



xasia

CrossAsia Integrated Text Repository (ITR) Ideas & Scope

Our Background at Staatsbibliothek zu Berlin:

Collection Development and the challenges of digitisation

long library tradition since 1661 and collection history for Asiatica

after the Second World War: nationwide special collection plan (Sondersammelgebietsplan, SSG) for literature published abroad

to ensure that every scientifically relevant work from abroad was present and available in Germany in at least one copy

from 1951 to 2015, the Staatsbibliothek zu Berlin was responsible for the East Asia Special Collection (SSG 6.25)

since 2006: CrossAsia starts as service

fundamental evaluation of the DFG's funding program for special subject collections



in the third funding phase as the Special Information Service Asia (Fachinformationsdienst Asien, FID Asien) since 2016

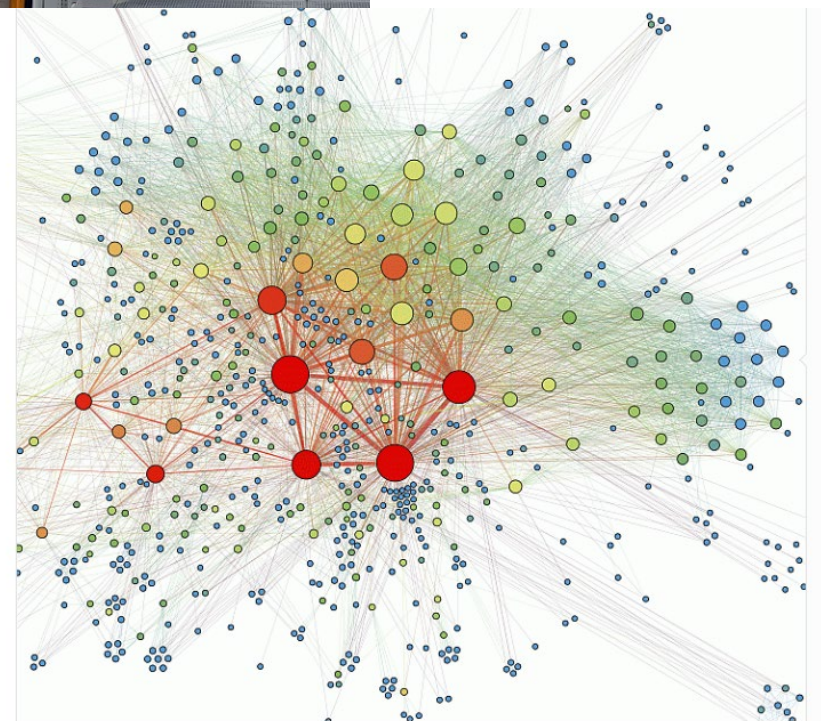
Our collections are partly financed within different projects



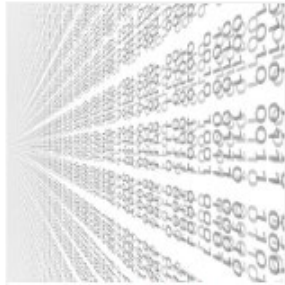
Today?
organizing access /
licensing



Yesterday ?



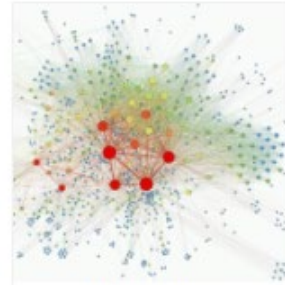
Tomorrow ?!



Data management



Search



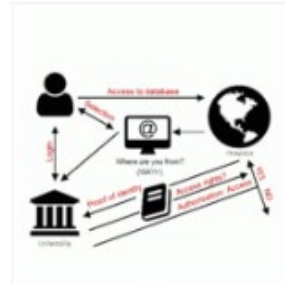
Digital Humanities



Licensing



FID Asia



Authentication



Media supply



Cataloguing



Digitisation





Licensing

CrossAsia and Staatsbibliothek zu Berlin have more than twelve years of hands-on experience in licensing of electronic resources. Today the Library provides nationwide remote access to a wide range of licensed **Asia-related databases**. From the beginning, licensing had the goal to support the German-based researchers within the scope of the DFG-funded Special Research Collection Programme (SSG) and subsequently the Specialised Information Service Programme (FID). In addition, individual European consortia have already been realised. The goal is to further internationalise CrossAsia's licensing activities.

