

Cluster of Excellence  
**ASIA AND EUROPE**  
IN A GLOBAL CONTEXT



**UNIVERSITÄT  
HEIDELBERG**  
ZUKUNFT  
SEIT 1386

# Wege zur Erschließung der frühen chinesischen Presse:

## Early Chinese Periodicals Online (ECPO)

Matthias Arnold | Uni Heidelberg · HCTS · HRA

# Research data – Chinese periodical press

- First decades of the 20th century
- Understudied, but dominated the contemporary print market and provide access to the "actual culture" (R. Williams, 1961)

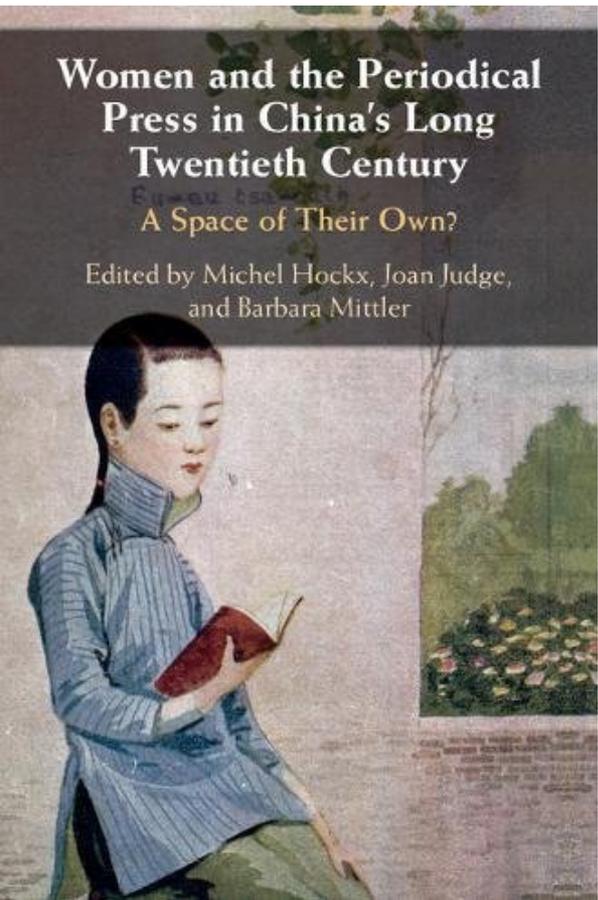
## Challenges

- Physically dispersed, often poorly preserved
- Voluminous (full runs, daily, up to >30 years)
- Multi-generic and intellectually demanding

## Approach

- Multi-disciplinary team, >10 researchers, 2 PhD theses
- *Women and the Periodical Press in China's Global Twentieth Century: A Space of Their Own?* Ed. by Joan Judge, Barbara Mittler and Michel Hockx, Cambridge University Press, 2018.

Database Early Chinese Periodicals Online (ECPO)



### Women and the Periodical Press in China's Long Twentieth Century A Space of Their Own?

Edited by Michel Hockx, Joan Judge,  
and Barbara Mittler

### TYPES OF PUBLICATIONS (239)

FILM	LITERATURE	XIAOBAO	POLITICAL
GENDERED	RELIGION RELATED	FAMILY LIFE	MISCELLA- NEOUS
LIFESTYLE	PICTORIALS	MEDICINE	FASHION
YOUTH	ART	FOREIGN PRESS	

<https://uni-heidelberg.de/ecpo>





anhui\_funu.jpg

anhui\_funu\_jikan.jpg

anqing\_nuzhong\_yuekan.jpg

beijing\_nuzi\_gdsf\_fs\_zooxsh\_zz.jpg

caifeng\_bao.jpg

chaochua\_yuekan.jpg

chunjiang\_huayu\_ebao.jpg

cijian\_funu.jpg

da\_bao.jpg

da\_shi\_jie.jpg

deyin\_banyuekan.jpg

diansheng\_ribao.jpg

diany...

dongfang\_xiaoshuo.jpg

fu\_er\_mo\_si.jpg

fujian\_funu\_yuekan.jpg

funu.jpg

funu\_gangwei.jpg

funu\_gongzuo.jpg

funu\_hezuo\_yundong.jpg

funu\_husheng\_fukan.jpg

funu\_jiating.jpg

funu\_jie.jpg

funu\_pinglun.jpg

funu\_pinglun\_zk.jpg

funu\_qia...

funu\_shenghuo.jpg

funu\_shenghuo\_yuekan.jpg

funu\_shijie.jpg

funu\_wenhua\_njl\_937.jpg

funu\_wenhua\_njl\_946.jpg

funu\_wenhua\_zstk.jpg

funu\_xianfeng.jpg

funu\_xinshenghuo\_yuekan.jpg

funu\_xinyun.jpg

funu\_xinyun\_tx.jpg

funu\_xunkan.jpg

funu\_xunkan\_km.jpg

funu\_yix...

funu\_yu\_jiating\_rk\_tjdg.jpg

funu\_yu\_jiating\_tj\_sb.jpg

funu\_yu\_jiating\_zk\_tjdg.jpg

funu\_yu\_jiating\_zz.jpg

funu\_yuebao.jpg

funu\_yuebao\_shj.jpg

funu\_yuekan.jpg

funu\_zhanxian.jpg

funu\_zhengyi\_zhoukan.jpg

funu\_zhishi.jpg

funu\_zhoukan.jpg

funue\_gongming\_banyuekan.jpg

funue\_g...

furen\_huabao.jpg

fusheng\_byk.jpg

fusheng\_yk.jpg

fuxiang\_tekan.jpg

fuxiang\_xunkan.jpg

fuying\_weisheng.jpg

fuyun\_yuekan.jpg

gansu\_funu.jpg

guangdong\_funu.jpg

guangxi\_funu.jpg

guangzhou\_funu\_tx.jpg

guangzhou\_nuqiangnian.jpg

gudao...

guoli\_nuzi\_shifan\_xueyuan\_yuankan.jpg

313 publications: 134 with items

jiating.jpg

jiating\_byk\_shenbao.jpg

jiating\_cd.jpg

jiating\_nk.jpg

jiating\_shdgb.jpg

jiating\_tjdg.jpg

jiating\_xingqi.jpg

jiating\_yanjiu.jpg

jiating\_yixue.jpg

jiating\_yiyao\_changshi.jpg

jiating\_yiyao\_yuekan.jpg

jiating\_yiyao\_zazhi.jpg

jiating\_y...

...

...

...

...

...

...

...

...

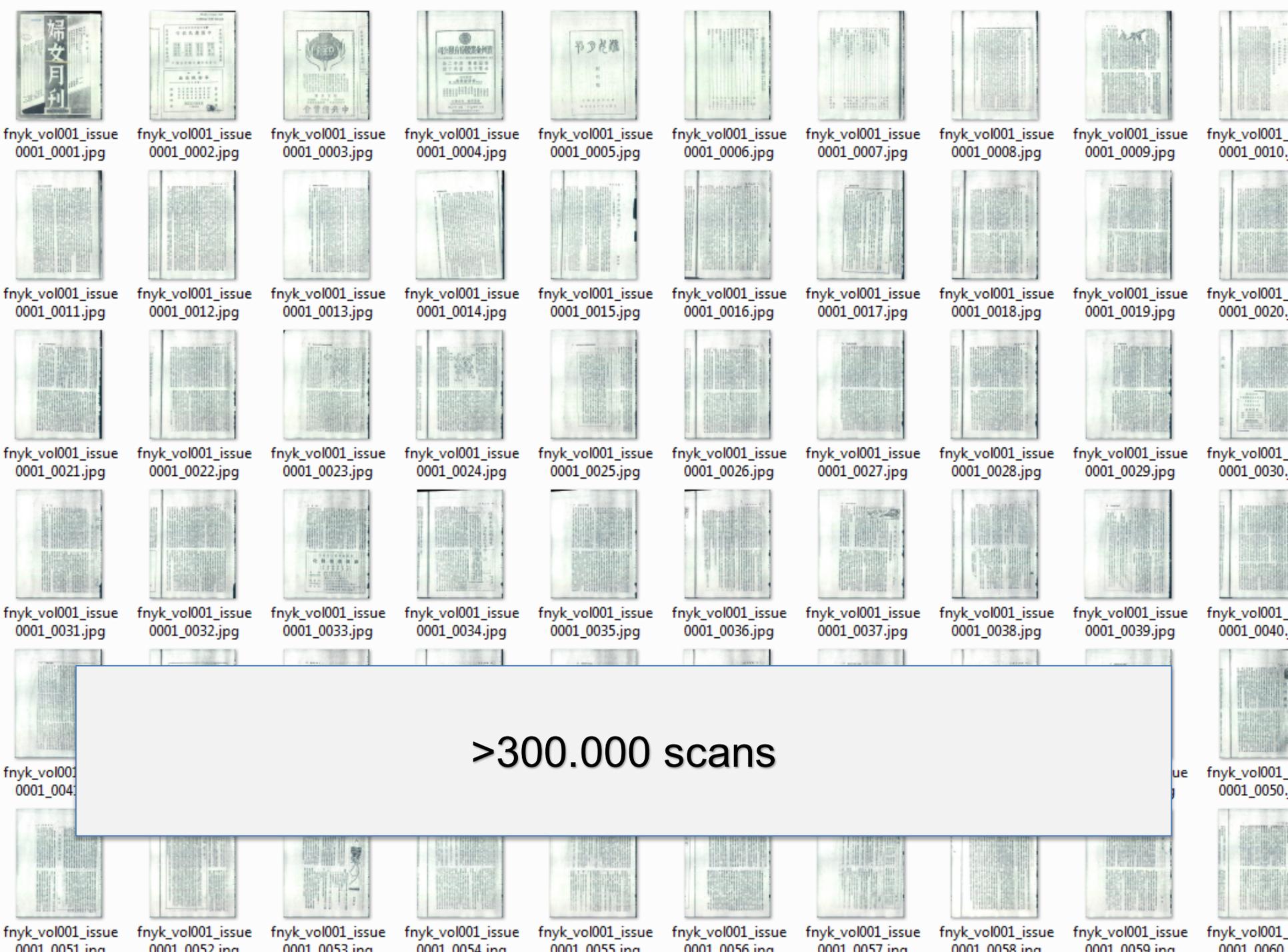
...

