Transforming data silos into knowledge: Early Chinese Periodicals Online (ECPO)

Matthias Arnold[1] and Lena Hessel[2]

[1]Heidelberg Research Architecture - Heidelberg Centre for Transcultural Studies, Universität Heidelberg;

[2]Institute of Chinese Studies, Universität Heidelberg

Abstract

This paper introduces the project "Early Chinese Periodicals Online (ECPO)" [1]. ECPO joins several important digital collections of the early Chinese press and puts them into a single overarching framework. In a first phase, several databases on early women's periodicals and entertainment publishing were created: "Chinese Women's Magazines in the Late Qing and Early Republican Period" (*WoMag*), "Chinese Entertainment Newspapers" (*Xiaobao*), and databases hosted at the Academia Sinica in Taiwan. These systems approach the material in two ways: in the *intensive approach* we record all articles, images, advertisements, and related agents and assign them to a complete set of scanned pages, while in the *extensive approach* we record the main characteristic features of publications.

ECPO has begun to join these various materials in a second, ongoing phase of the project. Today, ECPO provides open access to 267 publications comprising over 280.000 pages of print. A key aspect is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages. For 138 publications we also provide descriptions of individual items in Chinese with Pinyin transcription. These records also contain genre and column information, basic content analysis, as well as names and roles of agents associated with an item.

Our new cross-database agent service allows us to manage the approximately 47.000 names recorded in *WoMag* and ECPO: we a) merge identical names across databases, b) identify agents and assign names to them, and c) link agent records to authority data (GND, VIAF, Wikidata, Baidu, DBpedia). Besides creating a curated list of agents occurring in the publications, we also aim to add missing persons to authority files like the GND.

One crucial aspect of ECPO is full text capability. Unfortunately, OCR software cannot be used out-of-the-box, for a number of reasons: document analysis fails to recognize complex newspaper layout, character recognition fails when it faces emphasis marks next to characters, and recognized passages have to be grouped in the right semantic order.

Das hier beschriebene Paper ist in der Open Access-Plattform der Universität Heidelberg heiDOK unter der DOI `https://doi.org/10.11588/heidok.00027325` veröffentlicht.

The paper will discuss approaches to further exploring and analyzing the knowledge hidden in these publications, together with efforts to open the collection's data for reuse. We demonstrate workflows in the Agents service and cross-database record curation. We also present results from a crowdsourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, we introduce first ideas to create a module for encoding text in TEI and relate it to the database.

1. Introduction

Research data can occur in various forms and formats. In the humanities, the systematic collection of data driven by a set of distinct research questions very often is a challenge on its own: The research data collection becomes a major part of the research data output. This is also the case for the project we introduce in this paper:[1]"Early Chinese Periodicals Online (ECPO)"[1].

The project aims at a systematic examination of the Chinese periodical press during the first four decades of the 20th century. Chinese periodicals of this era remain understudied, even though they dominated the contemporary print market and provide access to what Raymond Williams has called "actual culture".[2] They present researchers with a number of challenges: They are physically dispersed and often poorly preserved, voluminous, multi-generic, and intellectually demanding. We approached these challenges in two ways. Firstly, we formed a multidisciplinary research team. Through a number of international interdisciplinary projects the research team engaged in over the past decade[3], we have developed a new methodology for approaching materials from the Chinese popular press.[4] Secondly, we created a set of unique digital resources developed by the Heidelberg Research Architecture (HRA) and located at the Heidelberg Centre for Transcultural Studies (HCTS). These resources formed the basis for conceptualizing and implementing a database prototype, ECPO -Early Chinese Periodicals Online.

ECPO is distinguished from other existing databases of Chinese periodicals in that it not only provides image scans, but also preserves materials often excluded in reprint, microfilm or digital editions, such as advertising inserts and illustrations. Our workflow is even designed to establish a definitive page sequence for citations and references. In addition, it incorporates a sophisticated body of metadata in both English and Chinese,

[1] This paper was made possible with funding from the Heidelberg Centre for Transcultural Studies and The Center of East Asian Studies at Heidelberg University. The authors wish to thank Prof. Barbara Mittler (Heidelberg University) and Prof. Joan Judge (York University) for their continuous support and invaluable input.

[2] Williams (1961) p.70.

[3] "A New Approach to the Popular Press in China: Gender and Cultural Production, 1904-1937" funded by the Canadian SSHRC and the German Humboldt-Foundation (TransCoop-Program); "The Stuff Stars are made of: International Politics, Mass Media and the Rise of dan Actors in the Republican Era (1910s–1930s)" funded by the Cluster of Excellence "Asia and Europe", the Krupp Foundation and the Institute of Chinese Studies, University of Heidelberg; "Building Early Chinese Periodicals Online (ECPO)—Expanding and Refining a Database Prototype for Historical Media Studies" funded by the Chiang Ching-kuo Foundation, Republic of China.

[4] See, for example, Hockx, et al (2018), and Sung, et al (2014).

including keywords and biographical information on agents—editors, authors and other individuals, groups, institutions, or corporations mentioned in articles and represented in illustrations and advertisements. We strongly believe in the importance of this manual editing of individual records for each item—article, image, advertisement, because this data analysis still offers information beyond what one might retrieve through a full text search, especially since much of the corpus is not yet available in full text for text mining. In addition, the bilingual analytical data makes these important sources available to non-readers of Chinese.

ECPO combines several important digital collections of the early Chinese press into a single overarching framework. To date, ECPO has focused on a body of rich but heretofore undervalued materials—women's and entertainment magazines. It is open to further additions: Currently, we are adding a selection of literary, art and women's magazines, e.g. Banyue半月 (The Half Moon Journal), *Tianyi* 天義(Tien yee), as well as western-language publications produced in China, e.g. The Canton Press.