[Licensing](#)

Licensing as integral part of collection building and development

In accordance with our national responsibility, we require and follow specific licensing policies and frameworks:

- e-preferred strategy
- access to licensed content beyond Staatsbibliothek zu Berlin - access to all researchers with an affiliation to a German university etc.
- metadata with stable identifiers and linking
- archiving and hosting rights
- text and datamining rights

Integrated Text-Repository – maintain & store licensed content in Berlin

Beyond Reading:

What comes after reading?

One of our answers:

In line with our licenses –

- managing images and text data
- storing data / content storage
- securing content and data
- metadata management content
- data management
- data control
- Follow **FAIR** principles
- design new data-based services

418,000 titles with 67,2 million pages

- [Area Studies: China and Southeast Asia](#) (Adam Matthew Digital)
- [Area Studies: Japan](#) (Adam Matthew Digital)
- [Airiti ebooks](#)
- [Asian Studies \(ISEAS Publishing\)](#)
- [China America Pacific](#) (Adam Matthew Digital)
- [China Comprehensive Gazetteers : 中國綜合方誌庫](#)
- [China Trade & Politics](#) (Adam Matthew Digital)
- [The Chinese Student Monthly \(1906-1931\)](#)
- [CNKI eBooks](#)
- [道藏輯要](#)
- [敦煌史料](#)
- [Foreign Office Files China](#) (Adam Matthew Digital)
- [Foreign Office Files Japan](#) (Adam Matthew Digital)
- [古今圖書集成 : Qing Imperial Encyclopedia](#)
- [Historical Newspapers of China and South China Morning Post \(1832-1998\)](#) (ProQuest)
- [The Japan Chronicle \(1902-1940\)](#)
- [Local Gazetteers \(Erudition\)](#)
- [民國圖書數據庫 : Early Twentieth Century Book in China \(1912-1949\)](#)
- [Meiji Japan \(only metadata\)](#) (Adam Matthew Digital)
- [Missionary, Sinology, and Literary Periodicals \(1817-1949\)](#)
- [Mobilizing East Asia \(1931-1954\)](#)
- [National Palace Museum periodicals \(Airiti\)](#)
- [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 \(Project\)](#)

Integrated Text-Repository – manage & archive licensed content in Berlin

```
namespace_msi.txt - Editor
Datei Bearbeiten Format Ansicht Hilfe
dcterms http://purl.org/dc/terms/
dc http://purl.org/dc/elements/1.1/
dcndl http://ndl.go.jp/dcndl/terms/
mods http://www.loc.gov/mods/v3
schema http://schema.org

DCNDL DOC: https://iss.ndl.go.jp/information/wp-content/uploads/2021/02/dcndl_simple_format_ver.2.0_20210104.pdf
*https://dl.ndl.go.jp/view/download/digidepo_8723210_po_ndl-term201112_en.pdf?contentNo=1

+Bibliographische Metadaten+

<dcterms:title> bzw. <dc:title>
<dcndl:titleTranscription> (Umschriften, Kurz-/Langzeichen etc.)
<dcterms:alternative> (Zusatz-, Paralleltitel, übersetzter Titel. Letzteres vereinfacht von <mods:title type="translated">). (4002)
<dcndl:otherName> をも見よ参照(別名) (4212 Vorgänger- NachfolgerTitel bei Zeitschriften etc.)

<dcterms:creator> bzw. <dc:creator>
<dcndl:creatorTranscription> (entsprechend dcndl:titleTranscription, hier sowohl für creator als auch contributor, da nur für Suche)
<dcterms:contributor> bzw. <dc:contributor> (weitere beteiligte Personen)
<mods:responsibility> 245 $h (verkürzt von <mods:note type="statement of responsibility"> Xuxiu-Diaolong: macht es so: note_tesim":["type=\"statement of res
[andere Möglichkeit komplette bibliographische Angaben abzulegen: <dcterms:bibliographicCitation>. Verwenden wir jetzt für Angaben zur Datenbank]

<dcterms:issued> (Publikationsdatum; "Date of formal issuance (e.g., publication) of the resource") > PP2 date
<schema:datePublished> Oliver hat issued als "Ganzzahl" definiert; deshalb musste hier für Artikel mit genauerem Datum ein neues Feld her ...
<dcterms:date> bzw. <dc:date> (Datum im Zusammenhang mit Objekt, z.B. Fertigstellung eines Buchs im Ggs. zu seiner Publikation - falls anwendbar. "A point or
<schema:startDate>
<schema:endDate>
<dcndl:publicationPlace>
<dcterms:publisher> bzw. <dc:publisher>
<dcndl:publisherTranscription> *
<schema:ISBN> (war <dcterms:identifier xsi:type="ISBN">, unnötig umständlich)
<schema:ISSN>
<dcterms:extent> (Umfang der Ressource)
<dcterms:format> (verschiedene z.B. 繁体 简体: 410 頁)
```




Integrated Text-Repository – manage & archive licensed content in Berlin

CrossAsia N-gram Service

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.

