



CrossAsia Integrated Textrepository
from data workflows
to service interfaces
to metadata workflows and services

Martina Siebert & Hou-leong (Brent) Ho

The screenshot shows a collection of logos for various digital resources and databases, including:

- cnki 中国知网 (www.cnki.net)
- CiNii Articles
- 大数据 中国政府 文献信息
- Buddhist Digital Resource Center
- JapanKnowledge Lib
- 한국 역사 문화 조사 자료 데이터베이스 (KOREAN HISTORY & CULTURE RESEARCH DATABASE)
- 中華經典古籍庫
- east view
- 皮书数据库
- 人民日报 (People's Daily)
- 雑誌記事索引データベース
- ざっさくプラス
- CEIC
- 法北 北大 (pkulaw.cn)
- WiseSearch
- AM Adam Matthew
- 前報中書館 (The Commercial Press)
- 漢籍電子文獻資料庫
- airiti Library 華藝線上圖書館
- ProQuest
- 万方数据 (WANFANG DATA)
- 沖縄タイムス
- 書同文 古籍数据库产品列表
- KPM
- CHINA DIGITAL LIBRARY

At the bottom, there is a navigation menu with the following items:

- xasia
- Ressourcen
- Service
- Information
- CrossAsia Blog
- CrossAsia Forum
- CrossAsia Classroom
- CrossAsia Regionen
- Über CrossAsia

A prominent red button in the bottom left corner reads: **CrossAsia Standard-Lizenzvertrag**.

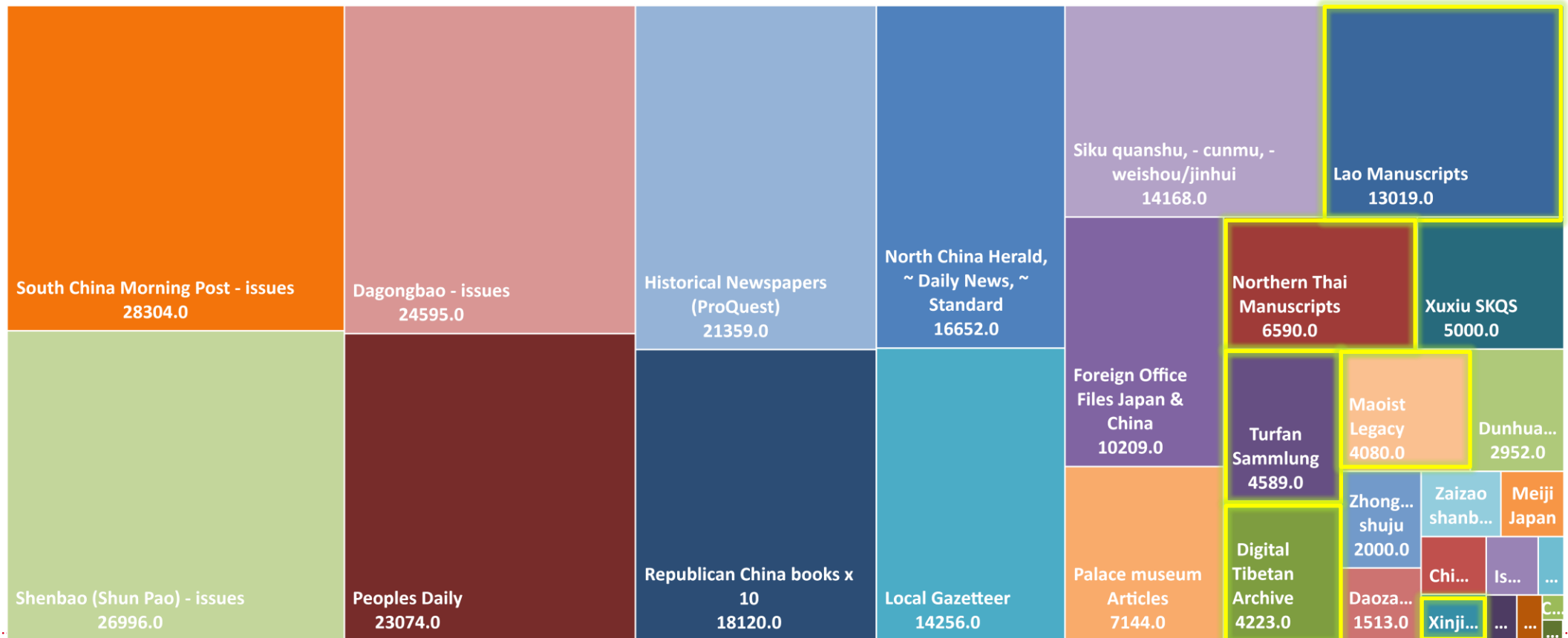


A screenshot of a file explorer window showing a list of files in a directory. The files are named with a unique ID and a .tar.gz extension. The columns are Name, Größe (Size), and Geändert (Modified).