The core of ECPO consists of several databases on early women's periodicals and entertainment publishing: "Chinese Women's Magazines in the Late Qing and Early Republican Period" (*WoMag*), "Chinese Entertainment Newspapers"(*Xiaobao*), and various databases hosted through the Academia Sinica in Taiwan.

WoMag [2] focuses on four influential women's magazines published between 1904 and 1937. It records all articles, images, advertisements, and related agents and assigns them to a complete set of scanned pages. This database is the model for what we have called the *intensive approach* within our database structure. *Xiaobao* [3] provides basic publication data and characteristics of the contents of some 22 entertainment newspapers (小報, *xiaobao)* from the late Qing and Republican periods. This database is the model for what we have called the *extensive approach* in our database structure.

The Academia Sinica, and in particular its Institute of Modern History, have in recent years digitized large parts of their collections of periodicals and built a database for the *Funü zazhi* 婦女雜誌 (The Lady's Journal) [4]. All these resources follow the model for what we call the *extensive approach*.

In a second, ongoing phase of the project, ECPO has begun to integrate these materials.[5] It is our aim to make the various individual digital collections accessible through a single search interface. We continue to acquire new publications to broaden the project's material base, and thus, to enrich and expand its potential for data analysis. We strive to open up our system to share data with the community, enable data re-use, and are adapting to principles of FAIR use.

As it currently stands, ECPO provides the research community with open access to more than 280 publications, mostly from the Early Republican period, comprising over 280.000 printed pages. A key and unique aspect of the project is to make entire issues available, front-to-back, including illustrations, advertisements, and even blank pages.

[5] During this phase, the project received a grant from the Chiang Ching-kuo Foundation 2012-2015 for a collaboration between Heidelberg University and the Academia Sinica. Since 2015, the Institute of Chinese Studies and the Heidelberg Centre for Transcultural Studies (HCTS) at Heidelberg University have continued to support the project. All technical development is coordinated through the Heidelberg ResearchArchitecture (HRA).

For approximately half of the publications, we provide descriptions and bibliographic metadata of individual items (articles, images, advertisements) in Chinese with Pinyin transcription.[6] These annotated records also contain genre and column information, basic content analysis through bilingual keywords, as well as the names and roles of agents associated with an item, including "mentioned in article" or "depicted in image". Overall, the project followed the five guidelines for the digital archiving of periodicals as formulated by Latham and Scholes.[7]

To further increase the impact of ECPO and in order to sustain the information, ECPO has begun to develop dynamic data services to provide data for re-use as open data. We implemented a MODS XML API [5] to provide bibliographic information of all annotated items in the database. In addition, we installed an IIIF image service for all page scans, and are developing an API to output each publication's detailed publishing information. Data sets will be published in heiDATA,[8] the Heidelberg research data repository.

2. The Agents Service

Before *WoMag* was linked to ECPO, both databases recorded personal names separately. At the same time, names were recorded without distinguishing between names and actual agents. We subsequently have begun to build a cross-database agent service, to create a single, central resource that hosts all agent-related data and can be referenced from other databases. Within this service, we combine the agent data from both *WoMag* and ECPO, keep the database open for the addition of new sets of data, and expand and refine the records of personal names into entries on agents, i.e. the persons behind a name, with biographical information, notes, and links to external authorities.

Our records of personal names were created manually by research assistants as part of the workflow to record individual articles, images, and advertisements. As with all manual input, some oversights and errors occurred in this process. Names were accidentally recorded more than once, name variants were not recognized as belonging to the same person, both of which led to the creation of more than one entry for a single individual. On the other hand, identical names were sometimes not recognized as belonging to different persons, resulting in only one entry for several persons. And sometimes, names were simply overlooked. In the process of joining ECPO and *WoMag* name records, we created a lot of duplicates, the same was the case when we ingested new data from other projects. Since all these issues have to be solved manually, we have implemented a number of tailored functions into the agent service to simplify these tasks.

The agent service allows us to: a) distinguish between "names" and "agents", b) assign names to agents, c) merge identical names (that refer to identical agents) across databases, and d) link agent records to authority data (GND, VIAF, Wikidata). At the moment

[6] Currently, (March 2019) 40.936 issues in total. Number of annotated items: 46.931 articles, 20.532 images, 18.639 advertisements.

[7] Latham and Scholes (2006), p. 524.

[8] heiDATA is an institutional repository for research data of Heidelberg University, see `https://heidata.uni-heidelberg.de/`. The ECPO data set will be available at `https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/Z3J0DV`.

(March 2019) we have 1220 Agents with references to VIAF, 985 with references to GND, 848 with references to Wikidata.[9] Recently, we started to also include Baidu Baike and DBpedia, and both authorities are referred to by 9 agents.

Before we can implement distinctions between "names" and "agents", i.e. separate names into different agent entries, merge them into one, or eliminate duplicates, we have to carry out research. In order to understand which person is meant by a particular name or visual representation in an item, we have to understand the item and its context; we have to find persons who share that name or that visual representation; and we have to know something of their biographical background. Only then can we identify an item as referring to a specific person and assign a respective agent entry to that item. Through this research process, we have often come across heretofore little known persons, events and phenomena.

