FULL-TEXT SEARCH AND N-GRAM SERVICE 

based on the CrossAsia ITR (Integrated Text Repository)

CrossAsia ITR
(Integrated Text Repository)

CrossAsia Full-text Search
(Beta version)

13.5 million pages

123,112 titles

15 different sources

Snippet-view of full-text

Full-text Search

“Guided” search

Hits are grouped and displayed by book

Books are ranked

by their number of pages with hits or

by ratio, i.e. ranked on top are titles with a

higher percentage of pages that contain the

search term

Only the full-text is searched

Mapping of Chinese characters: traditional

simplified

Functionality and ranking:

Step 1: Search a term

Step 2: Select a title with the search term

Step 3: Select a page with snippet-preview and open full-text in database (log-in required)

Visualization of N-gram overlapping in a set of books:

Set of books:

18 books

16 % 地名

7 % 山名

CrossAsia N-gram Service

Downloadable Dataset

Xuxiu Siku Quanshu

Local Gazetteers

Daozang

https://crossasia.org/en/service/crossasia-lab/crossasia-n-gram-service

Online N-gram analysis (prototype)

Search N-grams in pages of a collection

Fetch all bigrams in the set of pages with your N-gram

Define criteria to show bigrams in the set of pages with your N-gram (e.g. with regular expressions)

Combine two or more sets of pages with an N-gram (A and B, A and not B)

Show distribution of N-grams in pages of a book

CrossAsia

offers access to specialized information from the entire spectrum of humanities and social sciences from and about East, Southeast and Central Asia. CrossAsia supports the Asia-related studies in research and teaching in Germany and beyond, with a special focus on DH services.

Framework for DH activities:

- Serving the primary user group - researchers located at German research institutions who work on Asia - and beyond
- Experience in licensing electronic resources for about 15 years
- Subscriptions to more than 150 electronic databases
- License agreements include:
  - Metadata and the right to use the metadata (also via APIs)
  - Data and text mining rights
  - Archival and hosting rights

Future perspectives:

- Develop a DH infrastructure that is based on stable, trustworthy and reliable financial and IT-structures for the German Asian studies research community and beyond.
- Establish an international DH network, with cooperation and collaboration among research institutions, libraries, copyright holders, and data providers, based on stable and trustworthy structures.

Services developed by:

Hou leong (Brent) Ho
hou-leong.ho@sbb.spk-berlin.de

Contact us at:

x-asia@sbb.spk-berlin.de


Objectives:

- Organize and host a huge amount of data and digital objects (licensed / in the public domain)
- Offer the data for research and education (copyright will be protected)
- Analyze and annotate texts or text corpora
- Enhance and share texts (within the limitations of license rights)
- Integrate ontologies and all kinds of knowledge bases
- Provide the results to the research community
- Develop new DH services and connect these to the ITR via the CrossAsia ITR Application Programming Interfaces (API)

Data ingest:

- Development of routines for meta and object data ingestion, ETL: extract, transform, load
- Test the technology, formats and structure by using different kinds of materials:
  - Collection of classical texts
  - Newspaper articles
  - Archival materials

Serving the primary user group - researchers located at German research institutions who work on Asia - and beyond.