Here you can download 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. In addition, we will publish further datasets as well as some online services for analysing the N-grams, soon.

N-gram datasets



| Collection | Description | N-gram |
|---------------------------|---|---|
| Xuxiu Siku Quanshu 續修四庫全書 | 'Sequel to the Siku quanshu' of the late 18th century with more than 5,000 titles. |  |
| Local Gazetteers | Chinese geographical works from Tang dynasty to the Republican Era with about 8,000 titles. |  |
| Daozang Jiyao 道藏輯要 | 'Essentials of the Daoist Canon', collection of Daoist texts with about 300 titles. |  |

The datasets are released under a [Creative Commons Attribution 4.0 International License \(CC BY 4.0\)](#).

Recommended citation: Please follow the suggestions provided in the repository.














Integrated Text-Repository – manage & archive licensed content in Berlin

« 1 2 3 4 5 6 7 8 9 »

opium  

About 99k pages founds in 25.4k titles in CrossAsia full-text search

List of titles

- China International ... (484/838 pages)  
- China Opium and drug... (327/615 pages)**  
- Imperial Maritime Cu... (253/927 pages) 
- Imperial Maritime Cu... (235/972 pages) 
- Imperial Maritime Cu... (234/798 pages) 
- Imperial Maritime Cu... (233/884 pages) 
- The Maritime Customs... (232/960 pages) 
- The Maritime Customs... (230/912 pages) 
- China Kwangtung opiu... (224/280 pages)  
- Imperial Maritime Cu... (223/664 pages) 



« 1 2 3 4 5 6 7 8 9 »

China Opium and drug traffic (1917-1922). FO 671/452, The National Archives.



Collection:Foreign Office Files China

327 of 327 pages of this title contain the search term(s) "opium"



« 1 2 3 4 5 6 7 8 9 »

p.103  



of the Company's Fleet, Dear Sirs, **OPIUM** SMUGGLING. With a view to preventing the smuggling.... of **Opium** or ensuring that if **Opium** is on board it is discovered and handed over.... of the **Opium** and we wish every encouragement given to informers, whose identity must.... not be disclosed. Any **Opium** found on board is to be handed over to the Captain who will claim.... and no precautions already taken to prevent **Opium** smuggling are to be relaxed. Yours faithfully

p.104  

- 102 I. 2. **OPIUM** SMUGGLING. **Opium** smuggling is a very bad thing to do.... everyone must try to stop it. Everyone must look out for **opium** and if he can find it he.... The Captain will claim the reward for finding **Opium** and every man on board will share.... to find the **Opium**. # 5. The rewards paid by the Customs are on the following scales.... : - CUSTOMS' **OPIUM** SEIZURE REWARDS. 3- 4. ft per 100 Taels weight **Opium**, Raw, Foreign.... or Native **Opium**, Prepared, Foreign or Native **Opium**, Dross, Foreign or Native Morphia.... of Prepared **Opium** is found the Informer will receive \$g6o-and Crew will have \$96o-divided

p.130  

, Seizure of **Opium** on board the .a.- -Smyang" at Snan'hai. The sample of prepared **opium**.... shewn to me was "French **Opium**" i.e. the output of the Monopoly of Indo-China, **Opium**...., not as a general .rule to compete in Hongkong with the ' **opium** of the Government Monopoly.... seizures of "French **Opium**" for 191? were 25 ,000 taels out of a total of 27,000 taels.... prepared **opium** and for the first half of the current year 18,677 taels out of a total

p.183  

any new Rve.L aw as to **Opium** by embodying it in an International Regulation in **Opium** and as to the

Subject

Find filter item

- Essay (2619)
- Ostasiatica (2229)
- China (1585)

Date

Find filter item

- 1924 (440)
- 1925 (402)
- 1930 (379)

Language



- English (6276)
- Chinese (808)
- Dutch (518)

Author

Find filter item

- 善者不洋 (55)
- Rewi Alley (43)
- Institute Of Southeast

Collection

- North China Daily News (7566) 
- The North China Herald Online (4351) 
- Missionary, Sinology, And Literarv Periodicals (1817-

Integrated Text-Repository – maintain & store licensed content in Berlin

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.

