



復旦大學
FUDAN UNIVERSITY

Empowering Library Special Collections with AI: A Case Study of Fudan University Library

Song XUE¹, Qing MIAO¹, Yicheng XU², Le WANG¹

¹ Fudan University Library

² School of Data Science, Fudan University

Mar 26, 2026

Outline

- **Background: Opportunities and Challenges**
- **Approach: Phased and Incremental Initiatives**
- **Case Studies**
- **Summary and Outlook**

The Changing Landscape: AI and Special Collections

AI as a **national strategic priority** in China

- National policies emphasize high-quality datasets as the foundation of the “AI+” initiative
- Strong encouragement for discipline-specific datasets to support AI-driven research

New expectations for **academic libraries**

- Special collections as unique research infrastructure
- Libraries expected to provide AI-ready research data

Global developments

- Automated cataloging, metadata enrichment, and semantic linking
- AI-enabled content extraction and knowledge transformation
- “Collections as Data” initiatives
- AI-ready corpora built from special collections...

Fudan University's AI4S/AI4SSH Strategy

- **AI-Driven Research Ecosystem**
 - Promoting disciplinary integration and innovation in AI4S / AI4SS
 - Development of the intelligent computing platform: CFFF
 - Publication of the *Blue book on the Development of AI for Social Sciences and Humanities*

- **Research Programs**

- Frontier research on large-scale scientific facilities
- Interdisciplinary programs such as Early Chinese Civilization Studies
- ...



- **AI for Education**

- AI-driven reform of teaching and learning
- AI-BEST curriculum framework : AI-Basic, AI-Essential, AI-Subject, AI-Thematic Courses
- AI+ Double-Degree Programs...

Challenges in the Utilization of Special Collections/Unique Collections

Fudan University Library Special/Unique Collections

400,000 thread-bound ancient books
(including 60,000 rare editions)

- 100,000 books/journals from the Republic of China era
- Chinese Local Chronicles, Oriental Studies
- Egyptology Collections ,Indology Collections
- Personal archives, Manuscript, Letters, Old Map, Photograph
- Digitized Seal Stamps Collections
- Thesis/Dissertation, Research Data
- Donated Materials...



Challenges in the Utilization of Special Collections/Unique Collections

Fudan University Library Special/Unique

Collections

- 400,000 thread-bound ancient books (including 60,000 rare editions)
- 100,000 books/journals from the Republic of China era
- Chinese Local Chronicles, Oriental Studies
- Egyptology Collections ,Indology Collections
- Personal archives, Manuscript, Letters, Old Map, Photograph
- Digitized Seal Stamps Collections
- Thesis/Dissertation, Research Data
- Donated Materials...

Access Barriers

Physical space scarcity + limited metadata

Data Silos

Fragmented systems and services

Domain Expertise Gap

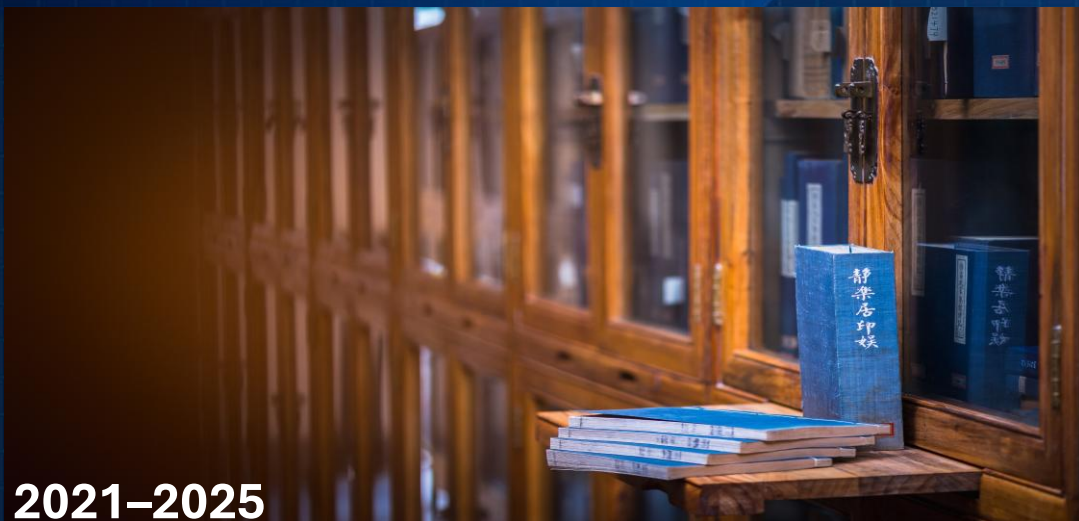
Impedes In-depth Special Collection Development