Name	Größe	Geändert
DLJ0010003.tar.gz	1.297 KB	24.07.2020 06:08:51
DLJ0010025.tar.gz	8.997 KB	24.07.2020 06:08:51
DLJ0010067.tar.gz	3.334 KB	24.07.2020 06:08:51
DLJ0010117.tar.gz	8.781 KB	24.07.2020 06:08:52
DLJ0010120.tar.gz	31.253 KB	24.07.2020 06:08:52
DLJ0010133.tar.gz	4.212 KB	24.07.2020 06:08:52
DLJ0010134.tar.gz	29.235 KB	24.07.2020 06:08:53
DLJ0010155.tar.gz	11.003 KB	24.07.2020 06:08:53
DLJ0010162.tar.gz	2.734 KB	24.07.2020 06:08:53
DLJ0010164.tar.gz	51.098 KB	24.07.2020 06:08:55
DLJ0010171.tar.gz	25.375 KB	24.07.2020 06:08:54
DLJ0010186.tar.gz	61.815 KB	24.07.2020 06:08:55
DLJ0010190.tar.gz	13.494 KB	24.07.2020 06:08:55
DLJ0010204.tar.gz	22.002 KB	24.07.2020 06:08:56
DLJ0010217.tar.gz	24.802 KB	24.07.2020 06:08:56
DLJ0010232.tar.gz	26.484 KB	24.07.2020 06:08:57
DLJ0010266.tar.gz	12.321 KB	24.07.2020 06:08:57
DLJ0010270.tar.gz	792 KB	24.07.2020 06:08:57
DLJ0010286.tar.gz	65.807 KB	24.07.2020 06:08:58
DLJ0010301.tar.gz	12.022 KB	24.07.2020 06:08:57
DLJ0010303.tar.gz	1.798 KB	24.07.2020 06:08:57
DLJ0010305.tar.gz	37.486 KB	24.07.2020 06:08:58
DLJ0010308.tar.gz	296 KB	24.07.2020 06:08:58
DLJ0010319.tar.gz	76.009 KB	24.07.2020 06:09:00
DLJ0010322.tar.gz	4.748 KB	24.07.2020 06:08:59
DLJ0010367.tar.gz	16.899 KB	24.07.2020 06:08:59

CROSSASIA ITR "COLLECTIONS"

- North China Herald, ~ Daily News, ~ Standard
- China: Trade and Politics
- CNKI E-Books
- Foreign Office Files Japan & China
- Local Gazetteer
- Meiji Japan
- Republican China books x 10
- Peoples Daily
- Qing history
- Sibu beiyao
- Xuxiu SKQS
- Daozang jiyao
- Zhonghua shuju
- Daozang
- Dunhuang
- Iseas (SOA)
- Taiwan eBooks
- Palace museum Articles
- Lao Manuscripts
- Northern Thai Manuscripts



Resources

- Adam Matthew - Area Studies: China and Southeast Asia
- Adam Matthew - Area Studies: Japan
- Adam Matthew - China America Pacific
- Adam Matthew - China Trade & Politics
- Adam Matthew - Foreign Office Files China
- Adam Matthew - Foreign Office Files Japan
- Adam Matthew - Meiji Japan (only metadata)
- Airiti ebooks
- Asian Studies (ISEAS Publishing)
- China Comprehensive Gazetteers : 中國綜合方誌庫
- The Chinese Student Monthly (1906-1931)
- CNKI eBooks
- 道藏輯要
- 敦煌史料
- 古今圖書集成 : Qing Imperial Encyclopedia
- The Japan Chronicle (1902-1940)
- Local Gazetteers
- 民國圖書數據庫 : Early Twentieth Century Book in China (1912-1949)
- Missionary, Sinology, and Literary Periodicals (1817-1949)
- Mobilizing East Asia (1931-1954)
- North China Daily News
- North China Herald
- North China Standard (1919-1927)
- 清代史料
- Records of the Maritime Customs Service of China (1854-1949)
- 人民日报 : People's daily
- 日本古典書籍 : Classical Works of Japan
- 四部備要
- 四部叢刊
- 四庫全書
- Ta Kung Pao 大公報 (1902-1949)
- 續修四庫全書
- 永樂大典
- 正統道藏
- 中國地方誌 一集
- 中國地方誌續集
- SBB digital : Western language Asia collection
- SBB digital : Asian language collection (selection)
- Fulltext search in print books (sample set)
- BuddhistRoad Papers (Project)
- The Maoist Legacy Database (Project)

fulltext of 367,000 titles with 56 Mio. pages



CrossAsia Integrated Textrepositary

East Asia Department /
CrossAsia / IT department

IT department / CrossAsia

IT department

data
(metadata, image +
text data)

- evaluating
- concatenating
- harmonising
- transforming

admin access

collection
ingest

identifier

Fedora 4
storage

metadata

- article + chapter titles
- non-permanent, subscribed resources
- suggestions for acquisition suggestions (PDA)

metadata
in PICA format

metadata
in PICA format

CBS (VZG/GBV)

- secure data hosting
- MARC21 export
- Discovery index

SBB CBS (VZG/GBV)