...

...

...

...



fnyk\_vol001\_issue\_0001\_0001.jpg

fnyk\_vol001\_issue\_0001\_0002.jpg

fnyk\_vol001\_issue\_0001\_0003.jpg

fnyk\_vol001\_issue\_0001\_0004.jpg

fnyk\_vol001\_issue\_0001\_0005.jpg

fnyk\_vol001\_issue\_0001\_0006.jpg

fnyk\_vol001\_issue\_0001\_0007.jpg

fnyk\_vol001\_issue\_0001\_0008.jpg

fnyk\_vol001\_issue\_0001\_0009.jpg

fnyk\_vol001\_issue\_0001\_0010.jpg

fnyk\_vol001\_issue\_0001\_0011.jpg

fnyk\_vol001\_issue\_0001\_0012.jpg

fnyk\_vol001\_issue\_0001\_0013.jpg

fnyk\_vol001\_issue\_0001\_0014.jpg

fnyk\_vol001\_issue\_0001\_0015.jpg

fnyk\_vol001\_issue\_0001\_0016.jpg

fnyk\_vol001\_issue\_0001\_0017.jpg

fnyk\_vol001\_issue\_0001\_0018.jpg

fnyk\_vol001\_issue\_0001\_0019.jpg

fnyk\_vol001\_issue\_0001\_0020.jpg

fnyk\_vol001\_issue\_0001\_0021.jpg

fnyk\_vol001\_issue\_0001\_0022.jpg

fnyk\_vol001\_issue\_0001\_0023.jpg

fnyk\_vol001\_issue\_0001\_0024.jpg

fnyk\_vol001\_issue\_0001\_0025.jpg

fnyk\_vol001\_issue\_0001\_0026.jpg

fnyk\_vol001\_issue\_0001\_0027.jpg

fnyk\_vol001\_issue\_0001\_0028.jpg

fnyk\_vol001\_issue\_0001\_0029.jpg

fnyk\_vol001\_issue\_0001\_0030.jpg

fnyk\_vol001\_issue\_0001\_0031.jpg

fnyk\_vol001\_issue\_0001\_0032.jpg

fnyk\_vol001\_issue\_0001\_0033.jpg

fnyk\_vol001\_issue\_0001\_0034.jpg

fnyk\_vol001\_issue\_0001\_0035.jpg

fnyk\_vol001\_issue\_0001\_0036.jpg

fnyk\_vol001\_issue\_0001\_0037.jpg

fnyk\_vol001\_issue\_0001\_0038.jpg

fnyk\_vol001\_issue\_0001\_0039.jpg

fnyk\_vol001\_issue\_0001\_0040.jpg

fnyk\_vol001\_issue\_0001\_0041.jpg

fnyk\_vol001\_issue\_0001\_0042.jpg

fnyk\_vol001\_issue\_0001\_0043.jpg

fnyk\_vol001\_issue\_0001\_0051.jpg

fnyk\_vol001\_issue\_0001\_0052.jpg

fnyk\_vol001\_issue\_0001\_0053.jpg

fnyk\_vol001\_issue\_0001\_0054.jpg

fnyk\_vol001\_issue\_0001\_0055.jpg

fnyk\_vol001\_issue\_0001\_0056.jpg

fnyk\_vol001\_issue\_0001\_0057.jpg

fnyk\_vol001\_issue\_0001\_0058.jpg

fnyk\_vol001\_issue\_0001\_0059.jpg

fnyk\_vol001\_issue\_0001\_0060.jpg

>300.000 scans

- ▶ Funu shenghuo yuekan
- ▶ Funu yuekan
  - ▶ data transfer
    - ▶ images\_renamed
      - ▶ volume001
        - issue001
        - issue002
        - issue003
        - issue004
        - issue005
        - issue006
      - ▶ volume002
        - issue001
        - issue002
        - issue003
        - issue004
        - issue005
        - issue006
      - ▶ volume003
        - issue001
        - issue002
        - issue003
        - issue004
        - issue005
        - issue006



40.936 issues: 46.931 articles, 20.532 images, 18.639 ads

- issue003
- issue004
- issue005
- ▶ volume006
  - issue001



**長篇小說 新上海春秋 (三三五) - CHANG PIAN XIAO SHUO XIN SHANGHAI CHUN QIU (SAN SAN WU)**

Article. 晶報 Jing bao, Volume 1, Issue 3843, Friday, 1939-04-21, Page 2

<b>TITLE:</b>	長篇小說 新上海春秋 (三三五) Chang pian xiao shuo xin Shanghai chun qiu (san san wu) Full-length novel: New Shanghai Year (335)
<b>PUBLICATION:</b>	晶報 Jing bao "The Crystal"
<b>PUBLICATION TYPE:</b>	NEWSPAPER
<b>DATE:</b>	21 April 1939
<b>PAGE:</b>	2
<b>PAGE NUMBER AS PRINTED:</b>	2
<b>SPECIAL POSITION:</b>	
<b>SEQUENCE:</b>	27
<b>DOCUMENT TYPE:</b>	Article
<b>KEYWORDS:</b>	law, upper class, bandit, lawyer
<b>AGENTS:</b>	周局長 Zhou ju zhang (mentioned in article), 唐少魯 Tang Shaolu (mentioned in article), 張延年 Zhang Yannian (mentioned in article), 楊賓秋 Yang Binqiu (mentioned in article), 沈家小姐 Shen jia xiao jie (mentioned in article), 董綬臣 Dong Shouchen (mentioned in article)
<b>ITEM LAST MODIFIED:</b>	
<b>DOCUMENT URL:</b>	<a href="http://ecpo.uni-hd.de/publications.php?magazin_id=1&amp;isid=20&amp;ispage=2&amp;itemid=299&amp;itype=2">http://ecpo.uni-hd.de/publications.php?magazin_id=1&amp;isid=20&amp;ispage=2&amp;itemid=299&amp;itype=2</a>



電聲日報  
DIAN SHENG RI BAO  
"RADIO MOVIE DAILY NEWS"



PUBLISHING INFORMATION



- Founding Date:** 1932-05-01
- End of publication:** 1933-12-28 (uncertain)
- Frequency:** Daily.
- Format/Size:** Size of a sheet: 40 x 28 cm.  
Format: 4 pages  
..... Details: medium-sized quadratic pages, 2 pages per 1 side of a sheet  
..... folding: vertically
- Print/Type:** - Original paper: issue no. 33 (1932-06-02): page 1: (本報) "是電影與無線電鼻祖；是彩紙四開報之創始者..."  
- Scans: are generally of bad quality, i.e. the images are very dark, rendering some texts and almost all of the pictures difficult or impossible to read.
- Price:** Standard price  
- issue no. 3 (1932-05-03): page 2: special (introductory?) price: 本埠零售：現特價僅售銅元六枚  
- issue no. 6 (1932-05-06): page 4: standard price: 本報價目：每日四頁四分；每月大洋七角；每三個月二元（寄費在內）  
  
Subscription price:  
- issue no. 4 (1932-05-04): page 2: subscription price: 優待訂閱：直接定戶每月連寄費原價一元一角現以創刊紀念特價每月僅收七角即小洋八角  
- issue no. 124 (1932-09-01): page 1: 本報自本日起、對於新訂閱者概照特價收費、每月僅七角、。。。
- Prominent agents:** 婦女日報 fu nu ri bao ()
- Summary of Content:** Columns/ sections:  
1. Dian ying 電影 Movie Critic  
... issue no. 0001 (1932-05-01) - last issue

# Search for Agent:

exact search  begin with  Filter Lang

[Agents](#) | [Find Duplicates](#) | [Reset Sort Order...](#) | [Add new Agent](#) | [Export Agents as csv](#)