One such phenomenon, which illustrates the effort we put into editing each agent entry, is the world tour of the Islington Corinthians Football Club.[10] Between 1937 and 1938, this London-based amateur football club sent a group of its players on a tour around the world, to engage in friendly matches in Europe, South Asia, East Asia and Northern America. In April 1938, their itinerary led them to China where they made headlines in a local entertainment newspaper.[11] In this instance, our task started with those newspaper articles and the question, which London-based football club in the 1930s was represented through the Chinese name *yi shi lin dun* 衣士林頓. Having found out that it likely refers to the Islington Corinthians F.C., we began to research which club members were sent on tour in order to create respective agent entries and match them to the Chinese version of their names; then, we looked into those club members' biographies in order to identify the people and organizations named in the Chinese newspaper articles. Finally, there were the names of the quasi-obscure foreign players and referees of the Shanghai Football Association against whom the Islington Corinthians played a match on April 3, 1938. Luckily in this case, we could rely on several outside digital resources,[12] such as a short Wikipedia entry for the Islington Corinthians F.C., contemporary newspaper articles, present-day web articles, and web sites of hobby historians and collectors of sports memorabilia. Our aim is to connect database users to these kinds of resources. Therefore, each agent entry is supposed to include assignments to ECPO items, short biographical information and references to the sources we used.

[9] VIAF: The Virtual International Authority File combines multiple national and international name authority files into a single OCLC-hosted name authority service; `https://viaf.org/`. GND: Gemeinsame Normdatei, or Integrated Authority File, is hosted by the German National Library; `https://www.dnb.de/EN/Standardisierung/GND/gnd_node.html` Wikidata is a collaboratively edited knowledge base hosted by the Wikimedia Foundation; `https://www.wikidata.org`. Baidu Baike 百度百科 is a Chinese-language encyclopedia by the Chinese search engine Baidu; http://baike.baidu.com/. DBpedia aims to allow users to semantically query structured information from Wikipedia resources; `http://dbpedia.org/`

[10] "Islington Corinthians F.C.", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=9814.`

[11] For example, "Ba qian qiu mi da wei sao xing: Yishilindun zu qiu dui zuo jing quan jun fu mei" 八千球迷大爲掃興：依士林頓足球隊昨竟全軍覆沒, Jing bao 晶報, April 4, 1938, p. 2, available at ECPO, `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=3841&ispage=2&itemid=5325&itype=2.`

[12] See above, "Islington Corinthians F.C.".

To provide this information in a sustainable way, we also offer permanent links to all of our digital sources as archived in the project OpenDACHS [6].

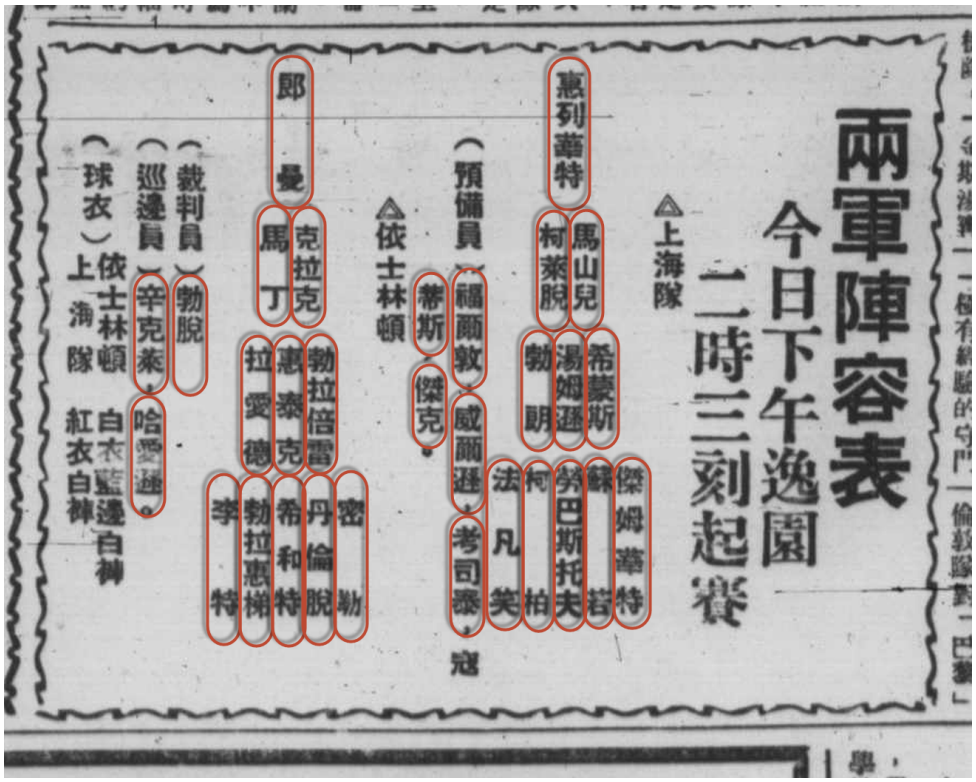


Figure 1.: The players' line-up for the April 3 match between the Islington Corinthians and an all-stars team of the Shanghai Football Association. All names are only printed in their Chinese versions. Cf. *Jing bao* 晶報(The Crystal), April 3, 1938, special issue, page 1, available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=4873` (emphasis markers by the authors).

Besides creating a curated list of agents occurring in ECPO items and connecting ECPO users to external resources, we also aim to open up and share our data. We plan to add missing persons to Authority files, using the German National Authority file (GND). The missing persons we plan to add range from people of minor, local importance, like the players of the Shanghai Football Association mentioned above; to people who played more prominent roles in world history, for example Sir Herbert Phillips,[13] Consul-General of the United Kingdom in Shanghai during the late 1930s, or Arminio de Mello Franco,[14] Brazilian ambassador to China in 1927-1928, neither of whom currently has a VIAF or GND record, or a Wikidata page. It is certainly neither meaningful nor feasible to add every individual from ECPO resources to the international authority.

[13] "Phillips, Herbert, Sir", ECPO Agents, last accessed on May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=9928`.