- secure data hosting
- MARC21 export
- Discovery index

>> **New CrossAsia Search (Discovery)**

New CrossAsia Search (Discovery) <<

original metadata

title_id	publication_title	title_url	publisher_name	Place	first_author	Remarks	version
xuxiu-01-01-001	馬王堆帛書周易經傳釋文	http://huntenq.com/anc	GET-HONG COMMUN	taiwan	廖名春釋文	續修四庫全書-經部-易類	
xuxiu-01-01-002	周易鄭注	http://huntenq.com/anc	GET-HONG COMMUN	taiwan	[宋]王應麟輯[清]丁	續修四庫全書-經部-易類	[清嘉慶二十四年蕭山陳氏刻湖海樓叢書本]
xuxiu-01-01-003	敦煌周易殘卷	http://huntenq.com/anc	GET-HONG COMMUN	taiwan	[魏]王弼注	續修四庫全書-經部-易類	[法國國家圖書館藏卷]

dcterms_identifier	dc title	mods_responsibility	dcndI_titleTrdcndI_titleI dc_creato xxx	Xxx
Diaolong_xuxiu-01-01-001	馬王堆帛書周易經傳釋文	廖名春釋文	ma wang dui 馬王堆帛書周易經傳 廖名春	Liao M
Diaolong_xuxiu-01-01-002	周易鄭注	[宋]王應麟輯[清]丁傑後定[清]張	zhou yi zhen 周易鄭注 王應麟 王應麟Sp宋	Wang Y
Diaolong_xuxiu-01-01-003	敦煌周易殘卷	[魏]王弼注	dun huang zh 敦煌周易殘 王弼注 王弼注Sp魏	Wang t
Diaolong_xuxiu-01-01-004	關氏易傳	題[後魏]關朗撰[唐]趙蕤註	guan shi yi zh 关氏易传 關朗 關朗Sp後魏	Guan L
Diaolong_xuxiu-01-01-005	周易經典釋文殘卷	[唐]陸德明撰	zhou yi jing c 周易经典释 陸德明 陸德明Sp唐	Lu Den
Diaolong_xuxiu-01-01-006	周易正義	[唐]孔穎達撰	zhou yi 明範氏天一閣刻本 1368###1644	明 [1368-1644] 明
Diaolong_xuxiu-01-01-007	周易注疏	[魏]王弼注[晉]韓康伯注[唐]孔穎	zhou yi 法國國家圖書館藏卷	
Diaolong_xuxiu-01-01-008	易經解	[宋]朱長文撰	yi jing ji 清嘉慶阮刻十三經注疏本 1796###1820	清嘉慶 [1796-1820] 清嘉慶
Diaolong_xuxiu-01-01-009	周易新講義	[宋]龔原撰	zhou yi 阮刻本	
Diaolong_xuxiu-01-01-010	晦庵先生校正周易系辭	[宋]呂祖謙編	hui an x 明崇禎四年刻本 1631	明崇禎四年 [1631] 明崇禎
Diaolong_xuxiu-01-01-011	古易音訓	[宋]呂祖謙撰[清]宋鹹熙輯	gu yi yin 日本文化五年活字印佚存叢書本 1808	日本文化 [1808] 日本文化

preparation for Solr
preparation for PICA

improve metadata

original delivery

- 01-01-001/ 2016-10-15 07:04
- 01-01-002/ 2016-00001.xml 2016-1
- 01-01-003/ 2016-00002.xml 2016-1
- 01-01-004/ 2016-00003.xml 2016-1
- 01-01-005/ 2016-00004.xml 2016-1
- 01-01-006/ 2016-00005.xml 2016-1
- 01-01-007/ 2016-00006.xml 2016-1
- 01-01-008/ 2016-00007.xml 2016-1
- 01-01-009/ 2016-00008.xml 2016-1
- 01-01-010/ 2016-00009.xml 2016-1
- 01-01-011/ 2016-00010.xml 2016-1
- 01-01-012/ 2016-00011.xml 2016-1
- 01-01-013/ 2016-00012.xml 2016-1
- 00013.xml 2016-1

```

<-region type="vtext" attr="" line="65535" direction="" id="119947" pos="20,20,2462,3490" FormatRange="20,20,2
<-row id="1" type="">
  <mark id="119951" type="title.start" value="馬王堆帛書周易經傳釋文" param="1" kind="TITLE"/>
  三三鍵元亨
  <mark id="119951" type="title.end"/>
  . 利貞. 初九. 滯龍勿用. 九二. 見龍在
  </row>
<-row id="2" type="">子終日鍵鍵. 夕沂若厲
  <-row id="3" type="">
  見大人. 尚九抗龍有慶. 遇九. 見羣龍无
  <mark type="small.start" kind="ATTRIBU
  一行
  <mark type="small.end"/>
  </row>
<-row id="4" type="">三三婦之非人. 不利.
<-row id="5" type="">二抱承小人吉大人不亨
<-row id="6" type="">休婦. 大人吉. 六亡.
<-row id="7" type="">三三掇. 亨. 小利貞.
<-row id="8" type="">莫之勝奪. 九三. 為擄
  
```

Extract fulltext (with all additional information it contains)