47245 results: [1-100](#), [101-200](#), [201-300](#), [301-400](#), [401-500](#), [501-600](#), [601-700](#), [701-800](#), [801-900](#), [901-1000](#), [1001-1100](#), [1101-1200](#), [1201-1300](#), [1301-1400](#), [1401-1500](#), [1501-1600](#), [1601-1700](#), [1701-1800](#), [1801-1900](#), [1901-2000](#), [2001-2100](#), [2101-2200](#), [2201-2300](#), [2301-2400](#), [2401-2500](#), [2501-2600](#), [2601-2700](#), [2701-2800](#), [2801-2900](#), [2901-3000](#), [3001-3100](#), [3101-3200](#), [3201-3300](#), [3301-3400](#), [3401-3500](#), [3501-3600](#), [3601-3700](#), [3701-3800](#), [3801-3900](#), [3901-4000](#), [4001-4100](#), [4101-4200](#), [4201-4300](#), [4301-4400](#), [4401-4500](#), [4501-4600](#), [4601-4700](#), [4701-4800](#), [4801-4900](#), [4901-5000](#), [5001-5100](#), [5101-5200](#), [5201-5300](#), [5301-5400](#), [5401-5500](#), [5501-5600](#), [5601-5700](#), [5701-5800](#), [5801-5900](#), [5901-6000](#), [6001-6100](#), [6101-6200](#), [6201-6300](#), [6301-6400](#), [6401-6500](#), [6501-6600](#), [6601-6700](#), [6701-6800](#), [6801-6900](#), [6901-7000](#), [7001-7100](#), [7101-7200](#), [7201-7300](#), [7301-7400](#), [7401-7500](#), [7501-7600](#), [7601-7700](#), [7701-7800](#), [7801-7900](#), [7901-8000](#), [8001-8100](#), [8101-8200](#)

number of results:  
[10](#) [20](#) [50](#) [100](#) [500](#)

Name	Pinyin ↑	Lang	Type	Gender	edit	ECPO	FZ	Names	P-ID /N-ID	P-ID to merge into
Cooper, Jackie		English	Given Name	male	<a href="#">edit detail</a>	6	1	5	8462/9632	<input type="text"/>
賈克哥根	Jiake Gegen	Chinese	Other Name, Variants						8462/8786	<a href="#">merge</a>
茄基苛伯	Qieji Kebo	Chinese	Other Name, Variants						8462/42862	
賈克柯根	Jiake Kegen	Chinese	Other Name, Variants						8462/51938	
賈克珂潑	Jiake Kebo	Chinese	Other Name, Variants						8462/51939	
Frederick, Pauline		English	Given Name	female	<a href="#">edit detail</a>	1	0	2	9147/9634	<input type="text"/>
保林佛特立	Baolin Foteli	Chinese	Given Name						9147/9633	<a href="#">merge</a>
Gilbert, John		English	Given Name	male	<a href="#">edit detail</a>	3	0	2	8983/9637	<input type="text"/>
約翰吉										<a href="#">merge</a>
McL...										<input type="text"/>
萊波爾										<a href="#">merge</a>
McLaglen, Victor		English	Given Name	male	<a href="#">edit detail</a>	1	0	2	9159/9650	<input type="text"/>
維多麥克拉倫	Weiduo Maikelalun	Chinese	Given Name						9159/9647	<a href="#">merge</a>
Negri, Pola		English	Given Name	female	<a href="#">edit detail</a>	1	0	2	9158/9651	<input type="text"/>

47.245 agents, 163.408 occurrences, 15 'languages'

# Opening the data silo

From static export to **dynamic data service**

- Output data using the Metadata Object Description Schema (MODS) - Open Access: <http://ecpo.uni-hd.de/api/mods/>

From static pre-rendered files to **dynamic image service**

- Implementation of International Image Interoperability Framework (IIIF) Image API <http://iiif.io/technical-details/>

From separate names to **cross-db agents service**

- Identify agent, assign names, link to authorities, structure information, feed data back to authority files (GND)



**Wege zum Volltext:**

**Vorarbeiten**

JADH 2018

第七百零三號

「特等」登於新聞之中價面議  
「頭等」封面分格計三英寸寬  
「二英寸高為一格每格每格洋



**定價報格**

每三日出一張增刊無定  
中國境內全年二元半年  
元一角每月二角日本同  
餘外國各埠加倍均作大  
計算報費先惠郵費在內  
票加一

今日一張售大洋二分

**新造 洋房 廉價**

茲有坐落法租界唐家灣  
靈樞萬路水福里內新造  
高大三層樓洋房一幢汽  
車間洋式白磁浴缸面盆  
冷熱水龍頭西式活動廁  
所電燈自流井水一應俱  
全外附華式大花廳連樓  
一幢油漆精雅地位合宜  
深合公館住宅之用租金

清儀閣所藏  
古器文物

全十册 夾紙 預十元 索樣請 郵分 附本閱

上海商務印書館啟

包天笑先生著

**上海春秋**

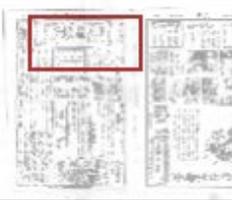
第二集出版

登場人物愈多  
情節愈加熱鬧  
文字愈加緊湊

欲知海上種種奇事豔事  
不可不讀此書

每集二册 每集一元二角

上海大東書局啟



上海春秋第二集出版 - SHANGHAI CHUNQIU DI ER JI CHU BAN  
Advertisement. 晶報 Jing bao, Volume 1, Issue 703, Sunday, 1925-01-04, Page 1

**TITLE:** 上海春秋第二集出版  
Shanghai Chunqiu di er ji chu ban

**PUBLICATION:** 晶報 Jing bao "The Crystal"

**PUBLICATION TYPE:** NEWSPAPER

**DATE:** 04 January 1925

**PAGE:** 1

**PAGE NUMBER AS PRINTED:** 1

**SPECIAL POSITION:**

**SEQUENCE:** 1

**DOCUMENT TYPE:** Advertisement

**PRODUCT BRAND:**

**ADVERTISEMENT CATEGORY:** 16

**KEYWORDS:** books, Shanghai

**AGENTS:** 包天笑 Bao Tianxiao (mentioned in advertisement), 大東書局 Dadong shu ju (mentioned in advertisement)

**NOTES ON ITEM:**

包天笑先生著上海春秋第二集出版登場人物愈多情節愈加熱鬧文字愈加緊湊欲知海上種種奇事豔事不可不獨此書每集二冊，每集一元二角上海大東書局啟

**ITEM LAST MODIFIED:**

**DOCUMENT URL:** [http://ecpo.uni-hd.de/publications.php?magazin\\_id=1&isid=282&](http://ecpo.uni-hd.de/publications.php?magazin_id=1&isid=282&ispage=1&itemid=4754&itype=4)

[ispage=1&itemid=4754&itype=4](http://ecpo.uni-hd.de/publications.php?magazin_id=1&isid=282&ispage=1&itemid=4754&itype=4)





**Wege zum Volltext:**

**Vorarbeiten - OCR**

JADH 2018

[Home](#)[Product Tour](#)[Plans and Pricing](#)[Support](#)[SLA](#)[Security](#)[Demo](#)[For Students](#)[About us](#)

Load source file (\*.png or \*.jpg):

jb\_0016\_1919-04-18\_0002+0003.jpg

I agree with the [Terms of use](#)Or paste url to source file (\*.png or \*.jpg): 

Recognition languages

Chinese Traditional

Profile

DocumentConversion

Image source

Auto

[More languages](#)

Source



Result





俳體西江月（並序）惡消聽也

通信社為報館輔助機關良裨社會  
苟能偵察隱私溫犀曷翅讚之不暇  
焉所用惡惟惡夫捕風捉影附會無  
根之事張皇不經之言名惟徒亂人  
意且令負責者代為分謗甯非一  
市之大蠹也哉戲填小詞以懲此輩

云爾

嚮壁最能虛造閉門儘可與謠非非是  
是一團糟墨白全然亂了 消息道聽  
塗說文章東抹 鈔郵花浪貼不辭勞  
還捏紛紛電報

俳  
通·信·社·  
苟·能·偵·  
焉·所·用·  
根·之·事·  
意·且·令·  
市·之·大·  
云·爾·  
嚮·壁·最·能·  
是·一·團·  
塗·說·文·  
還·捏·紛·

還·捏·紛·紛·電·報·  
塗·說·文·章·東·抹·  
是·一·團·糟·墨·白·全·然·亂·了·  
嚮·壁·最·能·虛·造·閉·門·儘·可·與·謠·非·非·是·  
消·息·道·聽·

俳辭一〓〓（並序）惡術聽也

通信社為報腋秬助餞關良裨社鈔  
苟〓〓察一〓泓庵〓〓郊謗之不暇  
忍所用惡惟惡夫捕風從影附會無  
根之事張皇不經之言名佞徒亂人  
意〓〓令負責寶者代為分謗甯#