[14] "Franco, Arminio de Mello", ECPO Agents, last accessed on May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=33930`.

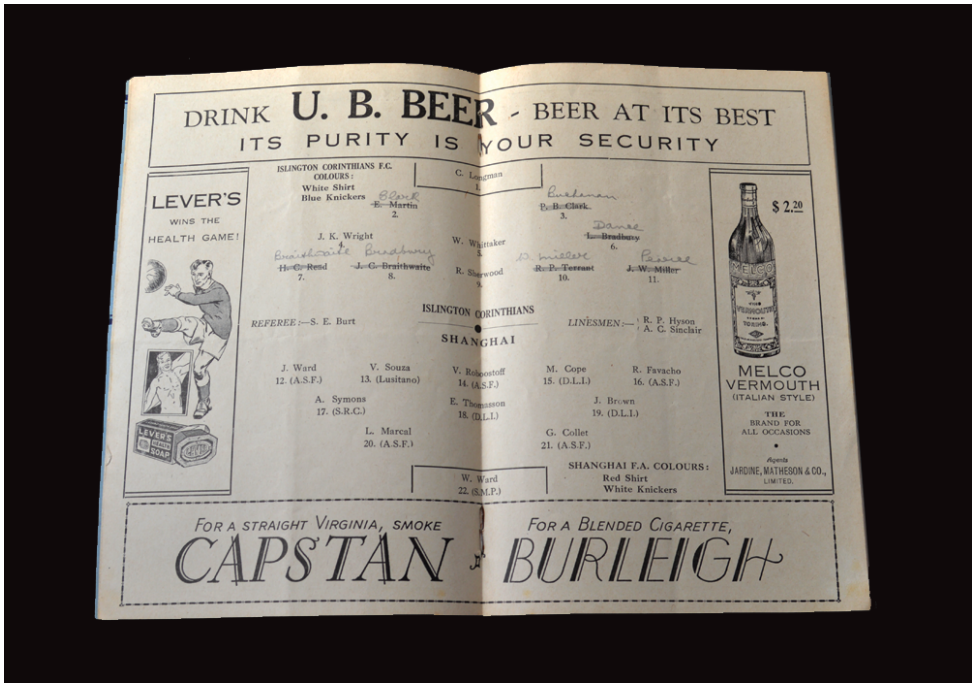
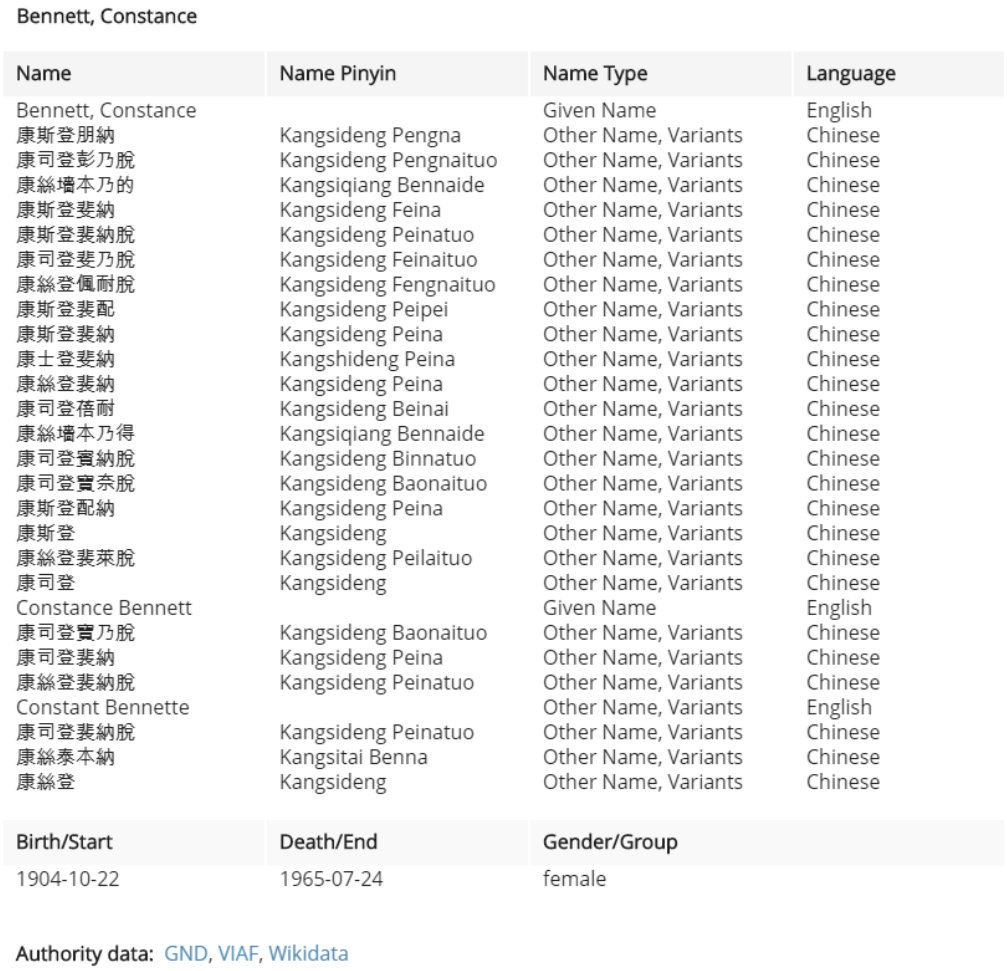


Figure 2.: The players' line-up for the April 3 match between the Islington Corinthians and an all-stars team of the Shanghai Football Association, as it was printed in the Shanghai Football Association's booklet for the game; all the participants' names are printed in Latin script. Photos of the booklet are hosted on the website 10footballs.com. They are part of an impressive private collection of football memorabilia, with a focus on the Islington Corinthians' world tour ("Islington Corinthians - Shanghai [team line up]", in *1938 - Islington Corinthians vs Shanghai: Sunday, 3rd April at 2.45 p.m. Canidrome*, edited by The Shanghai Football Association (Shanghai: The Printing Press, 1938), center pages, accessed February 7, 2019, `http://10footballs.com/wp-content/uploads/2018/10/063.png`).