三三鍵元亨. 利貞. 初九. 滯龍勿用. 九二. 見龍在田. 利見大人. 九三. 君子終日鍵鍵. 夕沂若厲. 九四. 或鑿在滯. 无咎. 九五. 聖賢在天. 利見大人. 尚九抗龍有慶. 遇九. 見羣龍无首. 不利. 君子貞. 大往小來. 初六友蓍茹. 以六寶. 貞吉. 亨. 六二抱承小人吉大人不亨六三三三婦之非人. 不利. 小人吉. 大人不亨. 六三. 掇. 亨. 小利貞. 用有攸往. 六二. 共之用黃牛之勒. 莫之勝奪. 九三. 為擄. 有疾. 厲. 畜僕妾. 吉. 九四. 好五. 嘉掇. 貞吉. 尚九. 肥掇. 无不利. (三行) 三三禮虎尾. 不真人. 亨. 初九. 錯禮. 往无咎. 人貞吉. 六三. 能視. 跛能利禮虎尾. 真人. 兇. 武人遇于大君. 九四禮虎尾. 朔期. 終吉. 九. 視禮. 巧. 貞吉. (四行) 三三訟. 有復誦寧. 克吉. 冬. 兇. 利用見大人. 不利涉大川. 利言吉. 冬. 九二. 不克訟. 歸而逋. 元邑人三百戶. 无首. 六三. 食德. 小人. 六三. 不克訟. 歸而逋. 元邑人三百戶. 无首. 六三. 食德. 小人.



Building data expertise within the team

China

Martina Siebert

- mass data from Solr/XML
- MARC-Daten



Catmandu scripts

China

Wang Xiaoming

Wang Bin

- historical eBooks
- XLS data, JSON data



OpenRefine
Breits „Knitting machine“

China

Franziska Trempler

- modern eBooks
- article data
- XLS data, JSON data



OpenRefine
OpenRefine Templating

Japan

Munkhtuul Schwedes

- PDA (China vendors)
- XLS data



OpenRefine
Breits „Knitting machine“

China/IT

Brent Ho

- data harvesting
- data routines
- „Toolz2 (date transformation, Pinyin, „Knitting machine“ etc.)

IDM

Marco Remy

- data harvesting routines
- CrossAsia Solr
- („old“) CrossAsia Search

Japan

Kathrin Weckerle

PICA format guardian



Project cooperation and support

CrossAsia :
Digital Collections
& Theme portals

Sie befinden sich hier: CrossAsia.org > Service > IIF

CrossAsia Digital Library of Northern Thai Manuscripts



Digital Library of Northern Thai Manuscripts (Lanna)
The Digital Library of Northern Thai Manuscripts is designed as a resource for the study of traditional
It assembles four digital collections:


CrossAsia Digital Library of Lao Manuscripts



Digital Library of Lao Manuscripts (DLLM)
The Digital Library of Lao Manuscripts (DLLM) makes images of about 13,000 texts from through
Staatsbibliothek zu Berlin Preußischer Kulturbesitz. The project is funded by the German Res

Sie befinden sich hier: CrossAsia.org > Service > IIF

CrossAsia Turfan



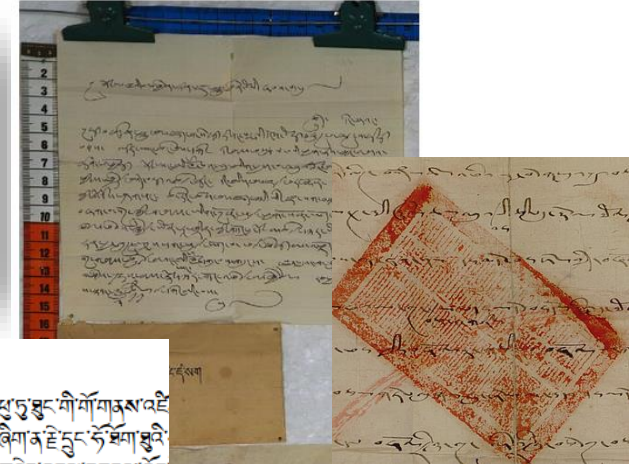
Turfan Sammlung: Akten, Fotos, Zeichnungen
Zwischen 1902 und 1914 fanden unter der Leitung von Albert Grünwedel

MADOC
digirati

xasia

in progress

**Digital Tibetan Archives
Bonn (DTAB)**



Main Text:

ཕུ་ཏུ་མུང་གི་གོ་གནས་འཛི
ཞིག་ནི་རྒྱུ་རྩེ་གི་མུ་ལུ་
གཞིར་འཇམ་དབུང་ས་འོད
བསྐྱེད་པ་དགེ་ལེགས་ བ

མི་ལུ་ཚེད་མངགས་གྱི་རྟེན་རྩམ་འབུལ་བར་ཡི་ཅིང་དུ་རྟོག་ཐོན་ལུ་
དགོས་སོང་བས་ལྷ་ལྷན་ནས་ དར་མདོ་བར་དགོས་བྱེད་བཞེན་རྟེ
བརྒྱུད་ དོས་ལལ་མ་ལུ་བརྩུང་གཞིས་ལང་དར་མདོ་བར་དོས་སྟེ
མི་གསུམ་ལ་ཡར་བྱེད་བཞེན་རྟེ་གཞིག་ བཞེན་པ་གཞིས་ལལ་མ་
བརྩུ་ཐམ་པ་བཅས་གཏོང་ལུགས་ལུ་དགོས་མ་ཟད་དར་མདོ་ནས་ཡི་ཅི