市之大茲也哉敝垣小試以彷彿此依

云蜩

箔壁最能虛逝阳門儘可典謠非弗是  
是一网糴墨打全然亂多消息遺舛  
淦說文帘束抹。鈔郵花浪貼不辭勞

通捏紛紛訊軋

嚮壁最能虛造閉門儘可與謠非非是

是一團糟墨白全然亂了 消息道聽

塗說文章束抹 鈔郵花浪貼不辭勞

還捏紛紛電報

嚮壁最能虛造閉門儘可與謠非非是

是一團糟墨白全然亂了 消息道聽

塗說文章束抹 鈔郵花浪貼不辭勞

還捏紛紛電報

俳

通信社

苟能信

焉所用

根之事

意且令

市之大

云爾

嚮壁最能虛造閉門儘可與謠非非是

是一團糟墨白全然亂了 消息道聽

塗說文章束抹 鈔郵花浪貼不辭勞

還捏紛紛電報

俳辭一々（並序）惡術聽也

通信社為報腋秭助餞關良裨社鈺

苟々々察Ⅰ泓庵3.2 郊謗之不暇

忍所用惡惟惡夫捕風從影附會無

根之事張皇不經之言名佞徒亂人

意。令負責寶者代為分謗甯# |

市之大茲也哉敝垣小試以彷彿依

云蜩

箔壁最能虛逝阳門儘可典謠非弗是

是一网糴墨打全然亂多消息道艸

淦說文帘束抹。鈔郵花浪貼不辭勞

通捏紛紛訊軋

嚮壁最能虛造閉門儘可與謠非非是

是一團糟墨白全然亂了 消息道聽

塗說文章束抹 鈔郵花浪貼不辭勞

還捏紛紛電報

嚮壁最能虛造閉門儘可與謠非非是

是一團糟墨白全然亂了 消息道聽

塗說文章束抹 鈔郵花浪貼不辭勞

還捏紛紛電報

俳

通信社

苟能信

焉所用

根之事

意見令

市之大

云爾

還捏紛紛電報

俳

通信社

苟能信

焉所用

根之事

意見令

市之大

云爾

還捏紛紛電報

ca. 63% correctly recognized

本報具有廿一年的歷史是小型報紙其祖籍路普通全國刊登廣告最有力

法界懸旗糾紛

今日可復業

被捕市民昨日均已釋放

【本報訊】法界各界，自被捕後，即由法界各界發起，組織法界救國會，並由法界救國會，向法界各界，發起救濟法界，並由法界救國會，向法界各界，發起救濟法界，並由法界救國會，向法界各界，發起救濟法界...

交涉經過

【本報訊】法界交涉，自被捕後，即由法界各界發起，組織法界救國會，並由法界救國會，向法界各界，發起救濟法界，並由法界救國會，向法界各界，發起救濟法界...

河內

【本報訊】河內各界，自被捕後，即由河內各界發起，組織河內救國會，並由河內救國會，向河內各界，發起救濟河內，並由河內救國會，向河內各界，發起救濟河內...

釋放經過

【本報訊】法界被捕者，自被捕後，即由法界各界發起，組織法界救國會，並由法界救國會，向法界各界，發起救濟法界，並由法界救國會，向法界各界，發起救濟法界...

法界救國會

【本報訊】法界救國會，自成立後，即由法界各界發起，組織法界救國會，並由法界救國會，向法界各界，發起救濟法界，並由法界救國會，向法界各界，發起救濟法界...

外交界動向

【本報訊】外交界動向，自被捕後，即由外交界發起，組織外交界救國會，並由外交界救國會，向外交界，發起救濟外交界，並由外交界救國會，向外交界，發起救濟外交界...

杏花酒樓 高貴禮堂 租費從廉 隨處小酌 大小通酌

好美華學 高貴禮堂 租費從廉 隨處小酌 大小通酌

陽痿 治限日 早洩 治限日 保能 治限日

上海第一流攝影專家 國泰藝術照相館 化裝與照相化濟經價定 機良影攝光春此際

報晶 The Crystal 發行人 A.L. Teodoro



中華民國二十八年四月廿一日 星期五 今日慶祝國慶四週年 本報創刊於三月八日

各路華軍大舉反攻 克復石龍包圍南昌 贛南高安發生激烈爭奪戰 綏首重要據點多處已收復



【本報訊】各路華軍，自反攻後，即由各路華軍發起，組織各路華軍救國會，並由各路華軍救國會，向各路華軍，發起救濟各路華軍，並由各路華軍救國會，向各路華軍，發起救濟各路華軍...

抱定抗戰決心 中國現決不談和 英當局亦聲明未提議調解 本報復刊小啓

爽前 治限日 早洩 治限日 保能 治限日

品需必之期暑 點特大四 汽鮮順 海制鮮流 潔華美消 漸製可蒸水 生機口潔



**Wege zum Volltext:**

**Binnenstrukturen**

HAASDZ 2018

本報具有廿一年的歷史是小型報紙鼻祖銷路普及全國刊登廣告最有效果

### 各商店遵勸 今日可復業

#### 被補市民昨日均已釋放

關於以後自由懸旗一節特待解決

上海市商會及法租界商會，為維持法租界秩序，並便利市民生活起見，特向各商店發出勸告，謂：(一)凡法租界內各商店，應即停止懸掛任何旗幟，(二)凡法租界內各商店，應即停止營業，(三)凡法租界內各商店，應即停止一切活動，(四)凡法租界內各商店，應即停止一切宣傳，(五)凡法租界內各商店，應即停止一切集會，(六)凡法租界內各商店，應即停止一切遊行，(七)凡法租界內各商店，應即停止一切示威，(八)凡法租界內各商店，應即停止一切騷擾，(九)凡法租界內各商店，應即停止一切破壞，(十)凡法租界內各商店，應即停止一切投訴。

### 河內 航郵已通

河南航郵，自開航以來，頗受各界歡迎，現因河內航線，業已恢復，航郵亦隨之恢復，每日上午八時，由鄭州開往河內，下午二時，由河內開往鄭州，航郵之恢復，對於河內之交通，大有裨益。

### 法租界 交涉經過

法租界交涉，自開辦以來，頗受各界歡迎，現因法租界內各商店，業已恢復，交涉亦隨之恢復，每日上午八時，由法租界開往法租界，下午二時，由法租界開往法租界，交涉之恢復，對於法租界之交通，大有裨益。

### 法租界 報載紀實

法租界報載紀實，自開辦以來，頗受各界歡迎，現因法租界內各商店，業已恢復，報載亦隨之恢復，每日上午八時，由法租界開往法租界，下午二時，由法租界開往法租界，報載之恢復，對於法租界之交通，大有裨益。

### 路馬四 杏花酒樓

路馬四里，杏花酒樓，華貴禮堂，租費從廉，大小筵席，隨意小酌，九四六二九。

### 克復石龍包圍南昌

#### 各路華軍大舉反攻

#### 贛南高安發生激烈爭奪戰

#### 緩督重要據點多處已收復

(重慶廿日電) 據駐贛軍消息，各路華軍，大舉反攻，克復石龍，包圍南昌，贛南高安，發生激烈爭奪戰，緩督重要據點，多處已收復。

### 抱定抗戰決心 中國現決不談和

#### 英當局亦聲明未提議調解

(重慶廿日電) 中國現決不談和，抱定抗戰決心，英當局亦聲明，未提議調解。

### 陽痿早洩 治癒如神

#### 早洩早洩 治癒如神

陽痿早洩，治癒如神，早洩早洩，治癒如神。

### 本報復刊啟事

本報復刊啟事，本報自開辦以來，頗受各界歡迎，現因本報內各商店，業已恢復，復刊亦隨之恢復，每日上午八時，由本報開往本報，下午二時，由本報開往本報，復刊之恢復，對於本報之交通，大有裨益。

# 報晶

## The Crystal

發行人 A. L. Teodoro

八二四三九路電 號九九二路口漢越館

前報之遺寄益後包無照錄後立號特准特政郵局中

《號四廿字〇號記登局部工租共公》

《號五十九百七千三第》

### 家專影攝流一第上海

### 館相照術藝泰國

化族貴相照化濟輕價定

### 機良影攝光春此際

號五五路京南址地

號三九三〇九話電

日三月三年八國民於前報本 重四分貳幣國管張貳日今 五期星 日一廿 四年捌拾貳國民華中

地圖

各路華軍大舉反攻

克復石龍包圍南昌

贛南高安發生激烈爭奪戰

緩督重要據點多處已收復

### 抱定抗戰決心 中國現決不談和

#### 英當局亦聲明未提議調解

(重慶廿日電) 中國現決不談和，抱定抗戰決心，英當局亦聲明，未提議調解。

### 本報復刊啟事

本報復刊啟事，本報自開辦以來，頗受各界歡迎，現因本報內各商店，業已恢復，復刊亦隨之恢復，每日上午八時，由本報開往本報，下午二時，由本報開往本報，復刊之恢復，對於本報之交通，大有裨益。

### 應有懸國旗之自由

#### 謝晉元團長昨發告同胞

謝晉元團長，昨發告同胞，應有懸國旗之自由。

### 陽痿早洩 治癒如神

#### 早洩早洩 治癒如神

陽痿早洩，治癒如神，早洩早洩，治癒如神。

### 本報復刊啟事

本報復刊啟事，本報自開辦以來，頗受各界歡迎，現因本報內各商店，業已恢復，復刊亦隨之恢復，每日上午八時，由本報開往本報，下午二時，由本報開往本報，復刊之恢復，對於本報之交通，大有裨益。





