs downloaded in this manner do not allow full-text searching. The recognized text of the files is available for download as one single text file per document. Selection of pages for this text-only download is not possible.

There are also two digital humanities tools as part of the database: a term frequency counter and a topic finder. Both tools are linked at the bottom of the China and the Modern World overview page and because of their positioning easily can be overlooked. Additionally, they are shown in any search result screen. The term frequency tool allows the user to count the number of occasions of individual terms and their relative popularity within all documents. The topic finder is supposed to help the user explore the topics covered in the various documents. It however has considerable limitations, as the tool only “takes the titles, subjects, and approximately the first 100 words from a subset of your top results and feeds them into an algorithm.” This works well with individual short documents. In the case of FO 17 and the way in which the files were made available (see Structure, text corpus and OCR/HTR-processing) the tool is almost useless as each of the files includes several hundred pages with very different content. It might therefore be better to use the export function of the database to run topic modeling in a more powerful standalone DH system.

Alternatives and conclusion
In conclusion, the FO 17 provides a very rich cache of material on Sino-British interaction in the 19th century. Despite the limitations of the OCR/HTR reliability, the option to search through this huge archival collection as a whole has the potential to make much previously hidden material accessible and discoverable for the first time, thus supporting many new areas of research.

Most of the FO 17 documents can also be requested free-of-charge as PDF documents directly from the UK National Archives at Kew after signing up for a free online account. The
National Archives allow a maximum number of 100 requests in a 30-day period.\textsuperscript{12} Similar to the PDFs downloaded from the database, the PDF files from the National Archives do not allow full-text searching within the PDFs as they have not been OCR/HTR-processed.

In the end the decision to acquire access to the databases will depend on the expected local use. If the goal is to use original archival material from the FO 17 series in a class and let students take their first steps in archival research, using the option of requesting free PDFs would not only be free but could also be a very instructive part of an archival history class.

If, however, British-Chinese relations of the 19\textsuperscript{th} century and surrounding topics are a research focus of the local faculty and students, then the acquisition of this database should be seriously considered. Given that the associated finding aids are split between the two parts, it would certainly be sensible to acquire access to both parts rather than just one.