**Esche
collection**

0000002.tif 	0000003.tif 	0000004.tif 	0000005.tif 
0000011.tif	0000012.tif	0000013.tif	0000014.tif



Focus = support of academic research (now and future): acquisition of printed and electronic material & cooperation and support/hosting of academic projects

integration into **CrossAsia Fulltext Search**
and **ITR services**

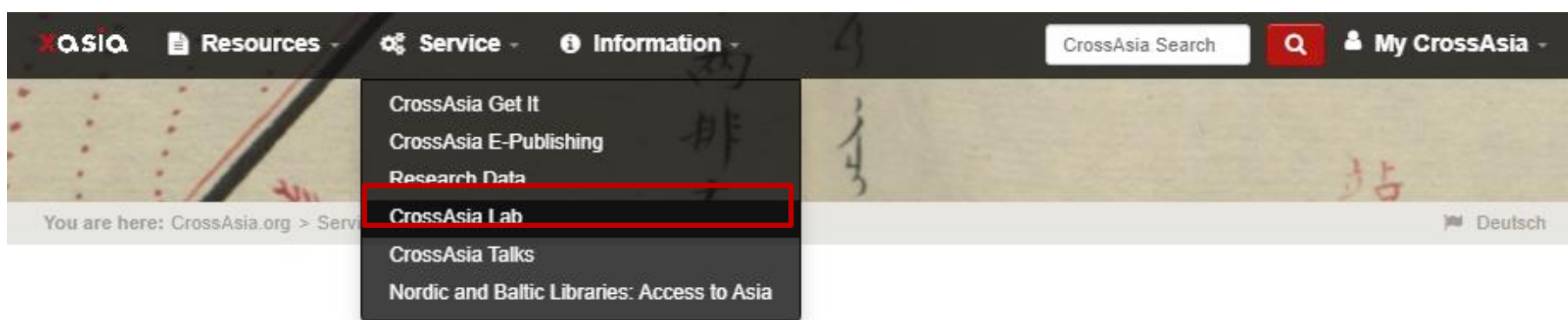
- **BuddhistRoad Papers** (Project)
- **The Maoist Legacy Database** (Project)
- Digital Tibetan Archives Bonn (Project)

database hosting
hosting of **data**
CrossAsia ITR

>> generic presentation and integration
into CrossAsia ITR services

- **Digital Tibetan Archives Bonn** (Prof. Schwieger)
from database hosting to transfer to MADOC
presentation and CrossAsia Search Services
- **Chinese Historical Healthcare Manuscripts-
Database** (Unschuld, Prackwieser, Kirk, Otte)
- **Collection Esche** > presentation in MADOC
- **Collection Kahlen** > presentation in MADOC

Services build upon CrossAsia Integrated Textrepository



CrossAsia LAB <https://crossasia.org/en/service/crossasia-lab/>

In the context of CrossAsia, various IT applications and tools have been created and are available free of charge in the CrossAsia LAB. These tools include N-gram data sets, a transliteration tool for selected Central Asian scripts and additional search functions for our full texts in the CrossAsia ITR (Integrated Text Repository). We invite all users and visitors to take advantage of these features and welcome your feedback at x-asia@sbb.spk-berlin.de. As part of the Digital Humanities, we are happy to realise new tools for you and your projects to support you in your research.

CrossAsia Fulltext Search



The CrossAsia Fulltext Search with its two Beta versions - a "guided" and an "explorative" fulltext search - is built from textual resources hosted in the CrossAsia Integrated Textrepository (ITR). Its aim is to help users find texts and sources relevant to their research

CrossAsia ITR Explorer



The CrossAsia ITR Explorer provides a different perspective on the CrossAsia ITR resources and is supplementary to the CrossAsia Fulltext Search. It allows to combine and compare search results and visualise them showing their overlaps and distribution over time.

Services build upon CrossAsia Integrated Textrepository

- ITR full text search
- ITR Explorer
- ITR Newspaper Explorer
- N-gram service

CrossAsia LAB

In the context of CrossAsia, various IT applications and tools have been created and are available free of charge in the CrossAsia LAB. These tools include N-gram data sets, a transliteration tool for selected Central Asian scripts and additional search functions for our full texts in the CrossAsia ITR (Integrated Text Repository). We invite all users and visitors to take advantage of these features and welcome your feedback at x-asia@sbb.spk-berlin.de. As part of the Digital Humanities, we are happy to realise new tools for you and your projects to support you in your research.

CrossAsia Fulltext Search



The CrossAsia Fulltext Search with its two Beta versions - a "guided" and an "explorative" fulltext search - is built from textual resources hosted in the CrossAsia Integrated Textrepository (ITR). Its aim is to help users find texts and sources relevant to their research questions by performing comprehensive fulltext searches across databases.

CrossAsia ITR Explorer



The CrossAsia ITR Explorer provides a different perspective on the CrossAsia ITR resources and is supplementary to the CrossAsia Fulltext Search. It allows to combine and compare search results and visualise them showing their overlaps and distribution over time.

CrossAsia N-gram Service



The full-texts stored in the CrossAsia ITR provide an excellent source for all who are interested in digital humanities and in statistical, linguistic and other computational analysis. We prepared the texts in a way that they can be downloaded as unrestricted N-gram datasets.