### 法租界懸旗糾紛

公文上海日  
北國駐滬總  
特生夫婦、  
在卽、特於  
(九)、在  
舉行茶會  
埠之比僑  
界國民精神總動員協會、於本  
二十二日、正式成立後、即電呈  
其精神總動員會蔣會長、報告成  
及工作狀況、頃該會接奉全國  
國民精神總動員會張秘書長

滬市工務林立、  
戰後人口突增、工  
業更見  
原料  
或游  
股市  
稻草  
上之

## 不堪勒索

### 瞿順昌喪心與日人朋

報載、比國駐滬總領事古特生夫婦、以離滬在即、特於昨日(十九)、在滬市工務林立、戰後人口突增、工業更見發達、原料或游股市、稻草上之

### 不堪勒索停業

#### 瞿順昌喪心與日人朋比為奸

業更見發達、而工  
業上必要之農產品  
或游擊隊、致一  
股市行情趨居奇、  
稻草一項、亦工  
立分支行、逐項推  
加後方生產、充實  
抗戰力量云、

### 日捕加薪

與西捕同待遇二十  
日法文上海日報載  
收、公共租界巡捕  
房之日薪標準、  
將增加薪水百分之  
三十五、而與西捕

### 滑稽一重

陰兵們到  
旁邊一個年紀大一點的、陪  
都已經弄得很亂了。他們的  
一來、吃不慣天上的東西、  
飯匙毒藥在內、所以無論什  
帶到了天上、燃料就無處  
中他們就把人家的桌子椅子  
這還有許多想講、就叫我們

統計、惟據捕房消息、曾有越捕及華捕  
各一人受傷、被捕華人、經納稅會交涉  
後、已於昨日下午四時許、陸續釋放、  
聞在捕房拘押時、多數均忍痛絕食、愛  
國熱情、充分流露、惟被捕諸人中有二  
名、捕房以其呼口號稱爲惡劣、初不允  
釋放、經再三交涉、亦允釋放、上海市  
銀行業同業公會、對於會員銀行之遭受  
騷擾干涉、一面致電對開、一面勸其先  
行復業、各銀行感以公會誠懇之勸慰、  
亦即照常營業、又黃金大戲院執照被吊  
銷後、聞亦發還、定即日起照常營業云  
廿日法文日報載、昨日法租  
界內、並無重大事件  
發生、外間之傳說、殊不足信、聞在昨  
日因大規模懸掛國旗事件、顯有政治性  
質之表示、故當局乃飭令各居民各自收  
旗、當時並無反抗情事、關於此點、法

報載、比國駐滬總領事古特生夫婦、以離滬在即、特於昨日(十九)、在滬市工務林立、戰後人口突增、工業更見發達、原料或游股市、稻草上之

國民精神總動員會張秘書長  
岳軍復電、慰勉有加、茲照  
錄如次(上略)：國民精神  
總動員、確屬抗戰建國之重  
要方略、滬市淪陷已久、敵  
寇誘脅甚甚、策動愛國之精  
神、爲抗戰之聲援、尤屬必  
要、貴會領導羣衆、組織協  
會、並利用各界固有之職業  
團體、普及分途推進之持久機關、深謀  
碩劃、殊堪嘉許、應准備案、仍希繼續  
努力、奉行不懈、是所厚望、張章叩

四行軍團團長謝晉元、昨  
出巡市內、緝獲掛國旗者同  
昨日爲上海市農工商  
各界響應國民政府精  
神總動員運動、租界  
出與、一致懸掛紀念、愛受  
法租界當局取締、發生糾紛  
、應旗爲代表一國之尊嚴  
、任何國民、均應掛具國  
旗之自由與權利、同胞乎、  
石等英勇果敢、愛護國家、  
擁護最高領袖之行為、世人  
同深欽仰、余前次發表「一  
平來觀感及對時局感想」文  
曾言：今後國民唯有本「國  
本興亡、匹夫有責之義、各  
守崗位、竭盡天職、以最高  
立分支行、逐項推  
加後方生產、充實  
抗戰力量云、

四馬路  
里錦豐  
專辦  
高尙粵菜  
華貴禮堂  
租費從廉

# 杏花酒樓

大小筵席  
隨意小酌  
電話  
九四六二九

美好良園

高尙  
娛樂  
妙演劇團  
新求樂  
之溫柔鄉

電話  
二〇八七號四〇二一

各路華軍大舉反攻

紙鼻祖銷路普

任何國國民

應有懸國旗之自  
謝晉元團長昨發告同胞

領袖之主張爲主  
意志、而腳踏實  
國、孤島之同胞  
年開始、願人人  
深望孤島同胞、  
效祖國之責任、  
有力出力、以完  
之基礎、余今  
向孤島同胞呼籲  
所受之苦痛、區  
之黑暗、一切惡  
華民族偉大光榮  
、切望孤島同  
有責、出錢出力  
、硬幹、苦幹、望  
負責、盡責、望  
按：、又有形同  
稅卡之機關出現於  
相率停業、致稻草  
來源驟減、各工廠  
受缺貨影響、對動  
索陋規深表憤慨、  
同時車船亦反對絡  
同業、一致反對云





# 報晶

## The Crystal

發行人 A. L. Teodoro

八二四三九路道 號九二路日漢法新

紙新之送寄妥包郵照按立號掛准特政郵華中

號五十九百七千三第

家專影攝流一第上海  
館相照術藝泰國

化族貴相照化濟經價定

機良影攝光春此際

號五五路京南址地

號三九三〇九話電

日三月三年八國民於前報本 厘四分幣國幣張貳日今 五期星 日一廿月四年捌拾貳國民華中

**國際時事**

德軍在波蘭境內，正積極地進行防禦，德軍在波蘭境內，正積極地進行防禦，德軍在波蘭境內，正積極地進行防禦...



**克復石龍包圍南昌**

贛南高安發生激烈爭奪戰 緩督重要據點多處已收復

**抱定抗戰決心 中國現決不談和**

英當局亦聲明未提議調停

**宿潼流寇各得 華軍甚得手**

開闢蘇北戰線

**桂軍未失陷**

桂軍未失陷，桂軍未失陷...

**本報復刊小啓**

本報自一月一日起，由美商發行，恢復舊觀...

**德機襲滬**

德機襲滬，德機襲滬...

**德機襲滬**

德機襲滬，德機襲滬...



華貴禮堂 租賃從廉

**杏花酒樓**

高貴禮堂 租賃從廉

華貴禮堂 租賃從廉

**高崗野地**

娛樂 妙筆出

娛樂 妙筆出

**陽痿**

治癒可以早洩

治癒可以早洩

**早洩**

治癒可以早洩

治癒可以早洩

**各商店遵勸 今日可復業**

被捕市民昨日均已釋放

【本報訊】上海各商店，自昨日（十九日）下午四時起，均已遵勸復業。據悉，昨日被捕之市民，均已獲釋。各商店復業後，秩序井然，市民生活亦趨正常。

**河內 航郵已通**

【本報訊】河內航線，自昨日（十九日）下午四時起，已恢復通航。航郵亦隨之恢復，市民通訊便利。

**不推勒索停業**

【本報訊】本市各商店，因不推勒索，已停止營業。市民對此表示不滿，呼籲政府採取措施，保護商民利益。

**任何國民 應有懸國旗之自由**

【本報訊】任何國民，均應有懸掛國旗之自由。此項自由，為國民之基本權利，不容忽視。

**淪陷區公產**

【本報訊】淪陷區之公產，應予保護。政府應採取措施，防止公產流失，維護國家利益。

**第一系統**

【本報訊】第一系統之各項工作，正積極進行中。政府將加強對第一系統之管理，提高行政效率。

**夏季之必需品**

順風牌 汽水

順風牌 汽水

本報具有廿一年的歷史是小型報紙鼻祖銷路普遍全國刊登廣告最有力

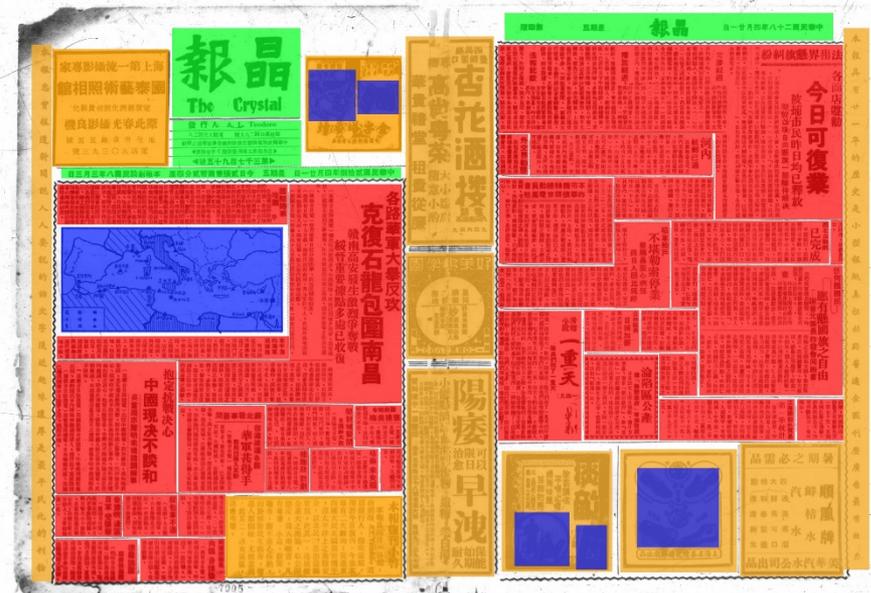
**Wege zum Volltext:**

**Manuelle Segmentierung (Crowd)**

JADH 2018

# Segmentation - I

- Page segmentation (pattern recognition/computer vision)
  - Analyze layout of page, use page-internal structures
  - Identify semantic units
  - Generate co-ordinates, relate them to items, store in DB



# Segmentation - II

---

- Page segmentation (crowdsourcing)
  - Pilot project with Pallas Ludens GmbH
  - Let the crowd help analyzing the pages
    - Identify and label four item types:
      - image/drawing
      - article
      - advertisement
      - additional information
    - Supervised
    - Non-Chinese speaking community!