But we will begin with more prominent figures and those agents that occur more frequently. Besides missing persons, we are preparing to add missing names to Authority files, especially Chinese variants of foreign names. In the Republican period, standardized Chinese renderings of foreign names did not yet exist. As a result, for some agents we have registered over twenty Chinese name variants. Most, if not all of these variants, are usually missing from Authority files such as GND and VIAF. Therefore, we are developing an API to provide our agents' data in machine readable format.

For example, the actress Constance Bennett[15] is mentioned by twenty-five different Chinese names in ECPO, and one variation of her English name.

[15] "Bennett, Constance", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=34635`.



Bennett, Constance

Name	Name Pinyin	Name Type	Language
Bennett, Constance		Given Name	English
康斯登朋納	Kangsideng Pengna	Other Name, Variants	Chinese
康司登彭乃脫	Kangsideng Pengnaituo	Other Name, Variants	Chinese
康絲墻本乃的	Kangsiqiang Bennaide	Other Name, Variants	Chinese
康斯登斐納	Kangsideng Feina	Other Name, Variants	Chinese
康斯登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
康司登斐乃脫	Kangsideng Feinaituo	Other Name, Variants	Chinese
康絲登佩耐脫	Kangsideng Fengnaituo	Other Name, Variants	Chinese
康斯登裴配	Kangsideng Peipei	Other Name, Variants	Chinese
康斯登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康士登斐納	Kangshideng Peina	Other Name, Variants	Chinese
康絲登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康司登蓓耐	Kangsideng Beinai	Other Name, Variants	Chinese
康絲墻本乃得	Kangsiqiang Bennaide	Other Name, Variants	Chinese
康司登賓納脫	Kangsideng Binnatuo	Other Name, Variants	Chinese
康司登寶奈脫	Kangsideng Baonaituo	Other Name, Variants	Chinese
康斯登配納	Kangsideng Peina	Other Name, Variants	Chinese
康斯登	Kangsideng	Other Name, Variants	Chinese
康絲登裴萊脫	Kangsideng Peilaituo	Other Name, Variants	Chinese
康司登	Kangsideng	Other Name, Variants	Chinese
Constance Bennett		Given Name	English
康司登寶乃脫	Kangsideng Baonaituo	Other Name, Variants	Chinese
康司登裴納	Kangsideng Peina	Other Name, Variants	Chinese
康絲登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
Constant Bennette		Other Name, Variants	English
康司登裴納脫	Kangsideng Peinatuo	Other Name, Variants	Chinese
康絲泰本納	Kangsitai Benna	Other Name, Variants	Chinese
康絲登	Kangsideng	Other Name, Variants	Chinese

Birth/Start	Death/End	Gender/Group
1904-10-22	1965-07-24	female

Authority data: GND, VIAF, Wikidata

Figure 3.: Cut-out of the agent entry for Constance Bennett ("Bennett, Constance", ECPO Agents, last accessed May 5, 2019, `https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/agent-information.php?agentid=34635`.)

In GND, she appears only under her English name.[16] In VIAF, she is listed with twenty-two name variants in multiple languages.[17] However, only one of those variants is Chinese. Constance Bennett's example highlights a less obvious contribution the agent service offers to researching foreigners in China. In order to find information in contemporary Chinese sources, the Chinese version of a foreigner's name needs to be known. Constance Bennett's Authority file would provide a wide range of possible Chinese name variants of any other person sharing the name "Bennett" (which is quite a common name), and thereby become a starting point for researching Republican sources. The agents service could act as a resource for Chinese name variants of common English or other foreign names.

[16] "Bennett, Constance", GND, last accessed May 5, 2019, `http://d-nb.info/gnd/129660760`.
[17] VIAF ID: 34718115 (Person), last accessed May 5, 2019, `http://viaf.org/viaf/34718115`.

3. Towards Full Text

ECPO is now also beginning to focus on full text capabilities, with the aim of producing machine readable texts that can then be used for further analysis, like text mining. While some records occasionally feature full text passages in the metadata, for example some advertisements[18], this task is too big to solve manually – we need automated workflows.

However, after a number of first experiments it soon became clear that we cannot use OCR software out-of-the-box, for a number of reasons:

a) In many cases we are working with secondary material in sometimes sub-optimal quality. Images can be blurry, or show noise, stains, scratches, etc.

b) Document analysis fails to recognize the complex and very densely set page layout. This is especially true for newspapers.

c) Character recognition fails when it faces special characters, like emphasis marks next to characters. Typically, titles (e.g. of articles) or texts within illustrations are handwritten or feature special calligraphic styles, which also cannot be recognized.

While the production of quality full text can be a challenge even with western language material[19]—Chinese characters are very complex, and significantly different from Latin-based script systems[20]—there are different ways to approach these problems. One way is double-keying, but this approach is extremely labor- and cost-intensive.[21] Only very few such endeavors have been undertaken even in China.[22] Although OCR has been significantly improved in recent years, it still largely fails with Chinese texts from before the 1970s.[23]

[18] ECPO already contains more than 5.000 advertisement records with full-text data, for example, the advertisement for a publication by Bao Tianxiao 上海春秋第二集出版(Shanghai Chunqiu di er ji chu ban), 晶報 *Jing bao*, volume 1, issue 706, Monday, 1925-01-12, page 1: `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=286&ispage=2&itemid=4754&itype=4`.