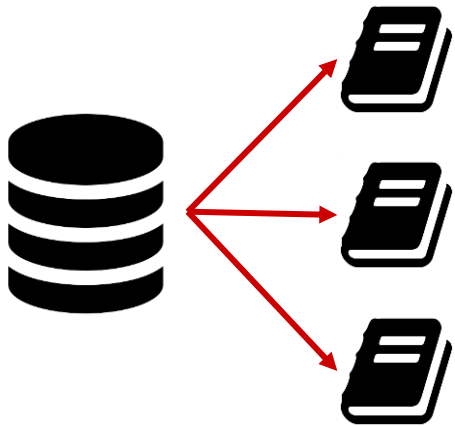
CrossAsia ITR Newspaper Explorer



The CrossAsia ITR Newspaper Explorer is an addition to the CrossAsia ITR Explorer that makes full use of the fine-granular data nature of newspapers. In addition to a Venn-diagram visualisation it provides "heatmaps" of the time distribution of search terms or combined search sets at different scales, from decade, year, month to day level.

General ranking mechanism in library system

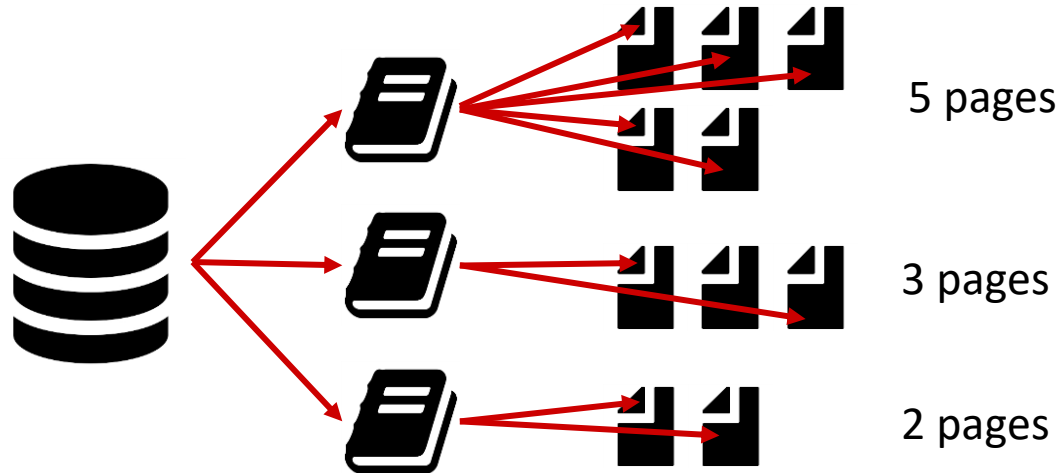
- General ranking mechanism in library system



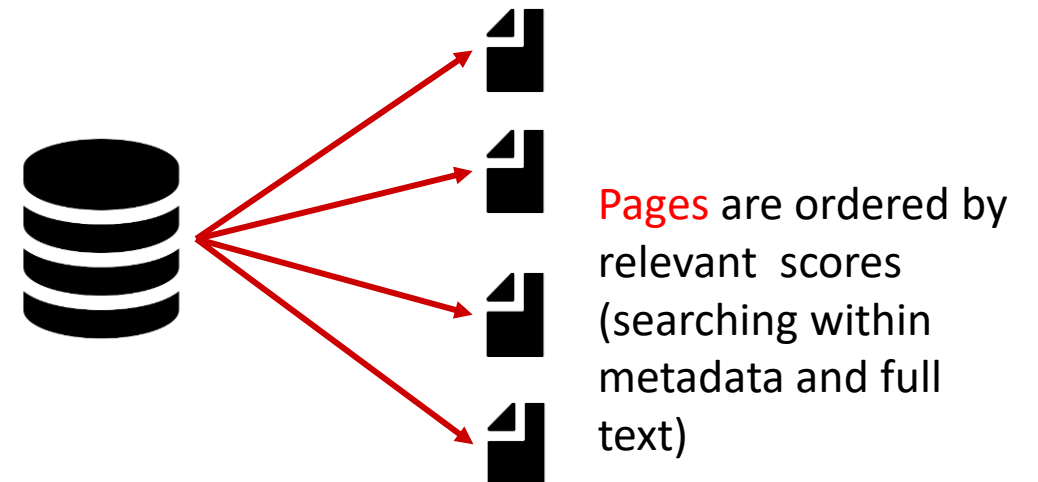
Titles are ordered by relevant scores (searching within Bibliograph metadata, such as **title, author, keyword, publisher, ... etc.**)

New ranking mechanism in ITR full text search

- Guided full text Search (Titles are ordered by numbers of **page hits**)



- Explorative full text Search (Pages are ordered by relevant scores)



ITR full text search

- Guided full text Search (order by **page hits**)

The screenshot shows the CrossAsia search interface. At the top, there is a navigation bar with 'CrossAsia Suche' and 'Mein CrossAsia'. Below the navigation bar, the search results for 'Cotton' are displayed. The search term 'Cotton' is entered in the search box, and the results are ordered by page hits. The first result is 'Imperial Maritime Customs 1 3 4 (1904)' with 412 pages. The search results are displayed in a list format, with the first result highlighted. The search results are ordered by page hits, and the search term 'Cotton' is highlighted in red. The search results are displayed in a list format, with the first result highlighted. The search results are ordered by page hits, and the search term 'Cotton' is highlighted in red.