# 晶報

**定報** 每月一元二角 半年七元 全年十二元  
**零售** 每份五分  
**廣告** 第一版每行一元 第二版每行八角 第三版每行六角 第四版每行四角  
**印刷** 每日下午四時以前送稿 逾期不候  
**地址** 上海南京路  
**電話** 二二八號

**廣告** 第一版每行一元 第二版每行八角 第三版每行六角 第四版每行四角  
**印刷** 每日下午四時以前送稿 逾期不候  
**地址** 上海南京路  
**電話** 二二八號

### 科發白濁丸

此丸專治白濁、淋病、尿道炎、膀胱炎、腎臟炎、前列腺炎、婦女經期不調、赤白帶下、子宮炎、附件炎、盆腔炎、不孕症、產後諸症、功效神速、無副作用、每盒一元、郵費在內。

### 請吸新出 長城牌國貨香煙



長城牌香煙，品質優良，口味醇厚，為國貨之光。每包十支，售價一元。

### 上海五洲大藥房

**自來血** 補血強身，治貧血、萎黃、失眠、健忘、心悸、頭暈、眼花、腰酸、背痛、四肢無力、精神不振、每瓶二元。

**樹皮丸** 治風濕、關節炎、腰酸背痛、手足麻木、每瓶一元。

**海波藥** 治咳嗽、氣喘、痰多、胸膈不舒、每瓶一元。

### 上海三馬路一號 中法儲蓄會

繁華世界 金錢如流 水東來西 去欲省而 不能今有 妙法每月 在消耗費 中節省一 份錢去投 入儲蓄會 裏將來可 以集成巨 數并且每 月有得獎 的希望中 法儲蓄會 成立最早 會員最多 當眾開獎 最靠得住 中國政府 立案法國 政府註冊 開獎在即 不要錯過

### 德六零六

專治梅毒、淋病、皮膚病、功效神速。每瓶一元，郵費在內。

### 西醫張世楷

專治梅毒、淋病、皮膚病、功效神速。每瓶一元，郵費在內。

### 美少年

美容聖品，使皮膚細嫩、容光煥發。每瓶一元。

### 萬應消癰丸

治一切瘡癤、疔毒、無名腫毒、功效神速。每瓶一元。

### 秋令咳嗽

治秋季咳嗽、氣喘、痰多、胸膈不舒。每瓶一元。

### 預防花柳奇藥 愛克憐

預防花柳、梅毒、淋病、功效神速。每瓶一元。

### 萬隆老棧

專營各省名產、綢緞、布匹、百貨。地址：上海南京路。

### 安兜氏補藥片

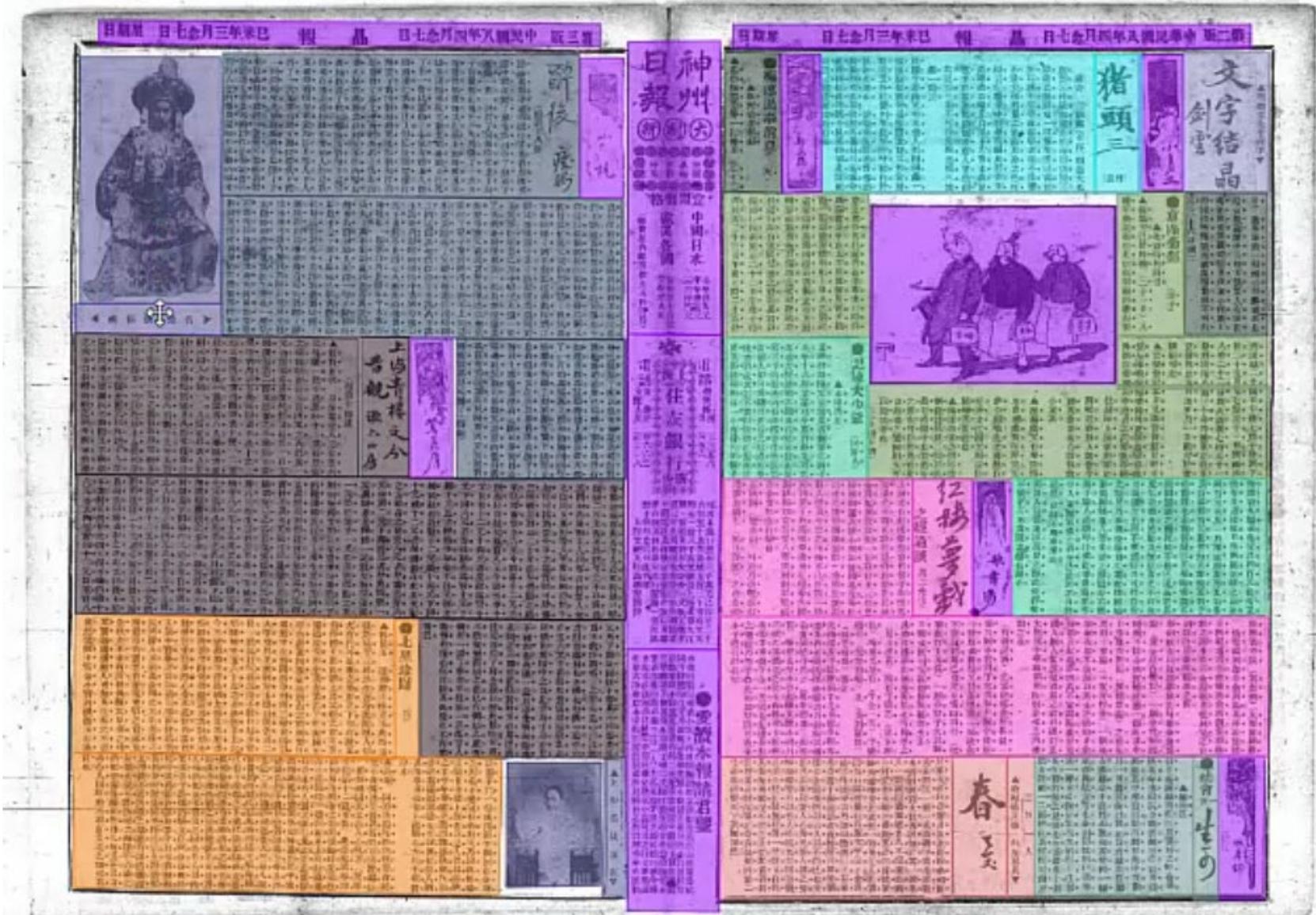
補血強身，治貧血、萎黃、失眠、健忘、心悸、頭暈、眼花、腰酸、背痛、四肢無力、精神不振。每瓶一元。



上海北京路安兜氏西藥公司謹啓

**寒雲主人好古知書深探三代漢魏之神髓**  
 求書者不暇應爰擬定書例如下  
 一、凡欲借書者請向本館領取借書卡一張，內註姓名、地址、職業、年齡、學歷、籍貫、電話、郵政信箱等項，以便寄書及催還。  
 二、借書卡內註明借書日期、書名、冊數、借閱期限等項。  
 三、借書卡內註明借書日期、書名、冊數、借閱期限等項。  
 四、借書卡內註明借書日期、書名、冊數、借閱期限等項。

# Grouping semantic units



# Outcome of segmentation pilot

---

1. Page segmentation can be outsourced to expert crowd
  - Requires supervision
  - Advanced user interfaces (high usability, efficiency)
  - Crowd should read Chinese (semantic grouping)
2. *Jingbao* 晶報 1919-21 completely segmented with qualified boxes, issues of April 1919 with semantic units
3. Further processing:
  - Partnership with Computational Knowledge Lab (知識計算實驗室), Department of Engineering Science and Ocean Engineering, Taiwan National University, <http://www.cklab.org/>
  - Seeking additional partners for collaboration!