[19] The German OCR-D initiative `http://ocr-d.de` was started to coordinate the developments OCR for printed historical texts with a focus on 16th to 19th ct. German language material. It continues the efforts of the EUC project IMPACT `http://www.impact-project.eu/` and can be seen as parallel approach to the current Horizon 2010 project READ `https://eadh.org/projects/read`, which focuses on handwritten text recognition of archival records.

[20] With all language specific variants, the 26 Latin letters form a group of about 500, while there are over 50.000 Chinese characters alone, without taking the many variants into account. In the current version 12 of the Unicode standard a total number of 96190 Ideographic code points, or 87.887 CJK Unified ideographs are defined. [7]

[21] A typical quote for a periodical of the early 20th century is 3 USD per 1000 characters. Adopting these figures to the *Jingbao* newspaper (21 years, about 10.000 double-pages) would result in an estimated quarter million USD.

[22] For example the new edition of late 19th ct. Shanghai Daily Shenbao produced by double-keying specialist Greenapple Changsha. The biggest non-commercial program is the full-text digitization of the complete Buddhist canon by the Dharma Drum Institute in Taiwan, which took more than 12 years to complete and was largely carried out by volunteers from the worldwide Buddhist community.

[23] Commercial double-keying agencies have separate (higher) pricing for material even dating to "before 1990."

Figure 4.: Detail of a digitized microfilm page, showing stains and scratches, as well as different fonts, illustrations, and emphasis characters. *Jing bao* 晶報 (The Crystal), March 6, 1919, page 2 (upper part), available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=649&ispage=2`.

Our tests with OCR software like Abbyy Cloud OCR SDK [8] showed that processing full pages does not work.[24] Even when pages were manually segmented and the segments then processed, the OCR failed if emphasis characters were present. This only changed for segments that were manually optimized: emphasis characters removed and images optimized. There the recognition rate went up to over 60%. Although an error rate of up to 40% is still far from a good recognition result, it still indicates the potential of this approach. Tests by a possible Chinese partner[25] show different options to develop automatic workflows for further processing these segments, detecting and removing emphasis characters, running image optimizations, detecting lines and characters, and recognizing characters.

[24] Only less than 10% of the characters were recognized.

[25] Computational Knowledge Lab, Department of Engineering Science and Ocean Engineering, Taiwan National University, `http://www.cklab.org/`.

Recently, other platforms have started promising projects with processing non-Latin character scripts.[26]

Within ECPO we therefore focused on segmentation. We began a pilot project with Pallas Ludens [9], a local start-up specializing in crowdsourcing solutions. The pilot was performed by a non-Chinese-speaking crowd, who manually identified information blocks on newspaper pages and qualified them with a label. Identification was very good, but as the crowd was unable to read Chinese, they were not able to identify semantic groups, e.g. decide, which segments belong to one article and which to another. The grouping of individual boxes into meaningful semantic units was then done in a second run by a reader of Chinese. This, too, turned out to be quite successful. Unfortunately for us, Pallas Ludens was then bought out by a larger company and had to stop all external co-operation—including the one with us.

