CONFERENCE REPORT

2023

CADAL
China Academic Digital Associative Library

小特藏与大资源

OPEN GLAM视域下的CADAL实践

2023-09-15

金佳丽

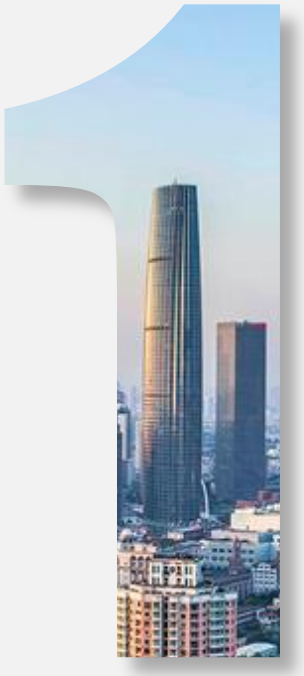
CADAL项目管理中心

CADAL

01 环境扫描：大模型背后的大数据

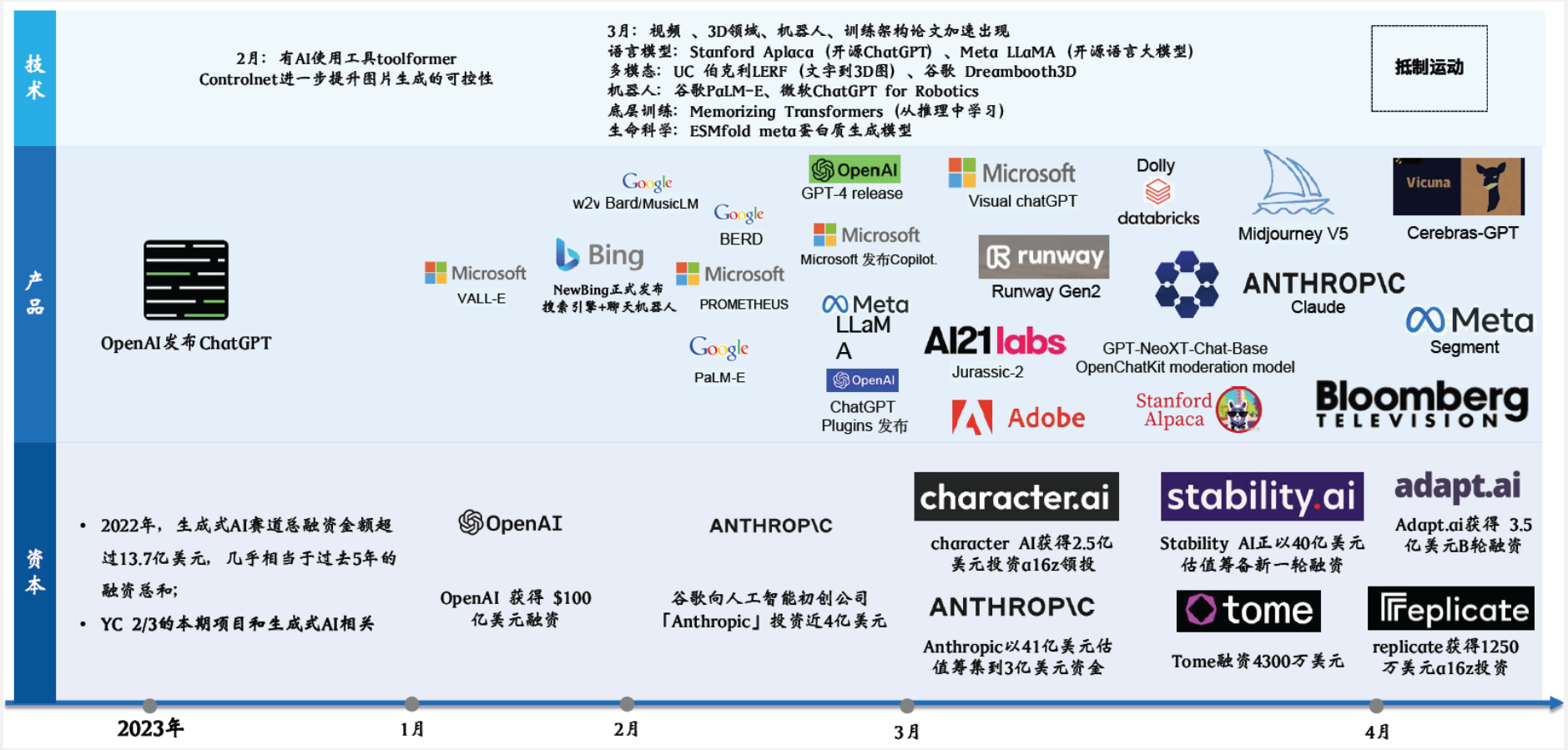
02 CADAL实践：小特藏走向大资源

03 未来计划：共建大资源



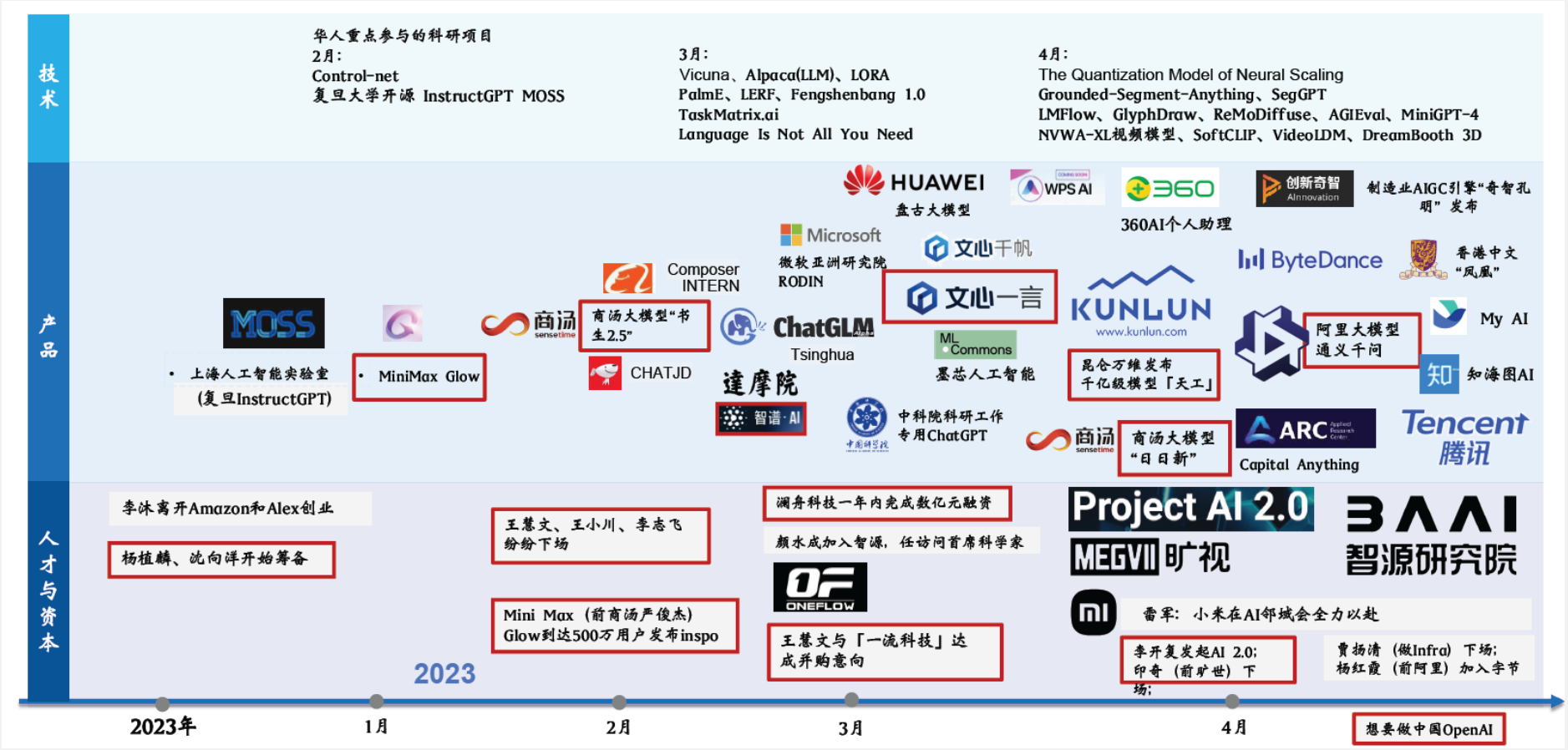
环境扫描：大模型背后的大数据

千模竞发：2023年以来的大模型浪潮（海外）



图片来源：陆奇《新范式 新时代 新机会》

千模竞发：2023年以来的大模型浪潮（国内）



图片来源：陆奇《新范式 新时代 新机会》

大模型训练数据来源统计（以GB为单位）

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GPT-1		4.6					4.6
GPT-2				40			40
GPT-3	11.4	21	101	50	570		753
The Pile v1	6	118	244	63	227	167	825
Megatron-11B	11.4	4.6		38	107		161
MT-NLG	6.4	118	77	63	983	127	1374
Gopher	12.5	2100	164.4		3450	4823	10550

*1.维基百科，是一个免费的多语言协作在线百科全书。
2. 书籍，包括Project Gutenberg和Smashwords (Toronto BookCorpus/BookCorpus)等，主要用于训练模型的故事讲述能力和反应能力。
3. 杂志期刊，包括预印本网站ArXiv和美国国家卫生研究院等已刊发的期刊论文。
4. Reddit链接，即WebText，它的数据是从社交媒体平台Reddit爬取而来，Reddit目前全球访问量排名第11，美国访问量排名第6。
5. Common Crawl，爬取了2008年以来网站信息的一个大型数据集，数据包含原始网页、元数据和文本提取，它的文本来自不同语言、不同领域。
6. 其他，主要包括开源代码社区GitHub等的代码数据集、StackExchange 等对话论坛和视频字幕数据集。

GPT-1

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB		4.6					4.6

Count	Genre	Book count	Percentage (Book count / 11038)
1	Romance	**2880**	**26.1%**
2	Fantasy	**1502**	**13.6%**
3	Science Fiction	**823**	**7.5%**