**Wege zum Volltext:**

**Ground truth**

# Funding for Ground Truth

---

Zusätzliche Mittel über Field of Focus 3 (ExStrat Heidelberg), Juli-  
Oktober 2019

Gemeinsam mit Duncan Paterson und 4 WiHi (typing)

- GroundTruth bounding boxes mit Labels
  - Koordinaten in json und web-annotation Format
  - Jingbao April 1939 und April 1938
  - Stand: 70 folds, 6335 shapes, Ø 90,5 shapes/fold
- GroundTruth Texte
  - Blind double-keying
  - Jingbao April 1939 - 10 issues, 40 folds
  - Stand: ca. 245.000 Zeichen, lokales XML Format, Ø 6100 Zeichen/fold (Bearbeitung ca. 10h/fold)

# Annotation tool



mode:select

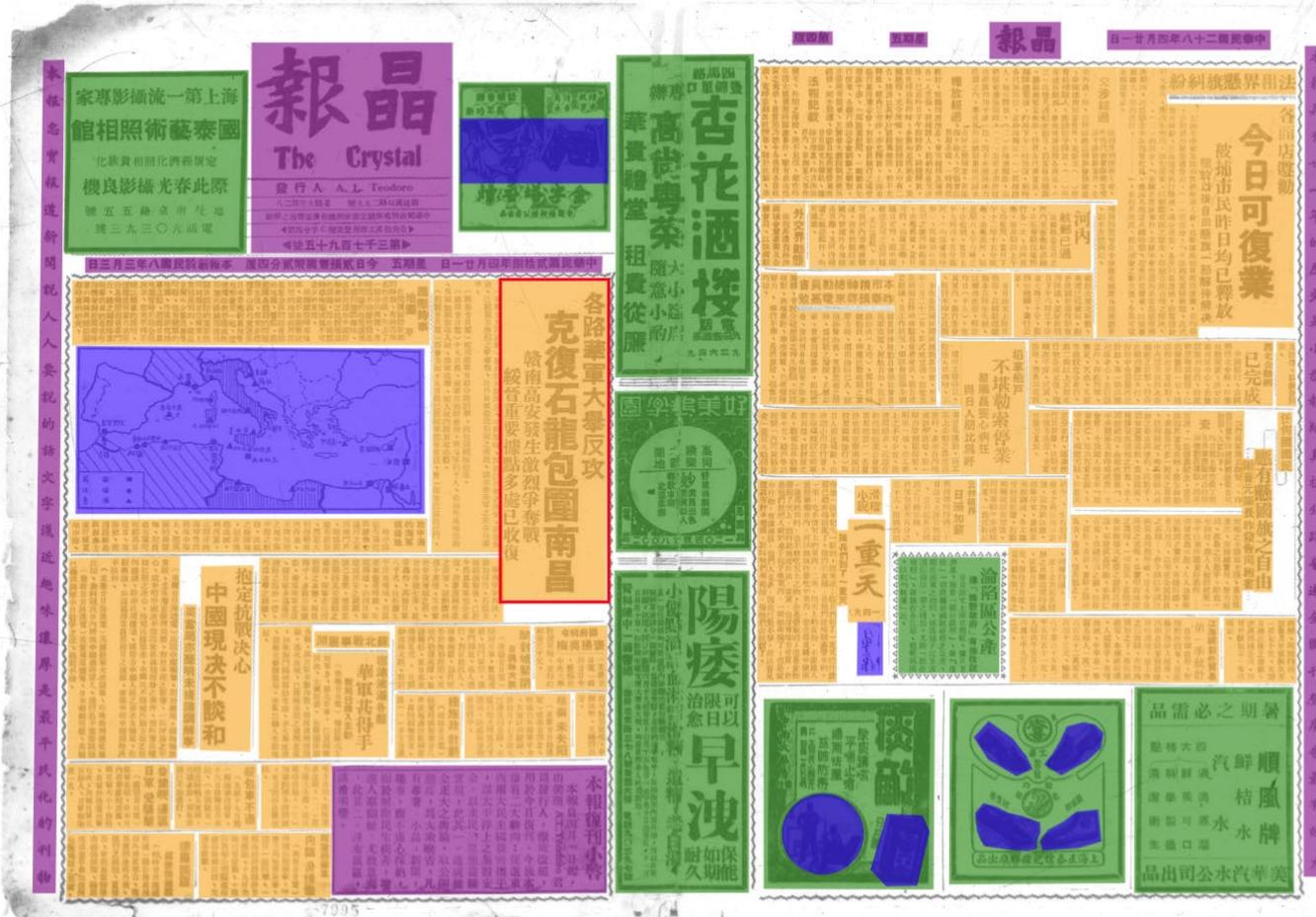


Image or Drawing

Article

Advertisement

Additional Information

SUBMIT



Entwicklung: eXist-Solutions

# Annotationen in json

jb\_3795\_1939-04-21\_0002+0003.json x

```
1 {
2   "items": [
3     {
4       "body": [
5         {
6           "value": {
7             "color": "purple",
8             "name": "additional",
9             "label": "Additional Information"
10          },
11          "type": "CategoryLabel"
12        }
13      ],
14      "created": "2019-07-10T16:14:26.772Z",
15      "target": [
16        {
17          "type": "SpecificResource",
18          "source": "https://kjc-sv002.kjc.uni-heidelberg.de:8080/fcgi-bin/iiprv.fcgi?IIIF=imageStorage/ecpo_new/jingbao/1939/04/jb_3795_1939-04-21_0002%252B0003.tif/full/full/0/default.jpg",
19          "id": "s-156277526676110",
20          "selector": {
21            "value": "<g transform=\"matrix(1 0 0 1 3462.22464 371.24256)\"><polygon points=\"-812.0176,-74.99264 812.0176,-74.99264 812.0176,74.99264 -812.0176,74.99264 \"/></g>",
22            "type": "SvgSelector"
23          }
24        }
25      ],
26      "type": "Annotation",
27      "id": "s-1562775266771"
28    },
29    {
30      "body": [
31        {
32          "value": {
33            "color": "purple",
34            "name": "additional",
35            "label": "Additional Information"
36          },
37          "type": "CategoryLabel"
38        }
39      ],
40      "created": "2019-07-10T16:14:48.631Z",
41      "target": [
42        {
43          "type": "SpecificResource",
44          "source": "https://kjc-sv002.kjc.uni-heidelberg.de:8080/fcgi-bin/iiprv.fcgi?IIIF=imageStorage/ecpo_new/jingbao/1939/04/jb_3795_1939-04-21_0002%252B0003.tif/full/full/0/default.jpg",
45          "id": "s-15627752886237",
46          "selector": {
47            "value": "<g transform=\"matrix(1 0 0 1 4331.59808 1719.37056)\"><polygon points=\"-26.18944,-1352.42848 53.00704,-1352.42848 26.23776,1352.42848 -53.00704,1352.42848 \"/></g>",
48            "type": "SvgSelector"
49          }
50        }
51      ],
52      "type": "Annotation",
53      "id": "s-1562775288631"
54    }
55  ]
56 }
```

# Texte auf GitHub - 1

---

## Lokales XML Schema

- `<fold>`  
mit `@xml:id` und recto/verso-Paar
- `<div>` (division)  
mit `page`, `article`, `image`, `advert`, `other`, `head` (for running head on top of page), `margin` (for margin)
- `<p>` (paragraph)
- `<lb/>` (line break) / `<pb/>` (page break)

## Encoding “UTF-8”:

- Zeichen nach Vorlage, wo immer Unicode codepoint verfügbar
- Unlesebar: `&gaiji;` (“`&#xe111;`“)
- Alle Schriftzeichen als double-space chars

# Texte auf GitHub - 2

---

## Besonderheit: Laufrichtung des Textes

`<div>` mit `@mode` und `@dir`

**Default:** `mode="vertical-rl"`

```
<div type='other' mode="vertical-rl">
  <div type='advert'>
    <!-- first contents -->
  </div>
  <div type='advert' mode="horizontal-tb" dir='ltr'>
    <!-- second contents -->
  </div>
  <div type='advert'>
    <!-- third contents -->
  </div>
</div>
```