- Explorative full text Search (in **page level**)

The screenshot shows the CrossAsia search interface. At the top, there is a navigation bar with 'CrossAsia Suche' and 'Mein CrossAsia'. Below the navigation bar, the search results for 'Cotton' are displayed. The search term 'Cotton' is entered in the search box, and the results are displayed in a list format. The search results are ordered by page level, and the search term 'Cotton' is highlighted in red. The search results are displayed in a list format, with the first result highlighted. The search results are ordered by page level, and the search term 'Cotton' is highlighted in red.

ITR Explorer

CrossAsia ITR Explorer (Beta version)

1. Select source(s) i

2. Enter search term or phrase i

search term or phrase



CJK Mapping

Phrase search

3. Show and combine result sets i

no data yet

4. Visualise your result sets i

VENN (OVERLAP) DIAGRAM

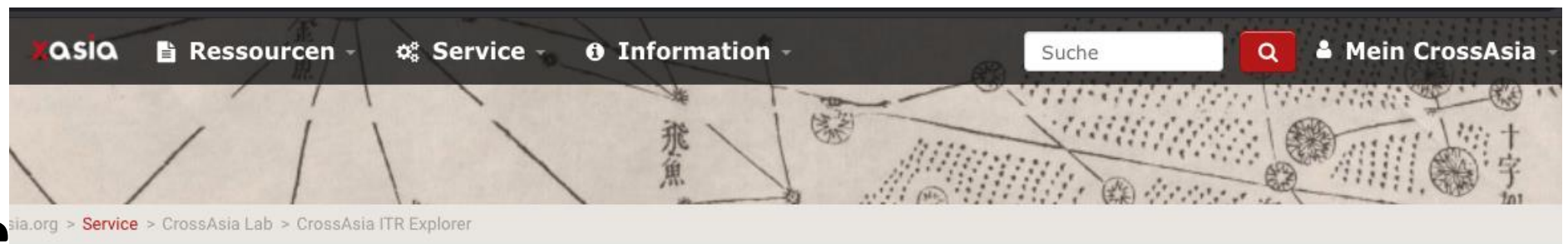
LINE CHART (TIME RANGE)

Select result set(s)

The **CrossAsia ITR Explorer** provides a different perspective on the CrossAsia ITR resources and is supplementary to the **CrossAsia Fulltext Search**. It allows to **combine** and **compare** search results and **visualize** them showing their overlaps or distribution over time. Results refer in general to the book or issue, only for two Chinese newspapers they refer to the article level. Resources are the same as in the CrossAsia Fulltext Search, CJK character mapping is selectable.

See the info i in each section for more details on usage; see a walk-through and some use cases [here](#).

ITR Explore



CrossAsia ITR Explorer (Beta version)

1. Select source(s) ⓘ

Chinese newspapers (old and new)

Ta Kung Pao 大公報 (1902-1929)

人民日報 (1949-2012)

Focus: Republican China

民國圖書數據庫 (1912-1949)

Ta Kung Pao 大公報 (1902-1929)

Modern academic eBooks

Airiti eBooks (licenced)

Asian Studies (ISEAS publishing)

CNKI eBooks (licenced)

The **CrossAsia ITR Explorer** provides a different perspective on the CrossAsia ITR resources and is supplementary to the **CrossAsia Fulltext Search**. It allows to **combine** and **compare** search results and **visualize** them showing their overlaps or distribution over time. Results refer in general to the book or issue, only for two Chinese newspapers they refer to the article level. Resources are the same as in the CrossAsia Fulltext Search, CJK character mapping is selectable.

See the info ⓘ in each section for more details on usage; see a walk-through and some use cases [here](#).

ITR Explorer - More than search

- Search results manipulation (Set operations: AND, OR, NOT)
- Search results (set) observations

Result sets: ✕

factory ✕ poverty ✕ women ✕ factory/poverty ✕ factory-poverty ✕

factory
12208 books

Combine with ...
poverty

NOT

− ∩

poverty
5431 books

Combine with ...

− ∩

women
20176 books

Combine with ...

− ∩

factory ∩ poverty
2569 books

Combine with ...

− ∩

factory-poverty
9637 books

Combine with ...

− ∩

Show overlap of sets. Please select sets for visualization.

women ✕ poverty ✕ factory ✕

(Set "factory" ∩ Set "poverty" ∩ Set "women") : 2213 books that is
 10 % of the set "women"
 40 % of the set "poverty"
 18 % of the set "factory"

VENN (OVERLAP) DIAGRAM LINE CHART (TIME RANGE)

Show sets over time, define time range. Please select sets for visualization.

cotton [set 1] ✕ women [set 2] ✕ poverty [set 3] ✕

From 1688-JAN to 1950-JAN ✳

• cotton [Set 1]'s sources : Missionary and Sinological Journals (1817-1949), SBB digital collection (Western language texts), North China Daily News (1864-1941/1946-1950), North China Standard (1919-1927), Japan Chronicle (1900-1941), Chinese Student's Monthly (1906-1931), North China Herald (1850-1940/1950)