4	New Adult	**766**	**6.9%**
5	Young Adult	**748**	**6.8%**
6	Thriller	**646**	**5.9%**
7	Mystery	**621**	**5.6%**
8	Vampires	**600**	**5.4%**
9	Horror	**448**	**4.1%**
10	Teen	**430**	**3.9%**
11	Adventure	**390**	**3.5%**
12	Other	**360**	**3.3%**
13	Literature	**330**	**3.0%**
14	Humor	**265**	**2.4%**
15	Historical	**178**	**1.6%**
16	Themes	**51**	**0.5%**
	Total	11038	100%

Books数据集中的书籍类型占比

BookCorpus以作家未出版的免费书籍为基础，这些书籍来自于世界上最大的独立电子书分销商之一的 Smashwords。这个数据集也被称为Toronto BookCorpus。经过几次重构之后，BookCorpus数据集的最终大小确定为4.6GB。

GPT-3

	Wikipedia	Books	Journals	Reddit links	CC	Other	Total
GB	*11.4*	*21*	*101*	*50*	**570**		753

Dataset	Tokens (billion)	Assumptions	Tokens per byte (Tokens / bytes)	Ratio	Size (GB)
Common Crawl (filtered)	**410B**	-	*0.71*	*1:1.9*	**570**
WebText2	**19B**	*25% > WebText*	*0.38*	*1:2.6*	*50*
Books1	**12B**	*Gutenberg*	*0.57*	*1:1.75*	*21*
Books2	**55B**	*Bibliotik*	*0.54*	*1:1.84*	*101*
Wikipedia	**3B**	*See RoBERTa*	*0.26*	*1:3.8*	*11.4*
Total	**499B**				*753.4GB*

Gutenberg Book：21GB。古腾堡书籍语料库，是电子书发明人 Michael Hart 创建的项目，也是世界上第一个免费电子书网站。

Bibliotik：101GB。Bib 是互联网最大的电子书站点，通过 P2P 方式分发下载，种子数量超 50 万。 EleutherAI 实验室在 2021 年为了训练 GPT-Neo 大模型，整合精选了该电子书数据集，占 EleutherAI 实验室最后使用的 Pile 数据集中全部数据的 12.07%。

语种信息茧房在大模型时代中依然存在

2022年，OpenAI公布了GPT-3的训练数据集规模，约为2045亿词。

其中:

英语单词占比高达 92%

法语占比 1.81%

德语占比1.47%

汉语占比不到0.1%

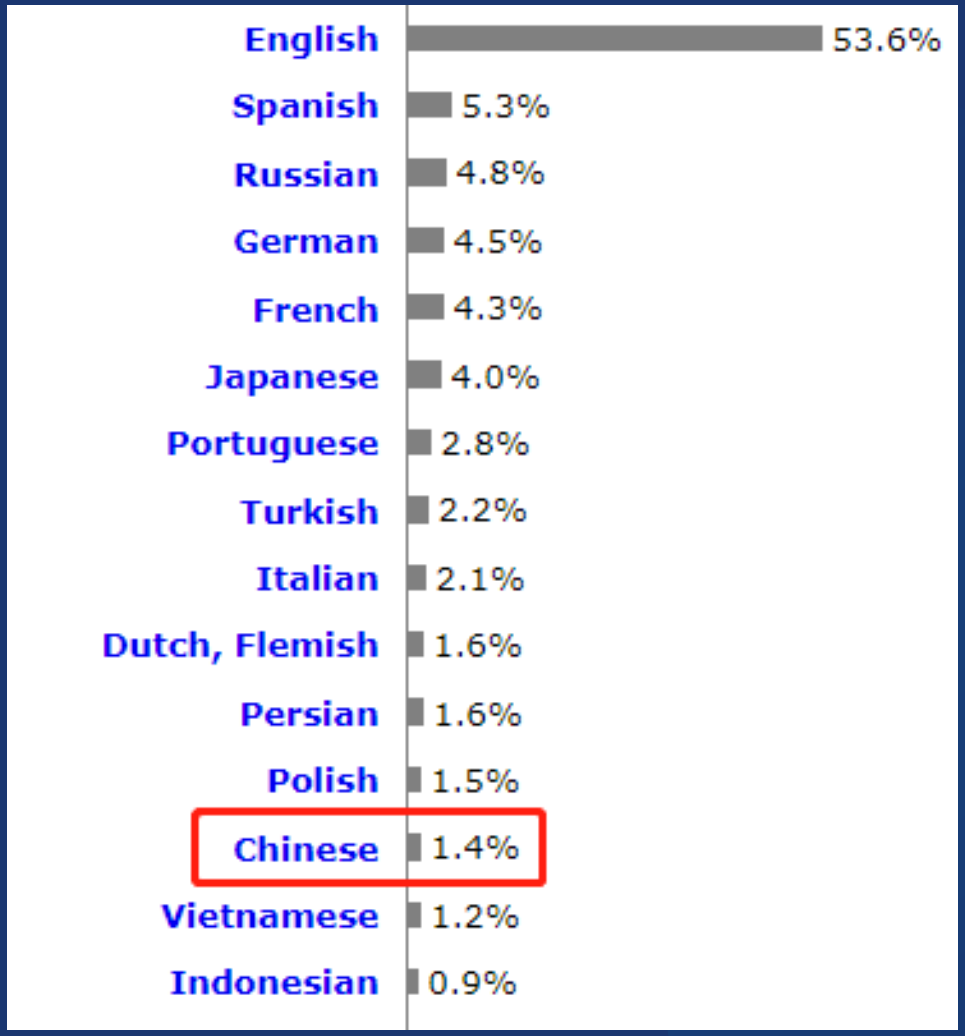
	language	number of words	percentage of total words
2	en	181014683608	92.64708%
3	fr	3553061536	1.81853%
4	de	2870869396	1.46937%
5	es	1510070974	0.77289%
6	it	1187784217	0.60793%
7	pt	1025413869	0.52483%
8	nl	669055061	0.34244%
9	ru	368157074	0.18843%
10	ro	308182352	0.15773%
11	pl	303812362	0.15550%
12	fi	221644679	0.11344%
13	da	221551540	0.11339%
14	sv	220920577	0.11307%
15	ja	217047918	0.11109%
16	no	212193299	0.10860%
17	zh	193517396	0.09905%
18	cs	139918438	0.07161%
19	hu	127224375	0.06512%
20	id	116930321	0.05985%

https://github.com/openai/GPT-3/blob/master/dataset_statistics/languages_by_word_count.csv

CONFERENCE REPORT

语种信息茧房在大模型时代中依然存在

截止到2023 年 9 月 7 日,全球浏览量前100万网页中，中文网站的占比是1.4%，排名第一的英文是53.6%。



https://w3techs.com/technologies/overview/content_language

中国大模型的数据集

- **华为“盘古”大模型：**

从5种来源的近 80TB 原始数据中清洗并构建了一个1.1TB的高质量中文语料数据集

▼表2 1.1 TB中文语料数据组成

数据来源	大小/GB	数据源	数据处理步骤
开放数据集	27.9	15个开放数据集，如DuReader、BaiDuQA、CAIL2018、Sogou-CA等	数据格式转换、文本去重
百科数据	22.0	百度百科、搜狗百科等百科类数据	文本去重
电子书籍	299.0	不同主题的电子书籍，如小说、历史、诗歌、古文等	敏感词过滤、基于模型的文本过滤
Common Crawl	714.9	2018年1月—2020年12月的Common Crawl网页数据	数据清洗、过滤、去重等所有数据处理步骤
新闻数据	35.5	1992—2011年的新闻数据	文本去重

*图片来源：曾炜,苏腾,王晖等.鹏程·盘古：大规模自回归中文预训练语言模型及应用[J].中兴通讯技术,2022,28(02):33-43.

- **百度“文心一言”大模型：**

万亿级网页数据，数十亿的搜索数据和图片数据，百亿级的语音日均调用数据以及5500亿事实的知识图谱等。

- **腾讯“混元”大模型：**特有的训练数据主要来自微信公众号、微信搜索等。

AI+X垂直领域大模型

8月21日，浙大推出了【智海】系列的垂直大模型：

◆ “智海-三乐”人工智能领域教育大模型

以阿里云通义千问7B通用模型为基座，基于核心教材、领域论文和学位论文等**教科书级高质量语料**和专业指令数据集继续预训练和微调。于今年9月起在全国13所高校应用，可提供智能问答、试题生成、学习导航、教学评估等能力。

◆ “智海-录问” 司法领域垂直大模型

由浙江大学联合阿里云、华院计算联合研制，已具备提供法律问答、知识检索增强问答、案情分析、意图识别、推理决策、法律文书生成等法律辅助服务功能。

◆ “智海-金磐”金融领域垂直大模型

训练用的高质量数据集涵盖了金融知识图谱、金融文本、金融对话等多种数据源，数据集为百亿级。能够实现对金融场景的精准理解和响应，为金融机构提供高效、智能、可信赖的语言服务，包括金融知识问答、金融文本生成、金融对话机器人等多种应用场景。

- **高质量语言数据集数据或将于2026年耗尽：**

Cornell University

We gratefully acknowledge supp

arXiv > cs > arXiv:2211.04325

Search...

Help | Advance

Computer Science > Machine Learning

[Submitted on 26 Oct 2022]

Will we run out of data? An analysis of the limits of scaling datasets in Machine Learning

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, Anson Ho

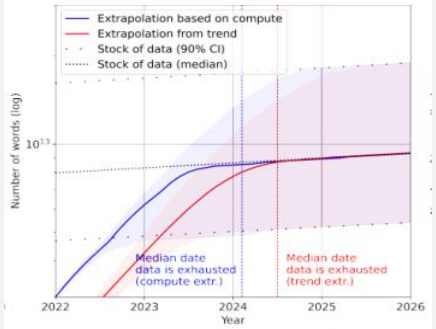
We analyze the growth of dataset sizes used in machine learning for natural language processing and computer vision, and extrapolate these using two methods; using the historical growth rate and estimating the compute-optimal dataset size for future predicted compute budgets. We investigate the growth in data usage by estimating the total stock of unlabeled data available on the internet over the coming decades. Our analysis indicates that the stock of high-quality language data will be exhausted soon; likely before 2026. By contrast, the stock of low-quality language data and image data will be exhausted only much later; between 2030 and 2050 (for low-quality language) and between 2030 and 2060 (for images). Our work suggests that the current trend of ever-growing ML models that rely on enormous datasets might slow down if data efficiency is not drastically improved or new sources of data become available.