<https://github.com/exc-asia-and-europe/ecpo>

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <?xml-model href="https://raw.githubusercontent.com/exc-asia-and-europe/ecpo/master/ecpo.rng" type="application/x
3 <!DOCTYPE ecpo [
4     <!ENTITY gaiji "&#xe111;">
5 ]>
6 <fold xml:id="jb_3795_1939-04-21_0001+0004">
7     <!-- start recto page -->
8     <div type="page" page="recto" n="1" mode="vertical-rl" dir="rtl">
9         <div type="head" dir="rtl" mode="horizontal-tb">
10             <div>
11                 晶報<lb/>
12             </div>
13             <div dir="ltr" mode="horizontal-tb">
14                 The Crystal<lb/>
15                 發行人 A. L. Teodoro<lb/>
16             </div>
17             <div>
18                 館址漢口路二九九號 電話九三四二八<lb/>
19                 中華郵政特准掛號立券按照總包優益寄送之報紙<lb/>
20                 ◀公共租界工部局登記證C字廿四號▶<lb/>
21                 ▶第三千七百九十五號◀<lb/>
22                 中華民國貳拾捌年四月廿一日 星期五 今日兩全張售國幣貳分四厘 本報創刊於民國八年三月三日<lb/>
23             </div>
24         </div>
25         <div type="advert">
26             <div>
27                 煙中<lb/>
28                 鐵軍<lb/>
29                 清香<lb/>
30                 雋永<lb/>
31             </div>
32             <div mode="horizontal-tb" dir="rtl">
33                 &gaiji;明者&gaiji;<lb/>
34                 &gaiji;不吃虧<lb/>
35             </div>
36             <div type="image"></div>
37             <div type="image"></div>
```

**Wege zum Volltext:**

**Machine learning**

# Impresso und dhSegment

---

DH2019 (Utrecht): Kontakt impresso

Folgegespräche insb. mit Sofia Oliveira

-> Einsatz dhSegment: “generic deep-learning framework for Historical Document Processing” <https://github.com/dhlab-epfl/dhSegment>

## Ziele

- Unterstützung bei Document segmentation
- Trainieren der Modelle auf Unterscheidung 3 Hauptbereiche
  - Paratext (Kopfzeilen/running head und Marginalia)
  - Werbung und Bilder
  - Textbereiche
- Unterstützung der Maschine durch Entfernung der Bereiche mit Spezialfonts und Bildern (Abarbeitung einzeln)

# Paratext

---

## Erster Schritt

- Weitgehend regelmäßig
- Nur geringe Unterschiede im Text (zb. Seitenzahlen, Datum)
- Daten können für Metadaten genutzt werden
  - Editors, Publisher
  - Identifikation der Seiten
- Sonderfonts und Layout behindern Seiten- und Fonterkennung

## Ziele

- Netzwerk trainieren auf Erkennung Header in ganz Jingbao (in Arbeit)
- Nutzung zum OCR Training (begrenzter Satz an Zeichen)
- Wenige unregelmäßige Zeichen und Fonts

# Illustrationen und Werbung

---

- Werbung meist im Falz, teilweise schlecht erhalten
- Viele Wiederholungen von Ads innerhalb bestimmter Zeiträume
- Derzeit noch kein tracking von Ads zwischen verschiedenen Issues / Publikationen

Herausforderung:

- Zeichen/Fonts extrem variabel – Sonderzeichen, Spezialfonts, invertierte Zeichen, Text-in-Bild-in-Text, etc

Ziele:

- Modell zur Erkennung einzelner Ads
- Über Einzel-Fold hinaus
- Metadaten

# Inhalte / Text

---

## Ideal

- Gruppierung der Sammlung nach Fonts (historisch wurden nur eine sehr begrenzte Anzahl von CJK Fonts genutzt)
- Texte in „natural language“, legt Einsatz von Wörterbüchern für die Verarbeitung nahe

## Aber

- historische Varianten (Zeichen, Wort) nicht in modernen Modellen
- Mischung von Schriften (Hant, Latn) und typische Wechsel in Laufrichtung führen zu Problemen in OCR
- GroundTruth Erstellung sehr zeitaufwändig, Verarbeitung mit extensivem pre-processing (comments, ornamentation, etc)

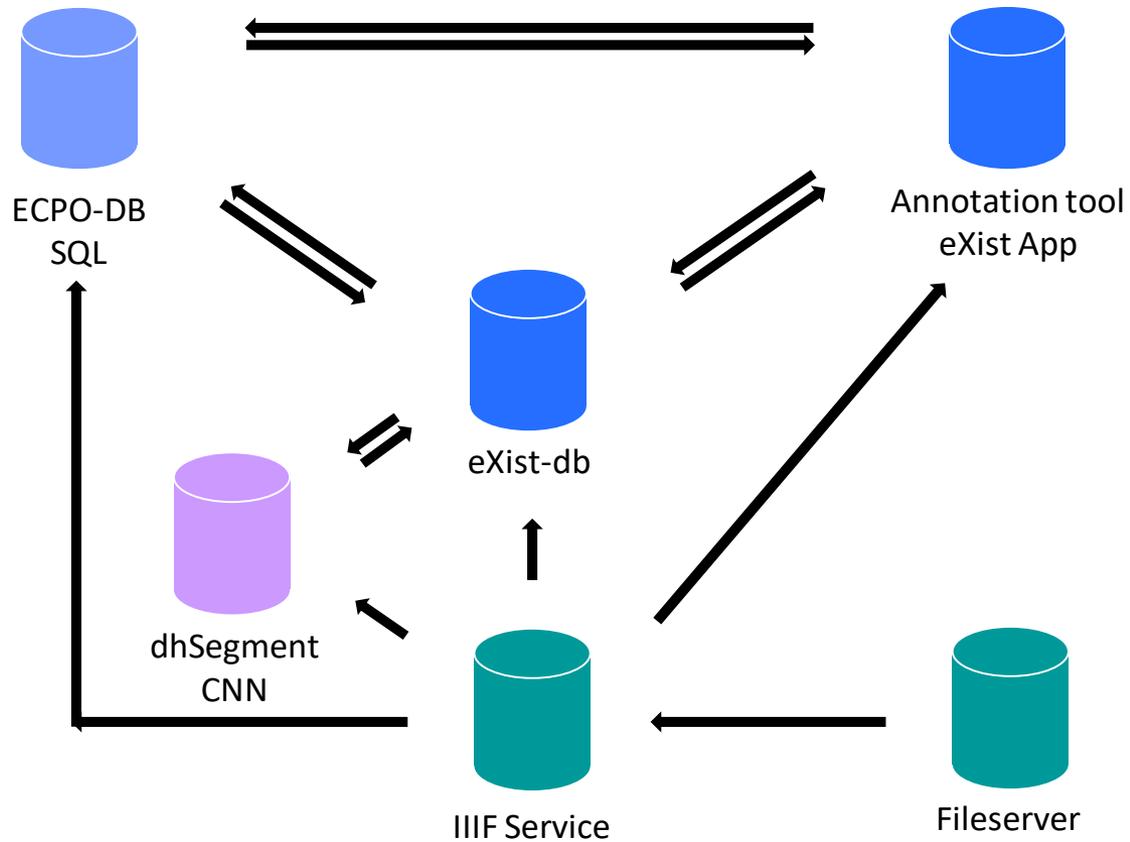
## Ziele

- Verarbeitung einzelner kompletter Jahrgänge und Einbau in ECPO Search
- Ausreichende OCR Qualität für einfache Suchen, incl. Zeichenvarianten

**Wege zum Volltext:**

**Nächste Schritte**

# ECPO – Erweiterte Infrastruktur



# Herausforderungen

---

- Rechenkapazitäten unzureichend für Entwicklung (GW und GPU Server)
- Gewachsene Teilmodule müssen in neuer Infrastruktur integriert werden
  - Unit-ID's, file encodings, etc.
- Outsourced Module teilweise mangelhaft dokumentiert

# Work in progress

---

- Generierung von .png aus Anno-Tool für CNN Training mit GroundTruth
- Verlinkung der Annotationen (manuell und Maschine) mit vorhandenen Annotationen in ECPO-SQL (Metadaten Artikel, Bild, Werbung)
- Stapelverarbeitung der Regionen im Scan nach Typ

## Mittelfristig

- Modell auf Wechsel der Laufrichtung innerhalb semantischer Einheiten trainieren
- Modell für vertikale Zeilenerkennung trainieren (RtL, TtB)
- Erstellung Wörterbuch für Republikzeitliches geschriebenes Chinesisch

## Langfristig

- Ausdehnung des Modells auf andere non-Latn script Materialien
- Verbesserung der CJK OCR für unterschiedliche hist. Perioden

# Contact

---

Matthias Arnold

Heidelberg Centre for Transcultural Studies | HCTS  
Karl Jaspers Centre  
Voßstr. 2 | Building 4400 | Room 005b  
69115 Heidelberg, Germany

Phone: +49 - 6221 - 54 4094  
eMail: [matthias.arnold@uni-hd.de](mailto:matthias.arnold@uni-hd.de)  
Web: <http://tinyurl.com/matthias-arnold>

# References

---

*Women and the Periodical Press in China's Global Twentieth Century: A Space of Their Own?* Ed. by Joan Judge, Barbara Mittler and Michel Hockx, Cambridge University Press, 2018.

Arnold, Matthias, und Lena Hessel. „Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO)“. Heidelberg: Heidelberg University Press, 2019.  
<https://doi.org/DOI:10.11588/heidok.00027325>.

Sung, Doris, Liying Sun, und Matthias Arnold. „The Birth of a Database of Historical Periodicals: Chinese Women's Magazines in the Late Qing and Early Republican Period“. *Tulsa Studies in Women's Literature* 33, Nr. 2 (2014): 227–37.  
<https://doi.org/10.1353/tsw.2014.0004>.

Oliveira, Sofia Ares, Benoit Seguin, and Frederic Kaplan. “dhSegment: A Generic Deep-Learning Approach for Document Segmentation.” *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, August 2018, 7–12.  
<https://doi.org/10.1109/ICFHR-2018.2018.00011>.