Approach: Phased and Incremental Initiatives

Strategic Roadmap for Special Collections Development

2021–2025

- Strengthening special collections
- Integrating heterogeneous metadata from multiple sources
- Enhancing scholarly information services and user experience



2026–2030

- Leveraging special collections as research data
- Developing domain-specific corpora
- Supporting AI-driven research and teaching



Approach: Phased and Incremental Initiatives

Data → Information → Knowledge → AI-enabled Services

Metadata
Integration



Metadata
Enrichment



Domain-
Specific
Corpora



AI-enabled
Knowledge
Services

These developments are supported by a systematic data governance strategy.

- Alignment with NLC Name Authorities
- IFLA Library Reference Model based unified metadata foundation

2022: Data governance for structured library data

- improving metadata quality
- establishing semantic relationships

2024–: AI-oriented data governance

- exploring AI-ready data infrastructures

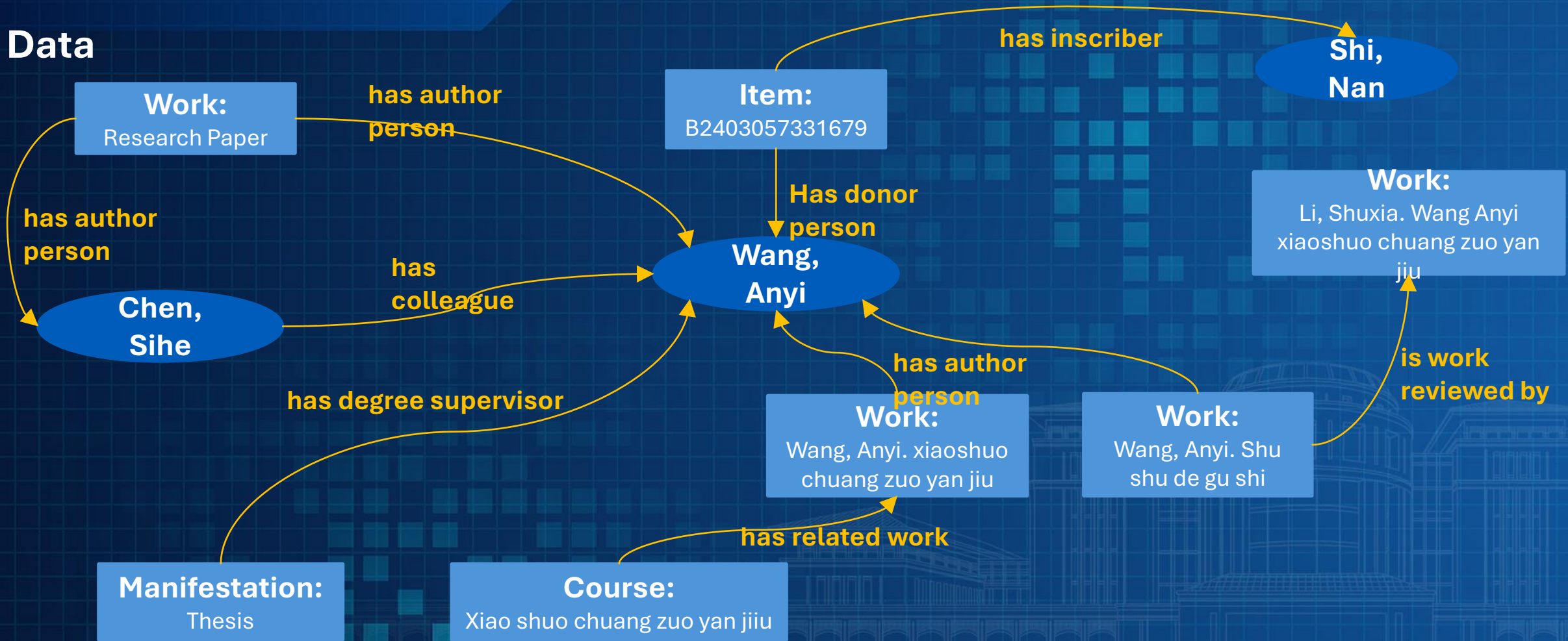
- Archaeological Texts Repository (2024-)
- Corpus of Archaeological Reports (2024-)
- Sanskrit Texts (2025-)
- ...