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI); Computation and Language (cs.CL); Computer Vision and Pattern Recognition (cs.CV); Computers and Society (cs.CY)

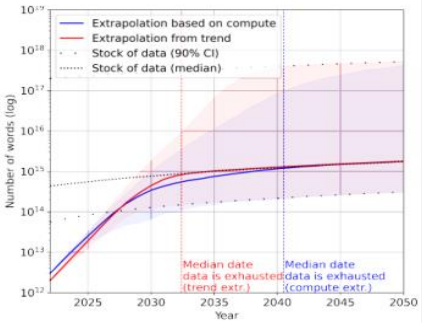
Cite as: arXiv:2211.04325 **[cs.LG]**

(or arXiv:2211.04325v1 **[cs.LG]** for this version)

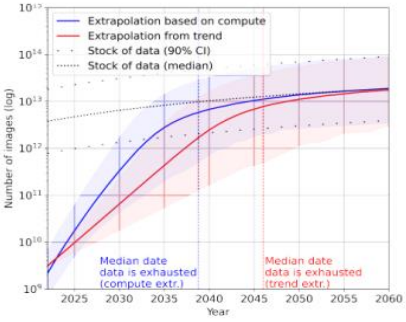
https://doi.org/10.48550/arXiv.2211.04325



(b) Projections for high-quality language data



(a) Projections for low-quality language data



(c) Projections for vision data

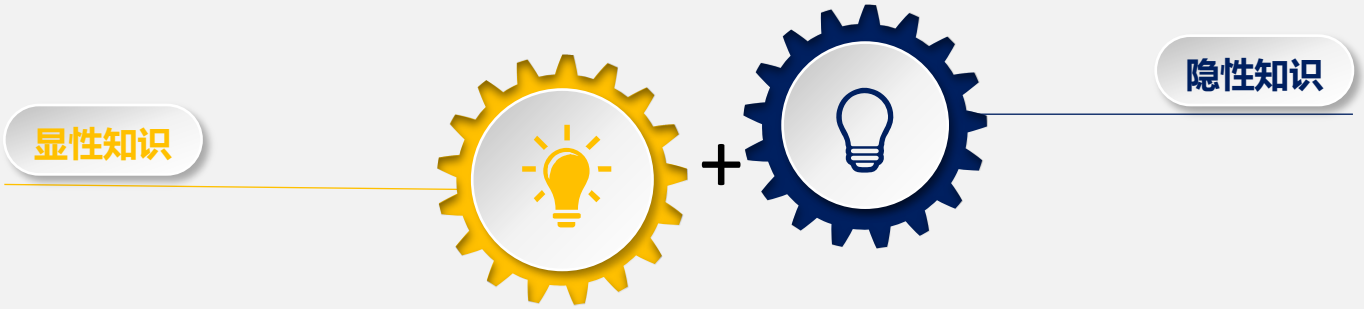
图书馆大有可为

浙江大学潘云鹤院士：

“人工智能将走向数据和知识双轮驱动”

“大数据、大模型固然重要，大知识同样重要”

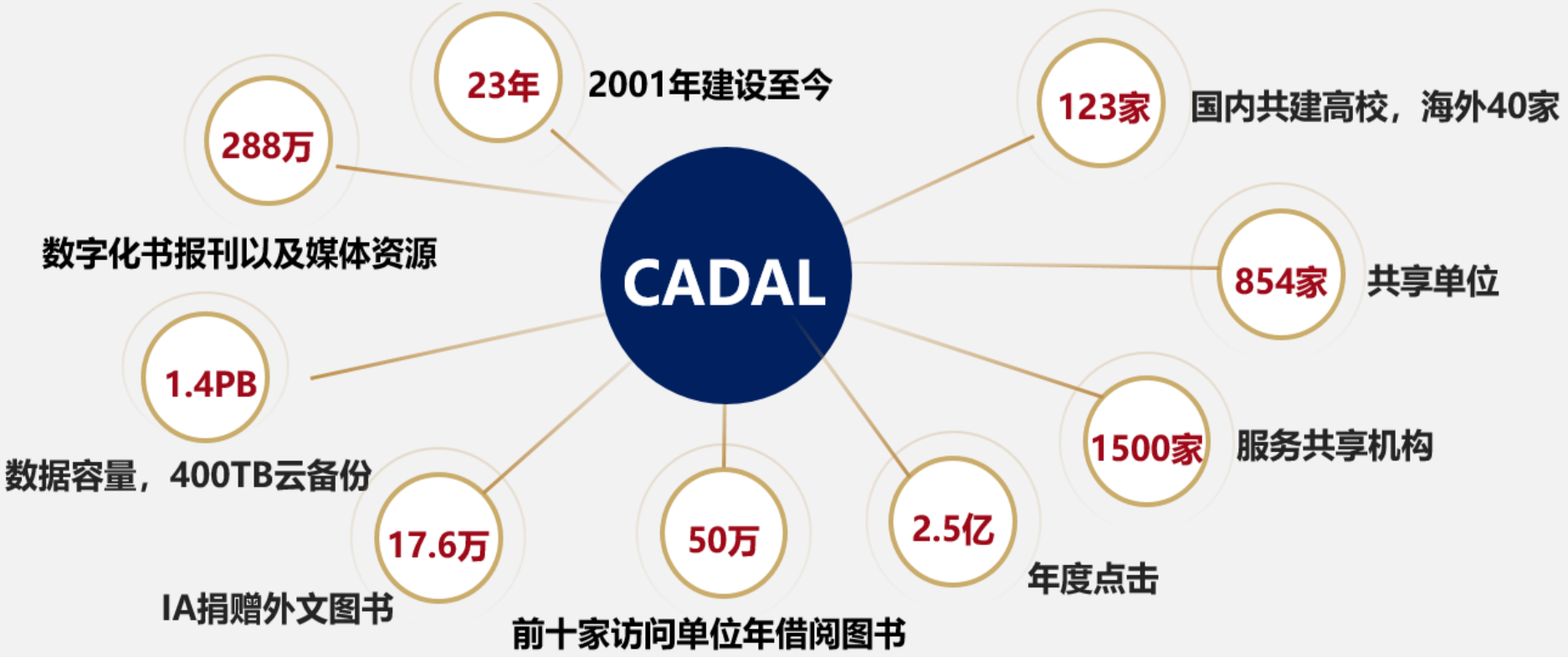
“未来的研究趋势将聚焦于知识、数据和大模型融合方式”