CrossAsia ITR Explorer (Beta version)

1. Select source(s) ⓘ
Chinese Students Monthly (1906-1991) (40 issues) X

2. Enter search term or phrase ⓘ
Search term or phrase X
CJK Mapping Phrase search

3. Show and combine result sets ⓘ
Result sets: X
opium X 鴉片 X
Combine with ... Combine with ...

4. Visualise your result sets ⓘ
VIEW (OVERLAP) CIRCULAR LINE CHART (TIME RANGE)
Show sets over time, define time range. Please select sets for visualization. X

From 1700-JAN to 1979-JAN ⚙

opium [Set 1] sources: Foreign Office File China (1919-1932), A3(1)8 (1940-2012), Foreign Office File Japan (1921-1932), ESE digital collection (Western language texts), Chinese Students Monthly (1906-1991)

5. List of matching titles ⓘ
Titles in selection (From 1700-JAN to 1979-JAN) X

| Title | Bits | Date | Creator | Subject | Crossfile | Thumbnail |
|---|------|------|---------|------------------|-----------|-----------|
| Relations between China and the Soviet Union June - August 1979 Folder 2 | ● | 1979 | | China; Soviet... | GO | GO |
| Relations between China and Japan countries | ● | 1979 | | China; Japan... | GO | GO |
| Relations between China and the Soviet Union January - June 1979 Folder 1 | ● | 1979 | | China; Soviet... | GO | GO |

How to support DH:

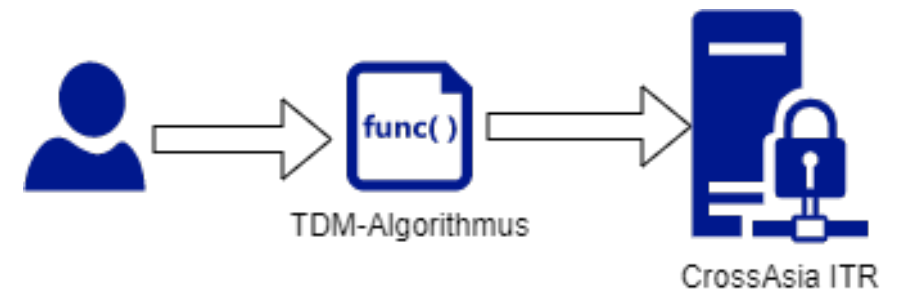
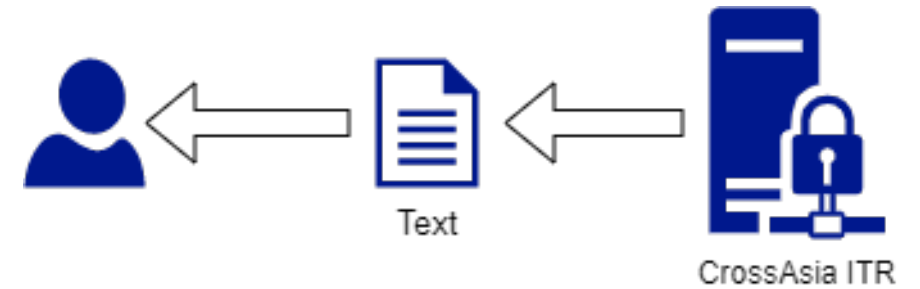
exploratory and probably also innovative Proof-of-Concept by using Compute-to-data within OCEAN Ledger

In the context of FID Asia and CrossAsia, resources managed as data objects in the ITR (currently approx. +50 million mostly in English and Chinese)

Current services: ITR full-text search, ITR Explorer

TDM rights exist

- Compute-To-Data: Algorithm goes to the data (not the other way around!)
- Execution of the algorithm in a protected environment
- Data itself will not be shared
- Only results of the algorithm are available for download
- Merging via data hub



Today & Tomorrow?

In all projects, relevance to our collections and services is important or crucial

Our experience:

In addition to the classic and previous tasks such as book selection and processing, new tasks are being added in the area of data and content management

Scaling challenge

We have the vision:

- in a trustworthy international network, partners agree on common data standards and on controlled access modes
- partners share tasks,
- share results with each other to avoid duplication of work, and
- support digital scholarship beyond their own institution through new operating models

Thank you very much for your interest and kind attention!

Matthias Kaun

Staatsbibliothek zu Berlin | Berlin State Library
East Asia Department

matthias.kaun@sbb.spk-berlin.de

<http://staatsbibliothek-berlin.de/>

<https://crossasia.org/>