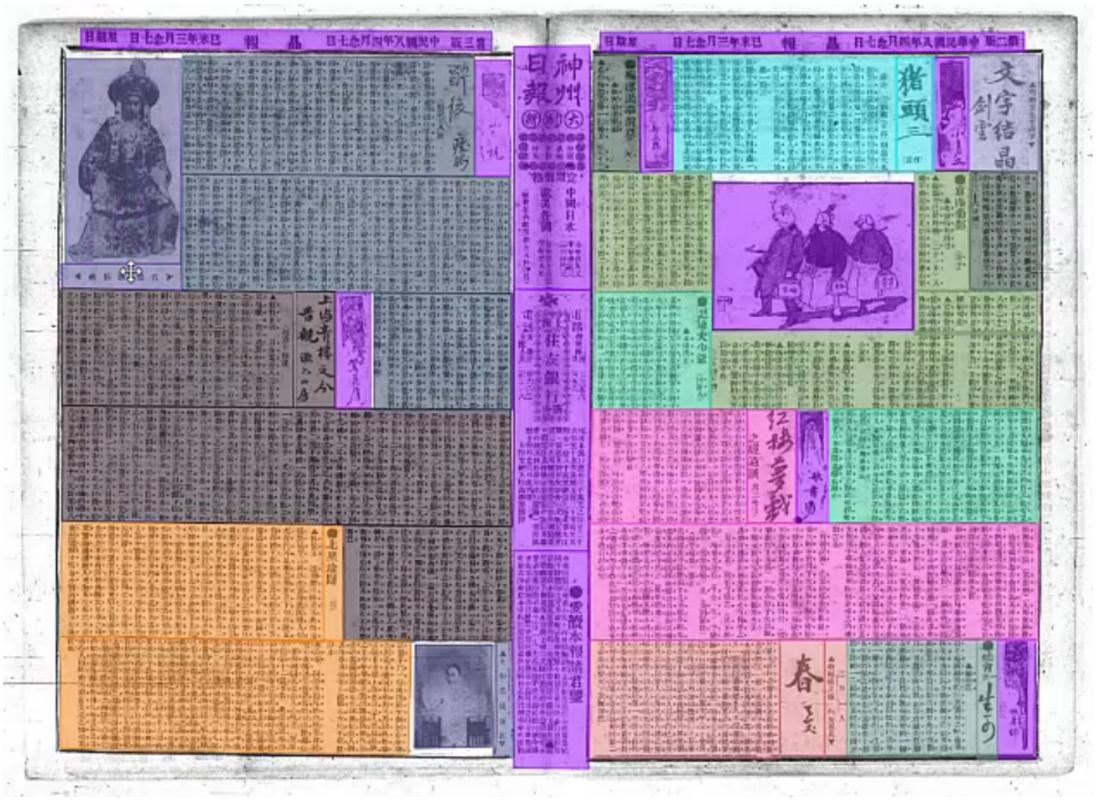


Figure 5.: A page where bounding boxes are grouped into semantic units, e.g. titles and text of articles, image and caption. From *Jing bao* 晶報(The Crystal), April 27, 1919, page 2, available via ECPO at `https://uni-heidelberg.de/ecpo/publications.php?magid=1&isid=9&ispage=2`.

Nevertheless, the outcome of the pilot taught us: crowdsourcing our material is not only possible, but can produce very good results, especially, if participants can read Chinese,

[26] For example, the *Naval Kishore Press – digital* project initiated by Heidelberg's South Asia Institute Library and Heidelberg University Library processed selected Hindi and Sanskrit titles in Devanagari script with Transkribus (`http://transkribus.eu/`). After training with ground truth from 200 pages an error rate of 5.59% was reached. See Merkel-Hilf (2018).

are supervised, and if user interfaces with excellent usability are provided.[27] We have just launched a new tool for annotating and semantic grouping that stores annotations with coordinates and labels in web annotation format in an XML database. The annotated segments can later be served in the database using the coordinates with our IIIF image service, and also be sent to a future OCR workflow. We are currently looking for partners to develop and implement these workflows to automate page segmentation and are in contact with computer vision labs and partners of the OCR-D initiative.

With the advent of basic segmentation and OCR workflows, we will soon have various full text segments. This calls for new solutions, as for example the text segments need to be linked with their respective metadata records, and the search has to be expanded to include unstructured texual data. The same is true when it comes to encoding full issues of magazines with structured markup, e.g. using TEI XML. The reason are a number of Asia-specific features, or perhaps even features of non-Western publications, for which good practice conventions still need to materialize. Within ECPO, we have started creating TEI records for the magazine *Tian yi* 天義(Tien yee).[28]

One major issue is the handling of emphasis characters in the text.[29] In *Tian yi*, for example, up to 6 different emphasis characters occur. Sometimes their use for emphasis is mixed with punctuation marks, in some texts almost every character is emphasized. Although there is a way to encode these characters using the list of "kenten" characters,[30] we eventually decided to not include these marks at all. The main reasons for this are: a) the workload for encoding and distinguishing emphasis characters and punctuation marks is huge, b) the added value of encoding these characters is limited, c) the focus currently rests on the creation of a machine readable text, and this additional level of information can be added at a later time. However, one type of emphasis is encoded by us, that is the font size. Texts can be set in double-size characters - indicating emphasized passages, in "normal" size, and in half-size characters - indicating comments or inserts.

[27] Outcome: *Jingbao* 1919-22 completely segmented with adjusted bounding boxes, for the first April issue (1919) all boxes are clustered in semantic groups.

[28] This amends an endeavor where we are adding research analysis of a project from the late 1990s into ECPO: As part of research focus "Frauenbewegungen - kultureller und sozialer Wandel" funded by the Hessian Ministry of Higher Education, Research and the Arts, Prof. Monika Übelhör (University of Marburg) received a grant for her project "Die Zeitschrift Tianyi (1907-1910) als Plattform für eine grundsätzliche Neubestimmung der Stellung der Frau in der chinesischen Gesellschaft"(The magazine Tianyi (1907-1910) as a platform for a fundemental redefinition of the position of women in Chinese society) in 1996. Project member Gabriele v. Sivers-Sattler, M.A. studied this publication and created a classified and annotated inventory. Cf. Sivers (2001) and Sivers-Sattler (2001). We are currently adopting their research outcome to the database ECPO database structure for data ingest.

[29] The authors are grateful for the valuable suggestions they got in various discussions with the following specialists: Duncan Paterson, Christian Wittern, Marcus Bingenheimer, Wolfgang Meier. Problematic cases and possible solutions are collected in an online document, cf. Arnold (2018).

[30] The list of *kenten* (jap.圏点) or "emphasis dots" comprise about 10 glyphs with their Unicode code points typically used for emphasis. Cf. the the W3C recommendations in Etemad and Ishii (2013), and the overview in the Japanese Wikipedia `https://ja.wikipedia.org/wiki/` 圏点 (page last changed 2019-03-30). Since CSS3 is still only a recommendation, browsers may render elements differently; some may not even be supported. For more information see Andrew (2017).

This is a typical feature in Asian texts, and we use CSS attributes with TEI elements `<emph> @type` and `<hi> @rend` to encode them.

Other editorial decisions we have to make are related to spaces. These can be indentations of full paragraphs, or single spaces. Indentation of longer passages is a common feature in magazines to indicate, for example different levels of argumentation, or longer commentaries. Individual spaces can also be used to emphasize the following character(s), but in some cases there may be technical reasons. While that is still a task for future research, there are multiple ways to encode these spaces. One is to use the `<space>` element with `@unit` and `@extent`, e.g.`<space unit="char" extent="1"/>`. Another suggestion is to use a U+3000 "Ideographic Space". At the moment we use the first method, but we are open for further discussion.