大模型时代，数据是新的生产资料，AI是新的生产力。

当前图书馆界讨论热烈的是 AI 在图书馆的应用，殊不知大模型赖以“涌现”的语料、知识等数据基础设施才是图书馆人大有可为的领域。

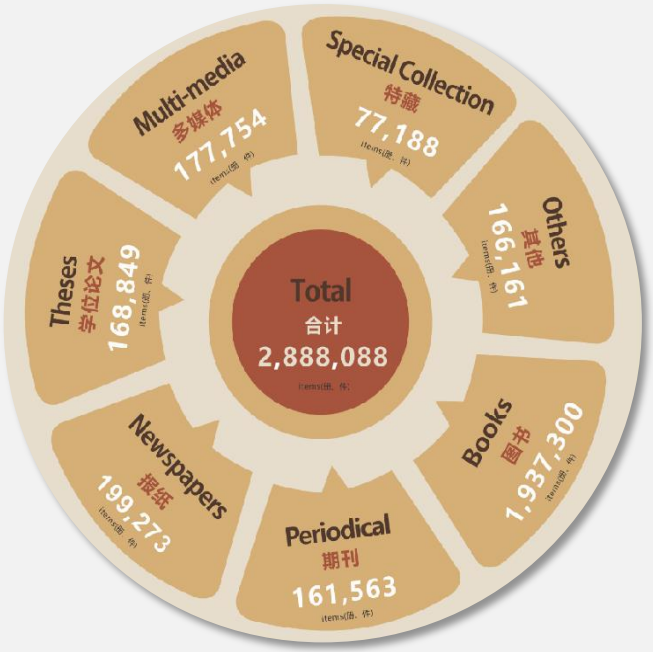
CADAL实践：小特藏走向大资源



2000年，中美计算机科学家与图书馆界联合发起“中美百万册数字图书馆国际合作计划”（China-America Digital Academic Library，简称CADAL），开始了全球最早的大规模数字化资源工程（Mass Digitization）。如今，已经建成了拥有超过288万册中英文电子图书的大型**公益性数字图书馆**。


China Academic Digital Associative Library
资源
新闻与公告
下载
关于我们
实验室
数字知识服务联盟
FAQ
中文简体
登录/注册
大学数字图书馆国际合作计划
CHINA ACADEMIC DIGITAL ASSOCIATIVE LIBRARY
全部
请输入搜索内容
CADAL
资源量 273万
共建共享单位 988家
当天检索量 9185次
当天阅读量 53650次
当天访问量 414389次
当天API访问量 304188次

资源建设呈现多类型、多模态、多语种、多学科



- 外文图书	713191
英文图书	674671
法文图书	2617
德文图书	7180
俄文图书	24879
其他外文图书	3844

首批运书的集装箱抵达

* 中美2个最大的公益性数字图书馆达成协议：Internet Archive向CADAL捐赠纸质英文图书70万册，这是建国以来数量最大的一项外文图书捐赠。

CADAL特藏类项目一览(2018—2022年）

序号	项目申请时间	申请单位名称	项目名称	建设进度
1	2018	吉林大学图书馆	**满铁**特藏资源合作共建	验收通过
2	2018	中国人民大学图书馆	馆藏**西文善本**数字化修复整理及利用推广	验收通过
3	2018	同济大学图书馆	**德文**文献资源数字化	验收通过
4	2018	东北师范大学图书馆	**俄文**图书资源数字化二期	验收通过
5	2018	中国美术学院图书馆	馆藏古籍**图像**资源库	验收通过
6	2018	清华大学图书馆	数字人文视角下的CADAL专题资料库共建共享研究	验收通过
7	2019	安徽师范大学图书馆	馆藏**徽州文书**数字化建设	验收中
8	2019	北京电影学院图书馆	**电影台本**数字化整理及利用推广	验收中
9	2019	浙江大学图书馆	**碑帖**数据库建设	验收中
10	2019	北京师范大学图书馆	高校人文社科获奖成果专题库建设	验收中
11	2019	中国人民大学图书馆	**红色文献**抢救性保护及数字化服务	验收中
12	2019	华东师范大学图书馆	馆藏碑志数字化建设	建设中
13	2020	东北师范大学图书馆	俄文特藏数字化共享项目	验收通过
14	2020	同济大学图书馆	德文文献资源数字化二期	验收通过
15	2020	清华大学图书馆	CADAL**金石拓片**资源共建共享研究	建设中
16	2020	四川大学图书馆	馆藏“张之洞捐俸置书”	验收通过
17	2020	大理大学民族文化研究院	云南**佛教碑刻拓片**专题数据库	建设中
18	2021	山西大学图书馆	馆藏石刻旧拓数字化建设项目	建设中
19	2021	浙江大学图书馆	中国**写本文献**数字资源库建设	建设中
20	2021	中山大学图书馆	馆藏碑刻拓本数据库建设	建设中
21	2021	中国美术学院图书馆	馆藏古籍图像资源建设二期	建设中
22	2022	清华大学图书馆	清华大学图书馆红色文献资源整理保护及数字化服务	建设中
23	2022	华东师范大学图书馆	红色文献特藏资源共建共享	建设中
24	2022	四川大学图书馆	四川大学特色馆藏“尊经书院学人著述”	建设中
25	2022	湖南大学图书馆	馆藏红色资源（1921-1949）数字化建设	建设中