- Unified resource discovery
- Graph-based navigation and faceted clustering
- GenIR Agent
- ...

Case Study 1: LRM-Based Integration of Heterogeneous Metadata

Application: Building Scholarly Networks from Integrated

Data



Case Study 2: Corpus of Sanskrit Texts

Sanskrit Collections at Fudan University Library

- Total Volume: 4,700+
- Subject : Primarily Indology Related
- Key Holdings:
 - 1,284 volumes from Prof. Witzel (Harvard University)
 - 608 volumes from Prof. Mette (LMU)
- Research Value:
Many titles belong to century-old specialized series published by leading academic institutions.

Access Challenges : Collections exist but are not easily discoverable or usable.

图书详情

प्रवचन-पराग



馆藏图书

责任者 शर्मा सत्यनारायण; पोपट श्रीमती यज्ञसेनी; लोहिया श्रीमती सुषमा

出版 वाराणसी: sārvaḥauma saṃskṛta pratiṣṭhāna Haujakatora

语种 梵语

主题 Hinduism

分类 哲学、宗教

中图分类号 B982

馆藏列表 你可能感兴趣的书

条码号	索书号	馆藏地	状态	流通类型	第二索书号	操作
B2211094902233	B982 /S531 /O(AM)	理科馆 - 理科馆2号密集书库	在馆	外文出借图书	SMJL2-008-1618	预约

Case Study 2: Corpus of Sanskrit Texts

● Goal

- AI-enabled data enrichment, corpus construction and knowledge services

● Key Components

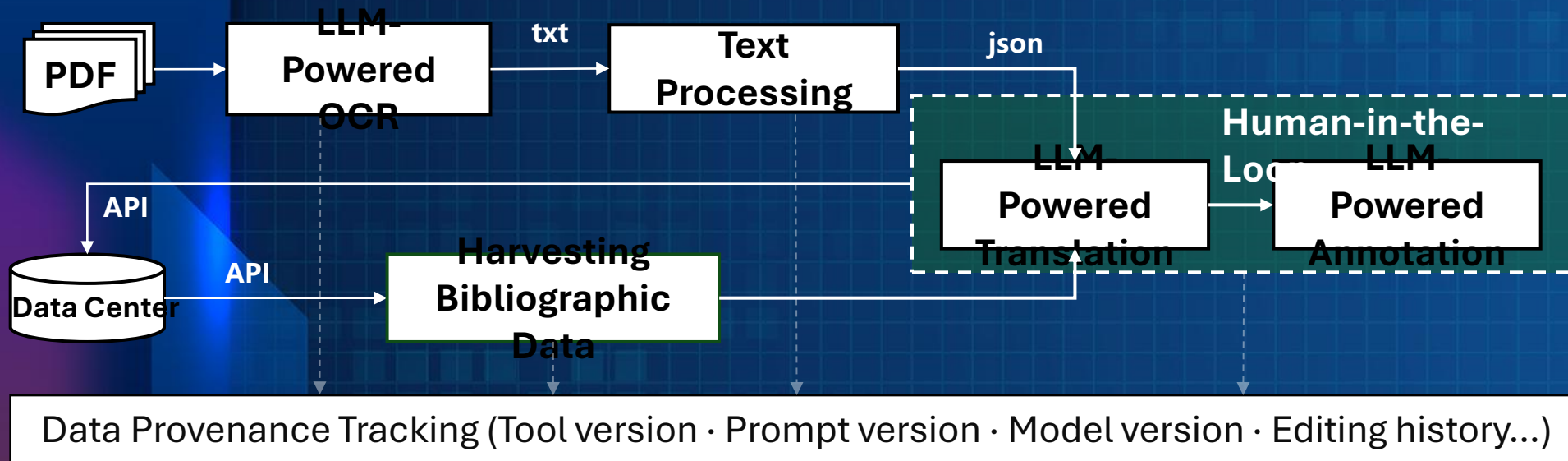
- Modular data processing pipeline
- Layered corpus architecture
- Intelligent reading assistant

● Collaboration

- Library · Faculty · Students · Tech Partner

Modular Data Processing Pipeline

Transform digitized library collections into a corpus through an AI-enabled processing pipeline.