• women [Set 2]'s sources : Missionary and Sinological Journals (1817-1949), SBB digital collection (Western language texts), North China Daily News (1864-1941/1946-1950), North China Standard (1919-1927), Japan Chronicle (1900-1941), Chinese Student's Monthly (1906-1931), North China Herald (1850-1940/1950)

• poverty [Set 3]'s sources : Missionary and Sinological Journals (1817-1949), SBB digital collection (Western language texts), North China Daily News (1864-1941/1946-1950), North China Standard (1919-1927), Japan Chronicle (1900-1941), Chinese Student's Monthly (1906-1931), North China Herald (1850-1940/1950)

- Downloadable metadata

Titles in selection ((Set "factory" ∩ Set "poverty" ∩ Set "women") : 2213 books)

Title	Sets	Date ↓	Creator	Subject	CrossAsia
The Rev George Smith DD Bishop of Victoria Hong Kong	● ● ●	n.d.	Headla...	Hong K...	↗
Bookshelf The English-language Press Networks of East Asia 1918-1945	● ● ●	2010-01-01			↗
澳門現代化進程與城市規劃	● ● ●	2007	楊允中	社會科學...	↗

ITR Newspaper Explorer

Explorer focus on timespan

Year

Month

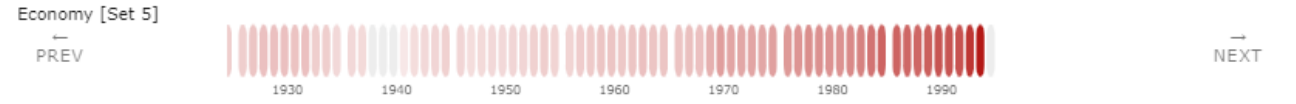
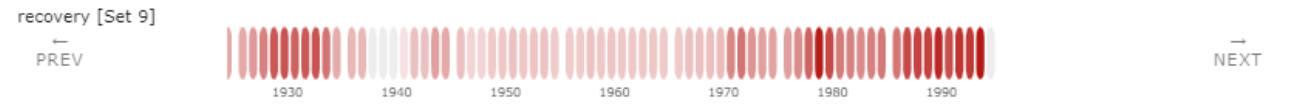
Date

4. Visualise your result sets

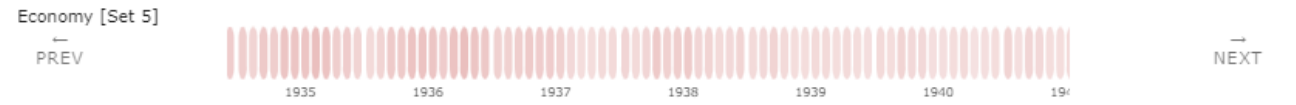
VENN (OVERLAP) DIAGRAM

HEAT MAP (CALENDAR)

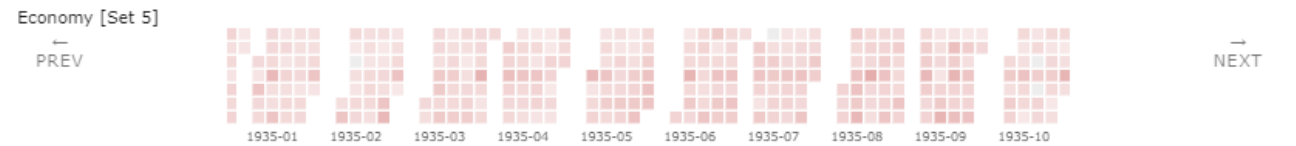
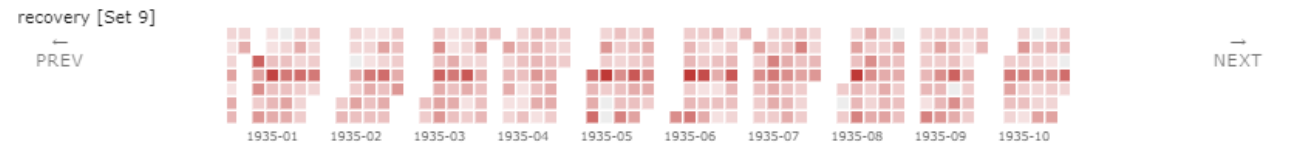
Select result set(s)
Economy [Set 5] recovery [Set 9]



decade - year decade - year
year - month year - month



year - month year - month
month - day month - day



CrossAsia N-gram Service

The full-texts stored in the CrossAsia ITR provide an excellent source for all who are interested in digital humanities and in statistical, linguistic and other computational analysis. Access to the licensed full texts is only available for registered users. By preparing the texts as N-gram datasets, i.e. splitting the texts into fragments - for Chinese texts into fragments with one, two or three character combinations - and presenting only the frequency of the respective fragments in the corpus, they can be downloaded as unrestricted N-gram datasets.

You can download the various datasets and explore them on your own computer and with your own tools. Currently, three datasets of Chinese text collections with N-grams per book are available, each with uni-, bi-, and trigrams.

If you are interested in N-gram sets of any other resources from the CrossAsia context, [please contact us!](#) These sets will then also be made available via this page.