民国文献大全

墓志拓片

中国写本文献数据库

甲骨数字化

老照片

蒋介石资料

侨批

满铁资料

地方志

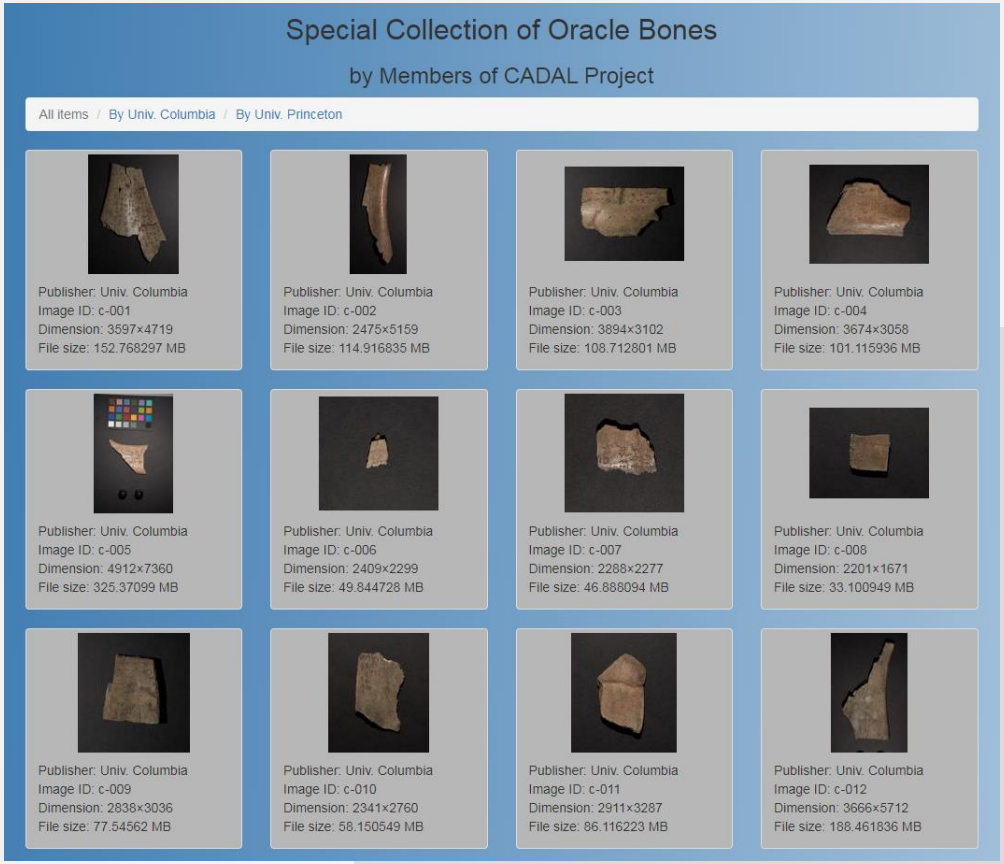
哥伦比亚大学图书馆民国文献缩微胶片

当代生活资料

民国法律法规数据库

知领特色产品

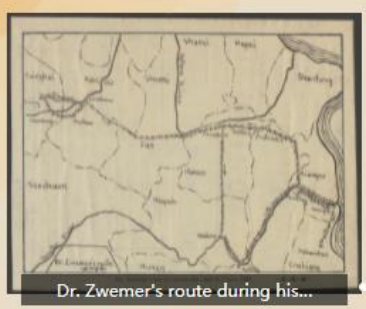
甲骨数字化


Special Collection of Oracle Bones
by Members of CADAL Project
All items / By Univ. Columbia / By Univ. Princeton
Publisher: Univ. Columbia
Image ID: c-001
Dimension: 3597×4719
File size: 152.768297 MB
Publisher: Univ. Columbia
Image ID: c-002
Dimension: 2475×5159
File size: 114.916835 MB
Publisher: Univ. Columbia
Image ID: c-003
Dimension: 3894×3102
File size: 108.712801 MB
Publisher: Univ. Columbia
Image ID: c-004
Dimension: 3674×3058
File size: 101.115936 MB
Publisher: Univ. Columbia
Image ID: c-005
Dimension: 4912×7360
File size: 325.37099 MB
Publisher: Univ. Columbia
Image ID: c-006
Dimension: 2409×2299
File size: 49.844728 MB
Publisher: Univ. Columbia
Image ID: c-007
Dimension: 2288×2277
File size: 46.888094 MB
Publisher: Univ. Columbia
Image ID: c-008
Dimension: 2201×1671
File size: 33.100949 MB
Publisher: Univ. Columbia
Image ID: c-009
Dimension: 2838×3036
File size: 77.54562 MB
Publisher: Univ. Columbia
Image ID: c-010
Dimension: 2341×2760
File size: 58.150549 MB
Publisher: Univ. Columbia
Image ID: c-011
Dimension: 2911×3287
File size: 86.116223 MB
Publisher: Univ. Columbia
Image ID: c-012
Dimension: 3666×5712
File size: 188.461836 MB

● 老照片

Pickens 老照片数据集于2016年10月从美国哈佛大学图书馆（Harvard University）引进，共计1007件。 该组老照片是20世纪初传教士拍摄中国穆斯林的纪实作品,记录了包括回族、撒拉族、东乡族等穆斯林同胞的社会百态。该组特辑保存了传教士从事中国伊斯兰文化研究的珍贵记录和历史文献资料，是有关中国穆斯林群体研究的宝库。

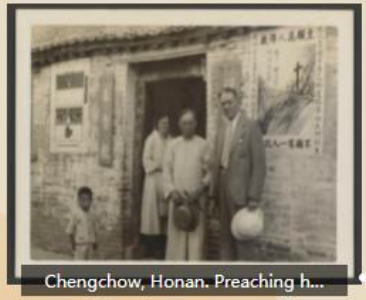

Shanghai Zhongguo Hui jiao sh...


Dr. Zwemer's route during his...


Chengchow, Honan. The Silva h...


Chengchow, Honan. Baptist Chu...


Chengchow, Honan. Baptist Chu...


Chengchow, Honan. Preaching h...


Chengchow, Honan. From one me...


Chengchow, Honan. Preparation...


Chengchow, Honan. Se Ahung; R...


Chengchow, Honan. The T'ang d...


Chengchow, Honan. The T'ang d...


Westward ho. Taxi!! Wenhsiang.

与哥伦比亚大学图书馆合作项目

共享资源：

高精度《玲珑》期刊

门神版画……

哥大藏民国文献

民国人物口述历史

海外研究中国史料

缩微胶片复制

非物质文化遗产合作

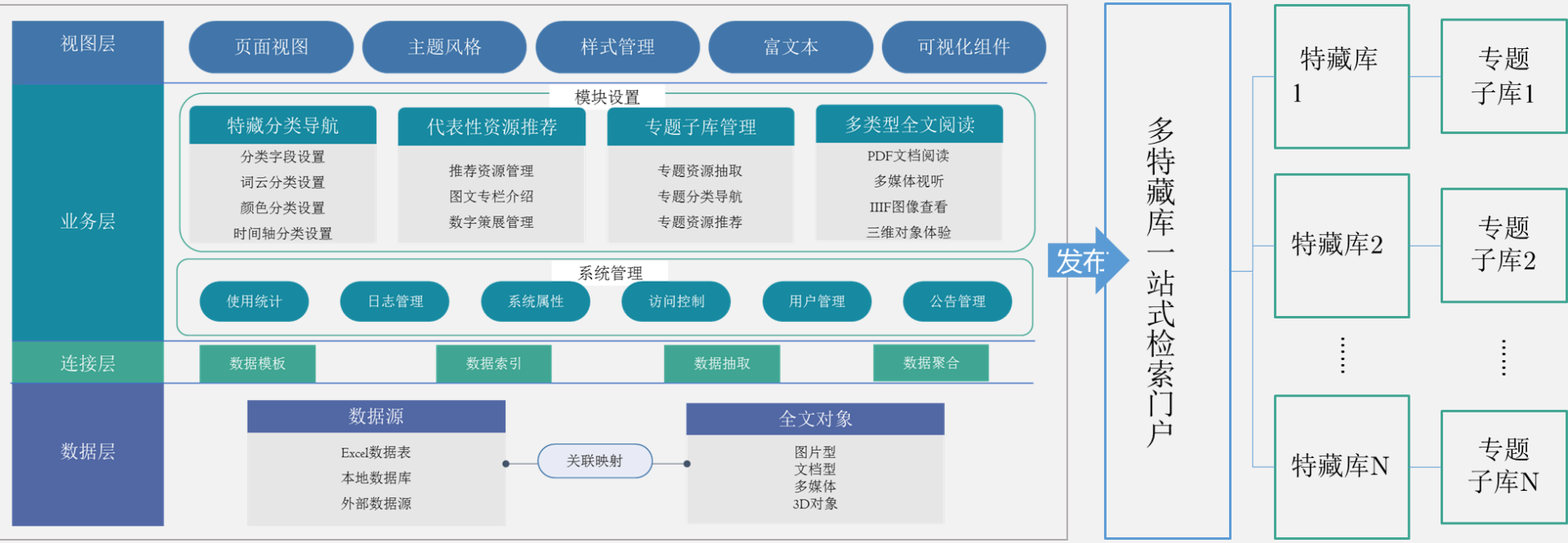
……

门神联展

2022年12月9日，《守护·传承——浙江大学图书馆、哥伦比亚大学图书馆馆藏民国门神画像巡展》先后在国家级传统村落松阳南岱村的问山美术馆、杭州市弥陀寺文化公园、杭州海塘遗址博物馆进行展览。