Modular Data Processing Pipeline

Example:

AI-assisted translation of bibliographic metadata

● 题目 ○ 出版年 降序 升序 < 1 ... 233 234 235 236 237 ... 239 > 10条/页 共 2383 条

萨提亚沙德所著吠陀仪式经सत्याषाढविरचितं श्रौतसूत्रम्. 数字化申请 访问馆藏信息 收起

English Information Sys No (Book Unique Identifier) : 002054467 Index No: B982 /S254s3 /no.21-29 /O Book Main Title: सत्याषाढविरचितं श्रौतसूत्रम्. Volume No: 21-29 / Responsibility: एतस्पुस्तकं बे० शा० सं० रा० शंकरशास्त्रीमारूलकर इत्येतैः संशोधितम्. Publish Place: [पुण्याख्यपत्तने] : Publisher: [आनन्दाश्रमुद्रणालये], Publish Year: [1929-1932] Physical Desc: 1 volume (various pagings) ;25 cm. Series Title: आनन्दाश्रमसंस्कृतग्रन्थावलि: Series Vol: ग्रन्थाङ्कः ५३ Note: \"पञ्चदशसमाप्तिपर्यन्तं मातृदत्तार्थविरचितवृत्तिसमेत तदग्रे च महादेवशास्त्रिसंकलितप्रयोगचन्द्रिकाव्यार्थासमेतम्.\ Subject Topic: Brahmanism.;Buddhist philosophy.	中文信息 系统号 (图书唯一标识) : 002054467 正题名: 萨提亚沙德所著吠陀仪式经 索书号: B982 /S254s3 /no.21-29 /O 卷册编号: 21-29 / 责任者信息 (编者/译者等) : 本书由贝·沙·桑·拉·商羯罗沙斯特里·马鲁尔卡尔等人校订 出版地: [浦那市]: 出版社: [阿南达阿什拉姆出版社], 出版年份: [1929-1932年] 载体形态 (页数、尺寸等) : 1册(页数不一);25厘米 丛书名称: 阿南达阿什拉姆梵文丛书 丛书卷号: 丛书编号53 一般性注释: \"包含直至第十五章结束的母论师注释及后续大自在沙斯特里所编《仪式月华疏》释义\" 学科主题 (分类术语) : 婆罗门教;佛教哲学
---	---

Modular Data Processing Toolkit

Example:

- Page-level browsing of images , OCR text and translation
- Cross-page sentence reconstruction
- Interactive editing and validation

执行翻译 原文 AI 审阅 译文 AI 审阅

内容修改与预览 (第 11 页)

标题:

B2312095104656_mitra-ocr 第 11 页

原文 (Paragraph):

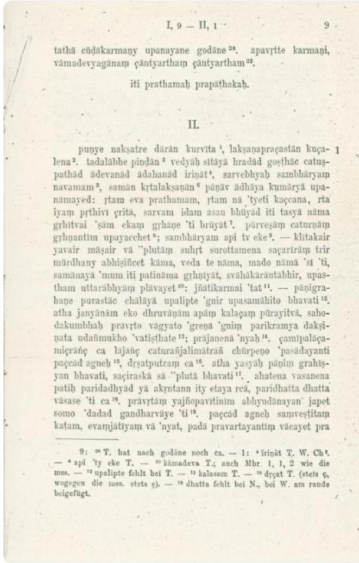
L, 9 - II, 1
9
tathā cūdākarmāny upanayane
godāne 28. apavṛtṭe karmaṇi,
vāmadevyagānam cāntyartham
cāntyartham 29.
iti prathamah prapathakah.
II.
punye nakṣatre dārān kurvīta',
lakṣaṇaprapācāstān kuṣa- 1
lena2. tadalābhe pindan vedyāḥ
sītāyā hradad goṣṭhāc catus-
pathād ādevanād ādahanād
irināt, sarvebhyah sambhāryam
navamam5. samān kītalaksanān

翻译/总结 (Summary):

L, 9 - II, 1
第九节
当剃发礼、入门礼、献牛礼等仪式
完成时，为祈求平安应吟诵《瓦玛
德维赞歌》——此为祈求平安之
仪。
以上是第一篇章。

第二篇章
第一节
宜在吉星高照时迎娶具吉祥体征的
女子。若不可得，可从祖坟、田
埂、十字路口、神坛、火葬场或荒
野获取土壤。当集齐九方之土，将
性状相近者置于掌心，引领少女完

保存修改



I, 9 - II, 1 9
tathā cūdākarmāny upanayane godāne 28. apavṛtṭe karmaṇi,
vāmadevyagānam cāntyartham cāntyartham 29.
iti prathamah prapathakah.
II.
punye nakṣatre dārān kurvīta', lakṣaṇaprapācāstān kuṣa- 1
lena2. tadalābhe pindan vedyāḥ sītāyā hradad goṣṭhāc catus-
pathād ādevanād ādahanād irināt, sarvebhyah sambhāryam
navamam5. samān kītalaksanān

切换正文/翻译 (Ctrl+.)