4. Summary

In this paper we discussed a number of approaches to further exploring and analyzing the contents of collected publications, together with efforts to open the collection's data for re-use. We demonstrated workflows in the Agents service, which assists in curating agent records across databases and forms the basis for enhancing authority records. We also presented results from a crowd-sourced approach to newspaper segmentation to generate segments that can easier be OCRed. In addition, we introduced our efforts to develop a proper markup for encoding full Chinese periodicals in TEI XML using *Tian yi* as example.

ECPO started as a typical information-silo: a sophisticated data structure, but no "outside" connections. While the collection of these materials itself is a huge contribution in providing the community with research data, we are continuously enhancing the metadata. Content analysis through keywords for each publication is amended by information about the publishing history, which is as comprehensive as possible. With the separation of meta-/data from the end-user interface we are able to start providing data sets in machine readable formats. We have also started to adopt FAIR principles: we are implementing DOI records for each publication, and connect our authority data to international authority files. We are publishing our resources and metadata Open Access, including metadata on publication, issue and individual item levels. We provide access to data sets through API's (e.g. bibliographic data in MODS format) and are preparing the publication of IIIF manifests. And we are publishing data sets on the Heidelberg research data platform heiDATA.

We still are expanding our data corpus: a project funded by the HCTS adds more than 100.000 pages from Foreign Press published in China. In co-operation with Erlangen University we are expanding our agent service with features like relations, and location services. We are also able to use ECPO to store and make available output from former research projects. In addition, ECPO will slowly grow into a data platform for other material from the CATS library. This will not only allow users to access the data, but also to further enhance and share it with the research community.

With its rich material base, ECPO is growing in different directions. A small project cannot do this alone. We are actively involved in initiatives like the DH-d Working Group

Newspapers and Journals, the non-Latn scripts interest group, the TEI East Asia SIG, and in close contact with the FID Asien (Cross Asia), as well as with members of OCR-d and READ/Transkribus. Only in collaboration with these groups and individuals can we successfully work with our own material and further develop ECPO.

Literature

Andrew (2017): Andrew, Rachel. „Christmas Gifts for Your Future Self: Testing the Web Platform". *24 Ways* (blog), 10. Dezember 2017. https://24ways.org/2017/testing-the-web-platform/.

Arnold (2018): Arnold, Matthias. "Tianyi bao - Questions related to TEI". 2018. https://docs.google.com/document/d/1xsE4kavEe-LdL7JKwDpaXtGZQbXLsRLtzSJ8sM4MTbw/edit?usp=sharing

Etemad and Ishii (2013): Elika J. Etemad and Koji Ishii (eds). "CSS Text Decoration Module Level 3", W3C Candidate Recommendation 1 August 2013, https://www.w3.org/TR/2013/CR-css-text-decor-3-20130801/#text-emphasis-style-property

Hockx, et al (2018): Hockx, Michel, Joan Judge, and Barbara Mittler, eds. *Women and the Periodical Press in China's Long Twentieth Century: A Space of Their Own?* Cambridge: Cambridge University Press, 2018.

Merkel-Hilf (2018): Merkel-Hilf, Nicole. „Naval Kishore Press – Digital: From Hidden Treasure to Open Access". *International Institute for Asian Studies - The Newsletter*, Autumn 2018. https://iias.asia/the-newsletter/article/naval-kishore-press-digital-hidden-treasure-open-access.

Latham and Scholes (2006): Sean Latham and Robert Scholes, "The Rise of Periodical Studies," *PMLA: Publications of the Modern Language Association*, 121 (2006), 517-531.

Sivers-Sattler (2001): Gabriele von Sivers-Sattler. "He Zhens Forderungen zur Namensgebung von Frauen im vorrevolutionären China: Untersuchungen zur anarchistischen Zeitschrift *Tian Yi (Naturgemäße Rechtlichkeit)* (1907-1908)". In Gimpel, Denise & Hanz, Melanie (editors). *Cheng - All in Sincerity: Festschrift in Honour of Monika Übelhör*, 275-284. [Hamburger Sinologische Schriften 2]. Hamburg: Hamburger Sinologische Gesellschaft e.V., 2001.

Sivers (2001): Gabriele von Sivers-Sattler. "Die mythische Figur Nügua in der anarchistischen Zeitschrift *Naturgemäße Rechtlichkeit (TianYi)*,1907-1908". In Übelhör, Monika (ed) *Zwischen Tradition und Revolution: Lebensentwürfe und Lebensvollzüge chinesischer Frauen an der Schwelle zur Moderne* [Beiträge zu einem Symposium des Fachgebietes Sinologie der Philipps-Universität Marburg vom 26. bis 28. November 1999], 105-130. [Schriften der Universitätsbibliothek Marburg, 107]. Marburg: Verlag der Universitätsbibliothek, 2001.

Sung, et al (2014): Sung, Doris, Liying Sun and Matthias Arnold. “The Birth of a Database of Historical Periodicals: Chinese Women’s Magazines in the Late Qing and Early Republican Period.” In *Tulsa Studies in Women’s Literature* 33, no. 2 (2014): pp. 227-37. http://muse.jhu.edu/article/564237.

Williams (1961): Williams, Raymond. *The Long Revolution*. Harmondsworth: Penguin Books, 1961.

Bibliography

[1] http://uni-heidelberg.de/ecpo

[2] http://uni-heidelberg.de/womag

[3] http://xiaobao.uni-hd.de/

[4] http://mhdb.mh.sinica.edu.tw/fnzz/

[5] http://kjc-sv034.kjc.uni-heidelberg.de/ecpo/api/mods

[6] http://www.asia-europe.uni-heidelberg.de/index.php?id=4425

[7] http://www.unicode.org/Public/UCD/latest/ucd/PropList.txt

[8] ABBYY Cloud OCR SDK http://ocrsdk.com/.

[9] Pallas Ludens http://pallas-ludens.com.