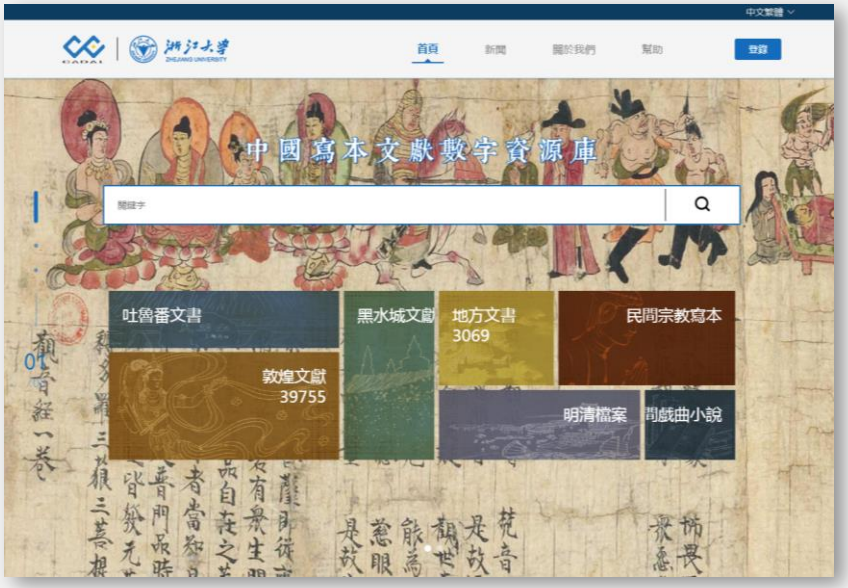
通用型特藏发布平台

2021年,CADAL联合西安交通大学图书馆完成了CADAL通用型特藏库发布平台的建设和试点应用。使成员馆可基于一套平台发布多个特藏库，满足对多种类型特藏资源存储、检索、阅读、利用、揭示的需要。



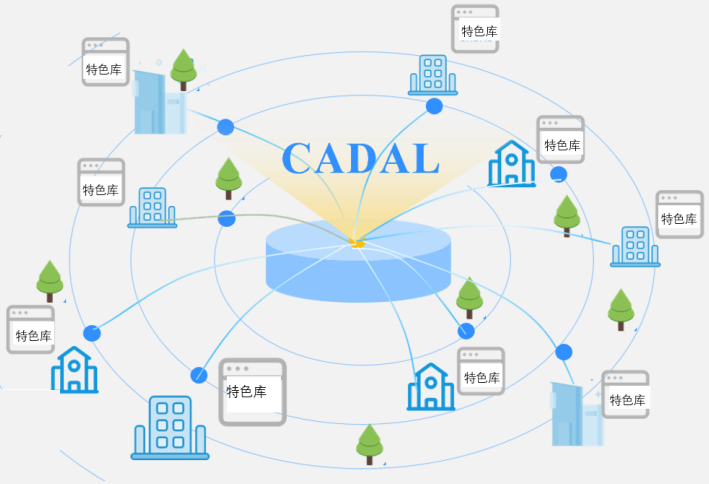
● 现有实践成果

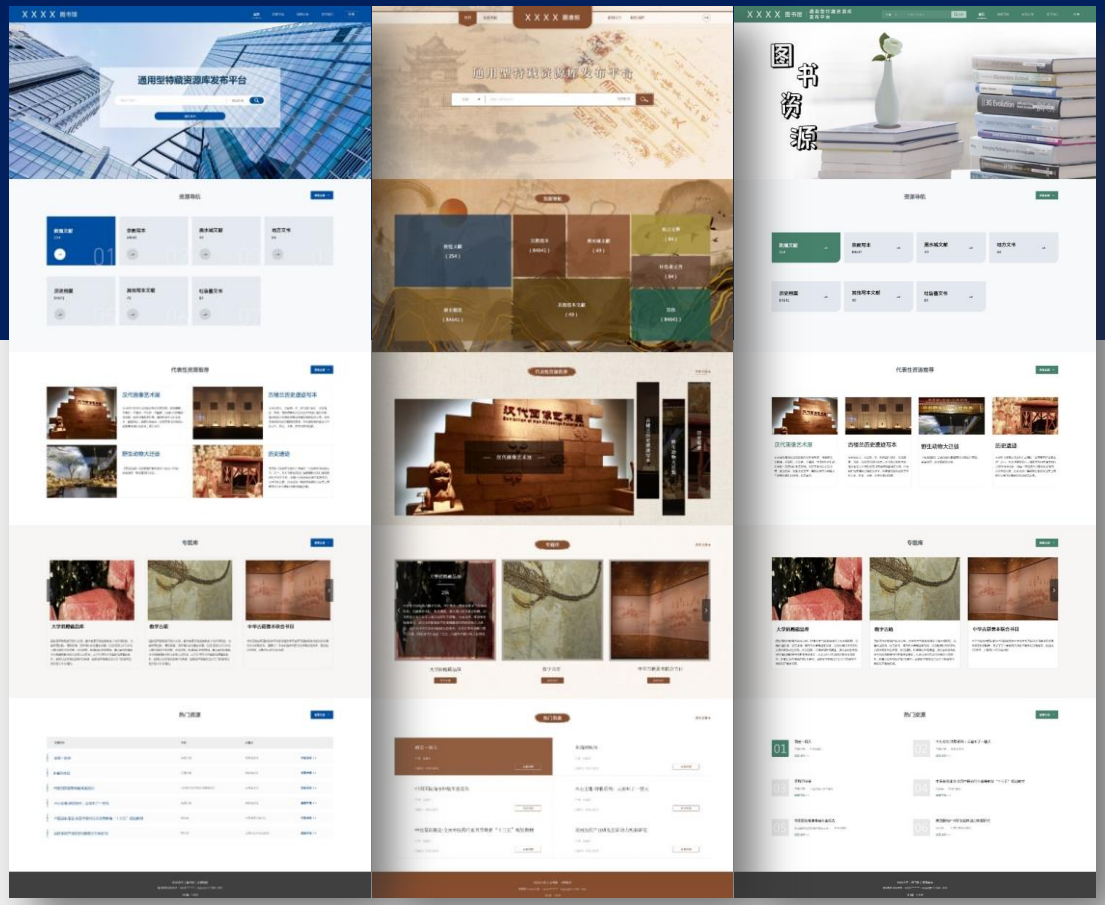
通用型特藏资源库发布平台目前已试点应用于 “中国写本文献数字资源平台”、“西安交通大学图书馆特藏资源发布平台” ，实现了古籍、民国书刊、写本文献、影像、音频等类型特藏资源的管理与揭示。