切换至原文 (Ctrl+;)

正文字体: A- 16px A+

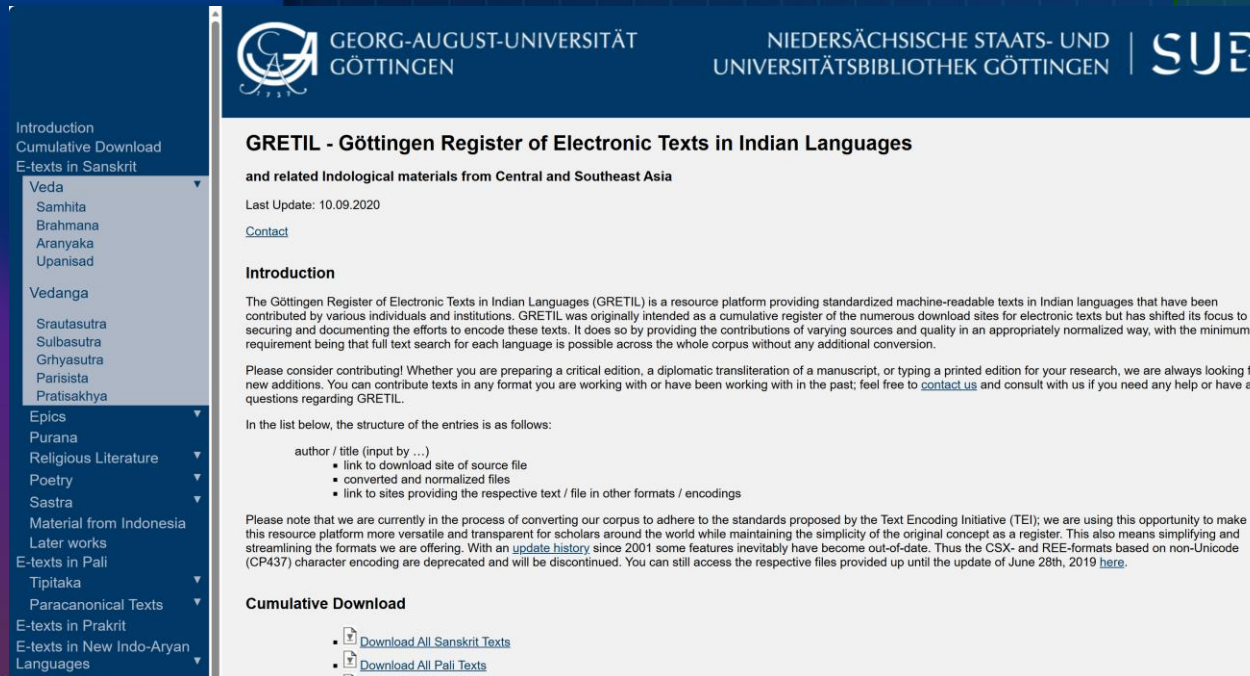
正文 (梵藏转写)

B2312095104656_mitra-ocr 第 11 页

L, 9 - II, 1 9 tathā
cūdākarmāny upanayane
godāne 28. apavṛtṭe
karmaṇi,
vāmadevyagānam
cāntyartham cāntyartham
29. iti prathamah
prapathakah. II. punye

Modular Data Processing Pipeline

AI-Assisted Subject Indexing via Domain Taxonomy (*Status: Experimental*)



The screenshot shows the GRETEL website interface. On the left is a navigation menu with categories like 'Introduction', 'Cumulative Download', 'E-texts in Sanskrit', 'Veda', 'Vedanga', 'Epics', 'Purana', 'Religious Literature', 'Poetry', 'Sastra', 'Material from Indonesia', 'Later works', 'E-texts in Pali', 'Tipitaka', 'Paracanonical Texts', 'E-texts in Prakrit', 'E-texts in New Indo-Aryan Languages', and 'Languages'. The main content area features the logos of Georg-August-Universität Göttingen and the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB). The title is 'GRETEL - Göttingen Register of Electronic Texts in Indian Languages and related Indological materials from Central and Southeast Asia'. It includes an introduction, a 'Cumulative Download' section with links for 'Download All Sanskrit Texts' and 'Download All Pali Texts', and a 'Contact' link.