● 下一步将实施云中部署

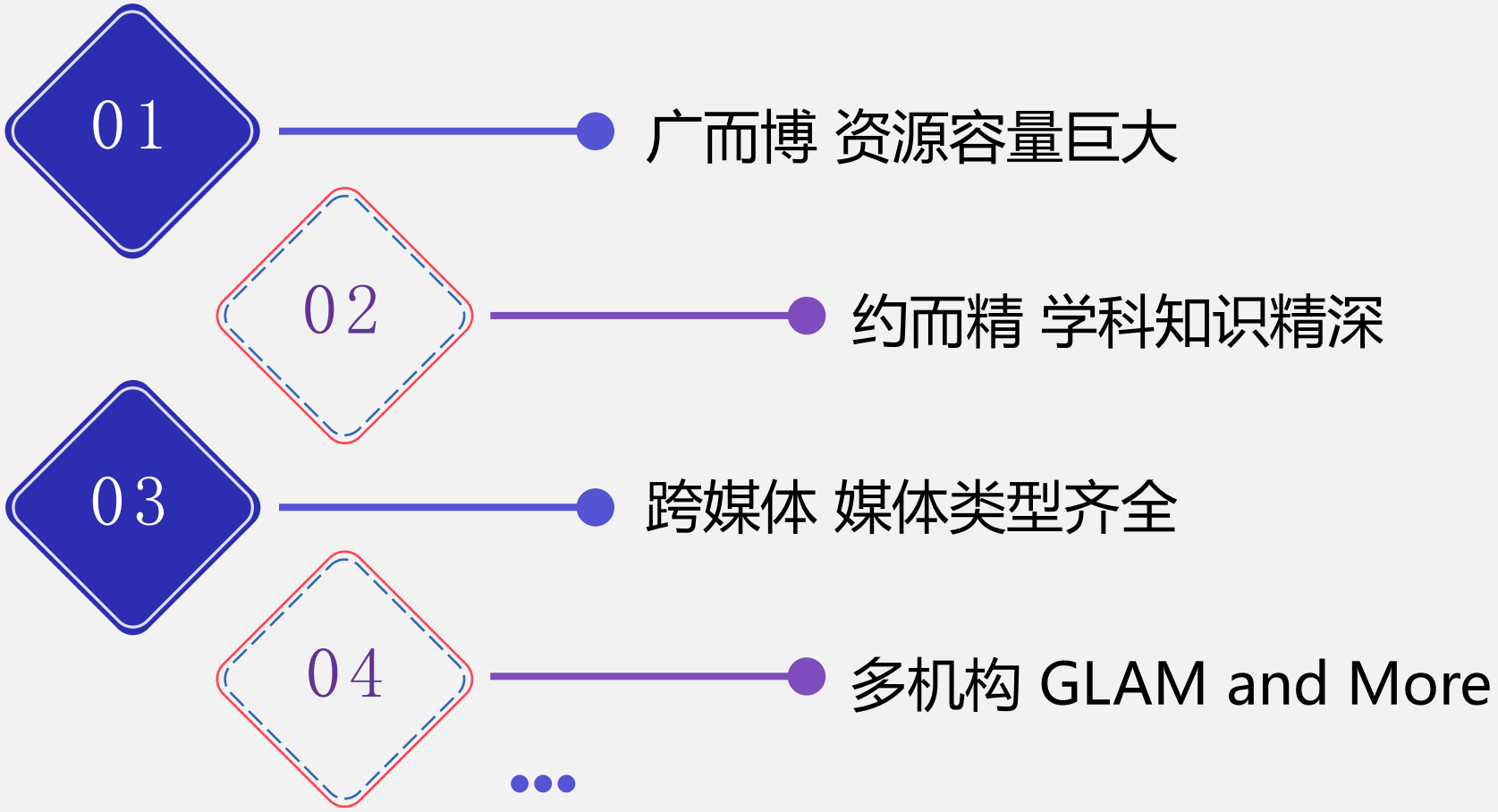
基于CADAL特藏库数据API接口可以通过数据共享，既能将成员馆特藏资源汇聚，实现向CADAL平台的深度融合，又扩大了成员馆特藏资源的揭示范围。





未来计划：共建大资源

何为大资源？



- **多措并举，盘活海量资源，为教学科研、学术/产业大模型训练提供可靠的中文知识和中华逻辑。**

扫描前的图书

修补后上架的图书

Calculus
函数(function) ↦ 切线(tangent) ↔ 微分(differential)↦ 中值定理(mean value theorem)↦ 拉格朗日中值定理(Lagrange mean value theorem) 函数(function)↦导数(derivative)↦微分方程(differential equation)↦ 一阶线性微分方程(linear first-order differential equation)
Data Structure
数据(data)↦树(tree)↦二叉树(binary tree)↦二叉搜索树(binary search tree)↦平衡二叉树(balanced binary tree) 图(graph)↦结点(node)↦最小生成树(minimal spanning tree)↦克鲁斯卡尔算法(Kruskal's algorithm)
Physics
速度(velocity)↦动能(kinetic energy)↦机械能(mechanical energy)↦机械能守恒定律(conservation law of mechanical energy) 电荷(charge)↦电流(electricity)↦电磁感应(electromagnetic induction)↦电磁感应定律(law of electromagnetic induction)

- 对海量图书按学科归类进行 OCR，形成专精的学科语料。
- 以知识图谱为抓手构建结构化显性知识，升级数字图书馆迈向知识中心。

- 共建**垂直学科**的特藏资源库，为垂直领域模型提供专业知识与数据。

垂直领域AI+X学科基座模型研究：

-敦煌（敦煌学研究）

-伏羲-万象（地理科学）

-大禹（智能会计）

- 共建**古籍资源库**，推动智慧古籍建设。

- **多力合作，共建标准、可信、开放的中文大资源，为 AI 时代的文化传承构建最坚实的中文资源基础设施。**

- 加深海内外图书馆交流，加速**海外资源的数字化回归**，深化资源共享与中华文化传承，携手共建多元化、多维度共享服务体系；
- 扩大**数字知识服务联盟**，打造“互联网+出版社+图书馆”的产业信息链融合发展生态圈，共同探索纸本数字并存时代或后纸本时代信息资源和知识资源共建共享的新模式；
- 成立**数字人文研究院**，以OPEN GLAM为理念，凝聚艺术馆、图书馆、档案馆和博物馆等文化机构，形成人文领域的大资源，进而支持大规模、跨学科、深度语义化的数字人文研究和数字文化项目打造。

CONFERENCE REPORT

2023

谢谢

OPEN GLAM视域下的CADAL实践

CADAL项目管理中心