Domain Taxonomy: GRETEL

AI-assisted indexing workflow :

1 Prompt engineering :

2 Batch annotation + human review :

3 Label persistence

Improve subject organization of Indological collections and enable **topic-based discovery and navigation.**

Modular Data Processing Pipeline

Example :

Title : 龙树《致友人书》 Nagarjuna's letter to a friend

Subtitle : 康楚仁波切释论 with commentary by Kangyur Rinpoche

Author : 莲花翻译小组译 translated by the Padmakara Translation Group

Note : 根本文本：《亲友书》（梵文）；释论：《圣者龙树大师亲友书注疏·教授甘露心要》（藏文）
Root text: Suḥr̥llekha (in Sanskrit); commentary: Slob dpon 'phags pa klu sgrub kyi bśes pa'i spri yig gi mchan 'grel źal rgyun bdud rts'i'i zegs ma skyabs rje bka' 'gyur rin po che'i gsu (in Tibetan)

Contents : 龙树《致友人书》--康楚仁波切《致友人书》释论--结构纲要
Nagarjuna's Letter to a friend -- Kangyur Rinpoche's commentary on the Letter to a friend -- Structural outline.

Classification given by DeepSeek : 宗教文献 (Religious Literature) >佛教 (Buddhist)

Reasons given by DeepSeek : The primary text Letter to a Friend was written by the Buddhist philosopher Nāgārjuna, and the commentary by Kangyur Rinpoche, a Tibetan Buddhist master. The work mainly interprets Buddhist teachings and provides guidance for practice, and is therefore classified under Buddhism within religious literature. Although philosophical ideas are present, the text is primarily oriented toward religious instruction and practice.

Layered Corpus Architecture

A multi-layer corpus design to support different research and service needs

Enhanced Corpus

- subject tags, Chinese metadata, full-text translation

Base Corpus

- cleaned and structured texts

Raw Corpus

- Original metadata, Digitized images, Raw OCR output

Provenance Metadata


- Processing history, Model & prompt versions, Editing records

Intelligent Reading Assistant

- Chinese–Sanskrit bilingual retrieval
- multilingual Parallel Reading
- Intelligent Transliteration
- Grammar Analysis
- Q&A

祭祀 | 检索 | 全文检索

● 题目 ○ 出版年 降序 升序 < 1 2 > 10条/页 共 20 条

 分类标签预留	火祭仪式规程 :Agnistoma paddhati :	数字化申请	访问馆藏信息	展开
	《鲍达亚那天启经》 Baudhāyana Śrautasūtra.	数字化申请	访问馆藏信息	展开
	昌多吉亚梵书Chāndogya brahmanam.	数字化申请	访问馆藏信息	展开
	古印度动物祭祀 :Das altindische Thieropfer :	数字化申请	访问馆藏信息	展开
	家庭祭祀经 /Grihya-Sūtra /	数字化申请	访问馆藏信息	展开
	贾米尼亚祭仪经注Jaiminiya-Śrauta-sūtra-vṛtti of Bhavatrāta /	数字化申请	访问馆藏信息	展开
	《阿耆尼火祭：》 L'agnihotra :	数字化申请	访问馆藏信息	展开

User Feedback and Impact

Boosts Researcher Efficiency

- “Solves 80% of Indology tasks (search/reading), freeing experts for high-end research.”

Benefits Relative Research Fields

- “Fast primary source discovery for scholars in Contemporary India, Borderlands, Comparative Linguistics, etc.”

Supports Fudan Sanskrit Course

- “Integrated with e-Learning, delivers textbooks and references, assists preview, review, and Q&A.”

...Future: Best AI Platform → Best Social Sciences & Humanities.

Summary & Outlook

A Practical Model for Empowering Special Collections with AI



Summary & Outlook

Integrated data foundations

Integrating heterogeneous metadata using models such as LRM.

1

AI-enhanced knowledge resources

Developing research corpora from special collections.

2

AI-enabled services for teaching and research

Supporting discovery, reading, and scholarly exploration.

3

Key Takeaways

Summary & Outlook



Collaboration

Multi-stakeholder co-creation around distinctive collections



Data

Prioritizing high-quality data as the foundation for effective AI applications



Standards

Corpus construction standards and data quality evaluation



Tools

Knowledge engineering pipelines and AI tools



Infrastructure

Campus-wide data and computing infrastructure

BEYOND BOUNDARIES:

New Modes of Access and Accessibility through Emerging Technologies Symposium



Thank You!

email : sxue@fudan.edu.cn

Organised by:

