大数据在CASHL资源发展中的应用

肖珑
CASHL管理中心
2015年6月，四川大学

CASHL

中国高校人文社会科学文献中心
China Academic Social Sciences and Humanities Library

成果来自

1. 教育部“高校人文社科外文文献资源的布局与保障研究”项目
2. 教育部“中国周边国家文献的国家保障研究”项目
3. CASHL“基于馆际互借与文献传递业务数据挖掘的读者行为模式研究”项目

合作研究馆：北京大学图书馆，复旦大学图书馆，武汉大学图书馆，厦门大学图书馆，中山大学图书馆，北京外国语大学图书馆，浙江大学图书馆，东北师范大学图书馆…

CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

主要内容

1. 背景：CASHL与大数据
2. 基于馆藏分析理论的大数据应用
3. 基于需求驱动采购的大数据分析
4. 基于科研支持的大数据挖掘
5. 结语：面对大数据

CASHL建设目标：文献渊薮，高品质服务

- 基础用户需求：文献倚赖型学科，文献需求跨度大
- 总体目标：以深入促进哲学社会科学繁荣为目标，以达到世界一流的文献保障水平为核心，努力建设国家人文社会科学信息资源平台。
  - 文献资源战略体系：覆盖全学科，覆盖全球重要学术期刊和图书，收藏率95%以上
  - 公共文献信息服务共享平台：高可用性，高保障率，移动环境下的服务，95%的国际一流文献保障率

CASHL

中国高校人文社会科学文献中心

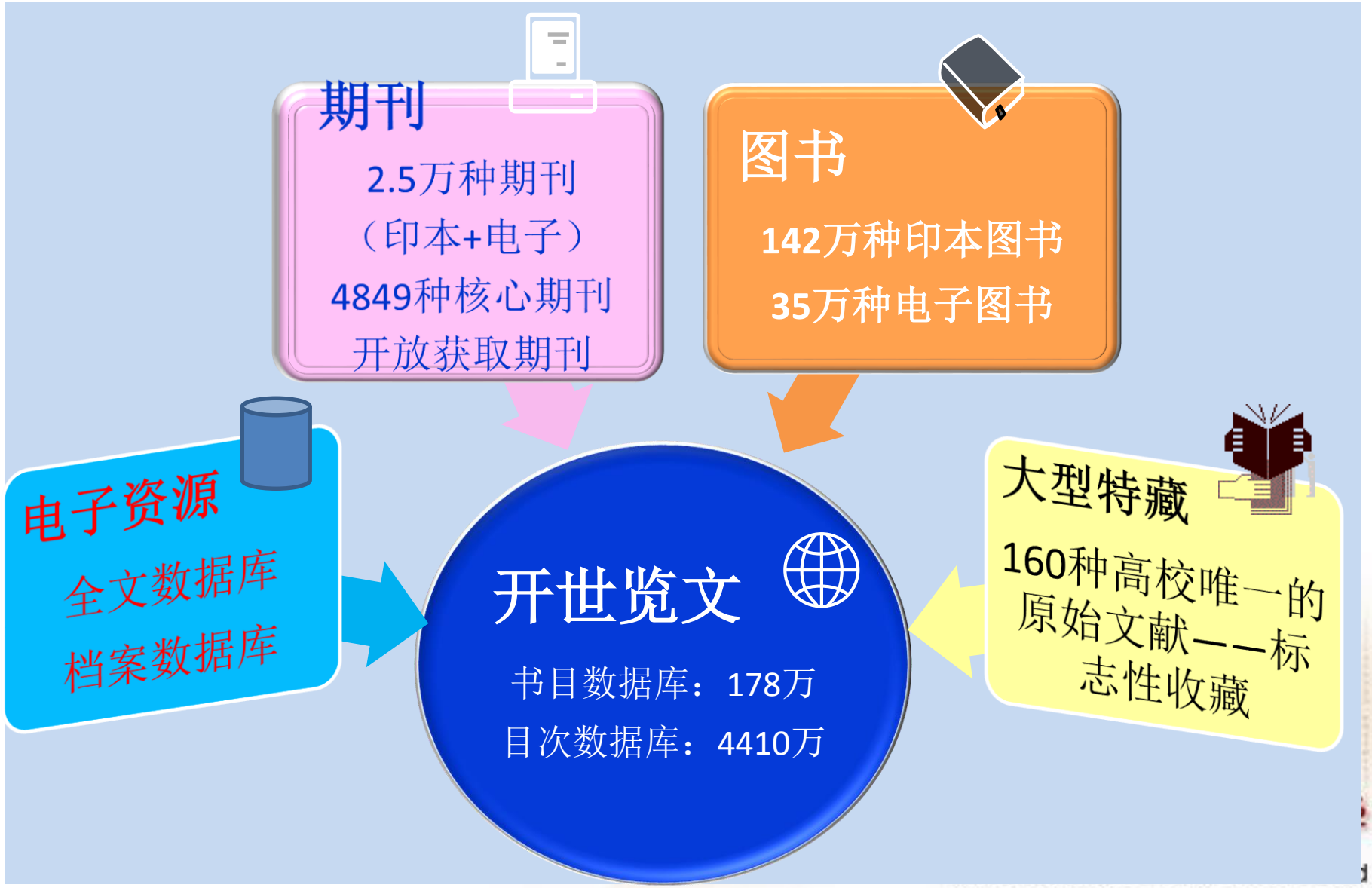
China Academic Social Sciences and Humanities Library

CASHL文科资源协调的学科格局

	文学/艺术	历史/考古	哲学/马列	政治/军事	理论经济	应用经济	法学	教育学	社会学	新闻传播	管理学	心理学	图书馆/情报/档案学	语言/文字	跨学科领域（如区域学）	大型文献
全国中心																
北京大学	√刊	√	√刊	√	√		√刊	刊	√刊			刊	刊		东方学	与图书相同
复旦大学	√	刊	√	√刊	√刊	√刊				√刊	刊			√刊	美国研究	
区域中心																
武汉大学					√		√刊		刊	√			√刊			与图书相同
吉林大学			√刊	√	√		√	刊							东北亚研究	
南京大学	√刊	√				√										
中山大学		√	√	刊							√	刊			港澳台研究	
四川大学		√刊	√	√	刊	刊				刊	刊			刊	民族学研究	
学科中心																
北京师范大学	√							√				√				与图书相同
东北师范大学		√						√								
华东师范大学								√						√	俄罗斯研究	
兰州大学		√													民族学研究	
南开大学				√		√										
山东大学	√		√													
清华大学	√										√					
厦门大学						√					√				东南亚研究	
浙江大学												√		√		
中国人民大学						√	√								欧洲研究	

70所文专院校的图书建设

CASHL资源体系不断拓展



CASHL

文献中心

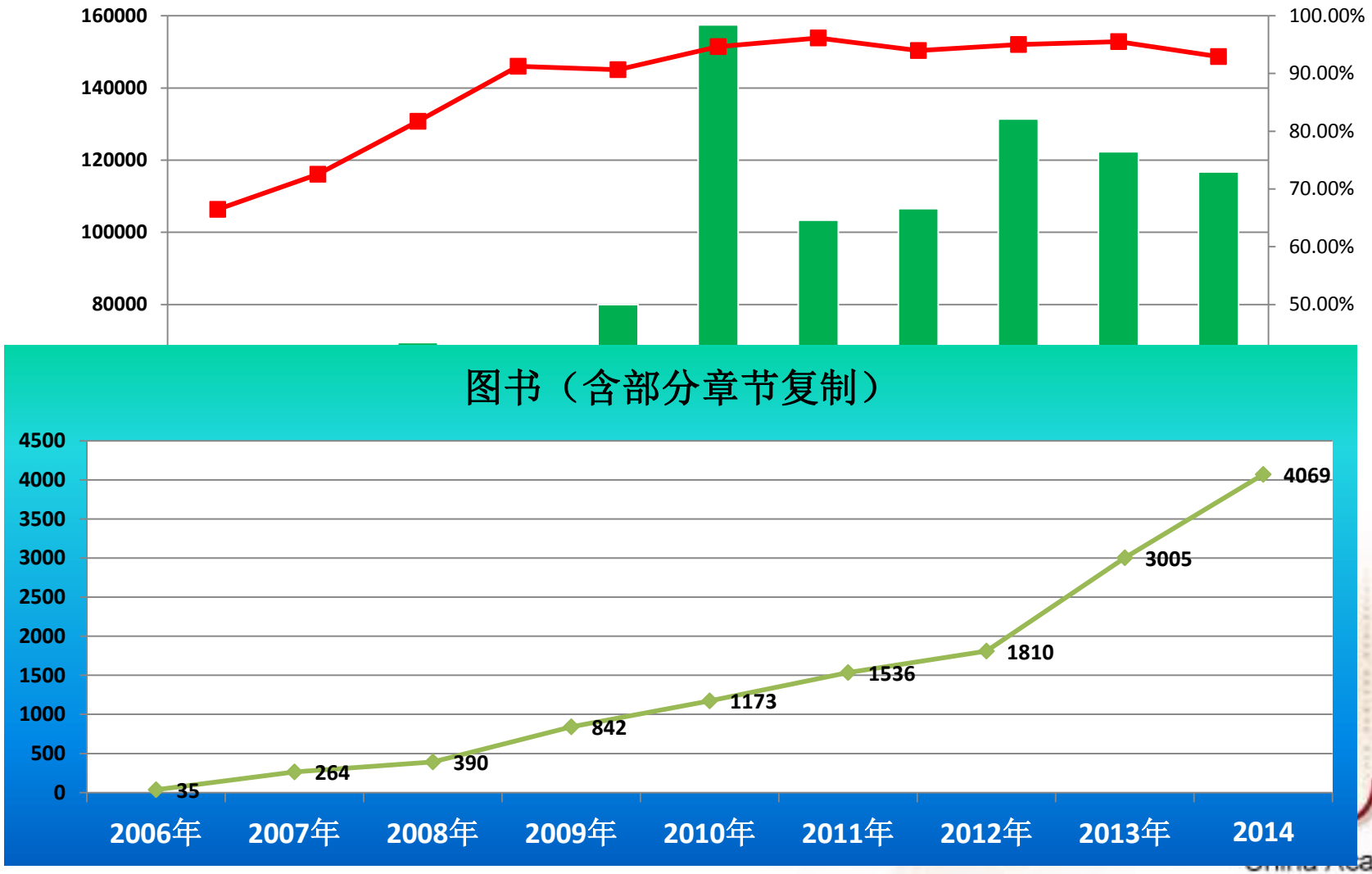
Humanities Library

有机生长的 CASHL公共服务体系

CASHL

中国高校人文社会科学文献中心
China Academic Social Sciences and Humanities Library

服务品质不断提升



为781所
高校服务

总服务量突
破100万笔

CASHL为图书馆提供了公益性云服务

资源增长：

- 2万余种外文期刊
- 178万种外文图书
- 160种大型特藏
- 联合书刊目录
- 目次数据库

服务共享：

- 相当于拥有外文资源发现系统和本馆ILL系统
- 开展服务
- 管理用户
- 自动结算

合作采购：

- 相当于拥有外文图书采购平台
- 采购书目数据查询
- 协调采购
- 自动查重
- 经费结算

编目外包：

- 相当于拥有了一个外文图书编目小组
- 文专院校仅需下载书目数据

资源长期保存：

- CASHL印本资源为高校图书馆提供资源长期保存服务
- 部分电子资源也开始发挥作用

人员培养：

- 馆员培养与交流合作项目
- 馆员资质培训
- 馆员国际出版项目

CASHL服务团队与合作伙伴

全国中心　北京大学（CASHL管理中心）　复旦大学

区域中心　武汉大学　吉林大学　中山大学　南京大学　四川大学　*代行华北中心*　*代行西北中心*　*代行华东南中心*

学科中心　北京师范大学　东北师范大学　华东师范大学　兰州大学　南开大学　山东大学　清华大学　厦门大学　浙江大学　中国人民大学

服务馆　来自于各文专院校

合作机构　CALIS管理中心 中国社科院
上海图书馆 澳门科技大学 BALIS管理中心

大数据4V

- 数量大（volume）
- 类型多（variety）
- 速度快（velocity）
- 有价值（value）

图书馆拥有哪些大数据

- 书目数据
- 馆藏数据
- 文献知识数据
- 用户数据
- 用户行为数据
- 服务数据
- 内部业务数据
- …

图书馆对大数据的应用

- 馆藏分析
- 资源整合
- 用户行为分析
- 用户需求挖掘
- 知识挖掘
- 建立新的业务模型
- …

主要内容

1. 背景：CASHL与大数据
2. 基于馆藏分析理论的大数据应用
3. 基于需求驱动采购的大数据分析
4. 基于科研支持的大数据挖掘
5. 结语：面对大数据

馆藏分析理论与模型

- 文献收藏率=一定时期内文献收藏种数/一定时期内文献出版种数×100%
- 文献缺藏率=一定时期内文献缺藏种数/一定时期内文献出版种数×100%
- 文献保障率=一定时期内可提供文献种数/一定时期内用户使用文献种数×100%
- 文献满足率=一定时期内可满足文献种数/一定时期内用户文献需求总量×100%

数据来源

- 国外各出版商出版书目
- OCLC WorldCat数据库的国外部分一流高校书目数据（哈佛大学、牛津大学、哥伦比亚大学、耶鲁大学、剑桥大学、普林斯顿大学）
- 日本高校图书馆文献收藏数据（CINII学术信息检索平台）
- 台湾地区学术研究机构订购西文纸本期刊资料库
- 中国高等教育文献保障系统（CALIS）联合目录数据库
- 中国高校人文社会科学文献中心（CASHL）联合目录数据库
- 全国高校图书馆进口报刊预订联合目录
- 各类引文数据库（如SSCI、A&HCI、CPCI-SSH、SCI、CPCI-S）
- 用户发表成果目录等。

CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

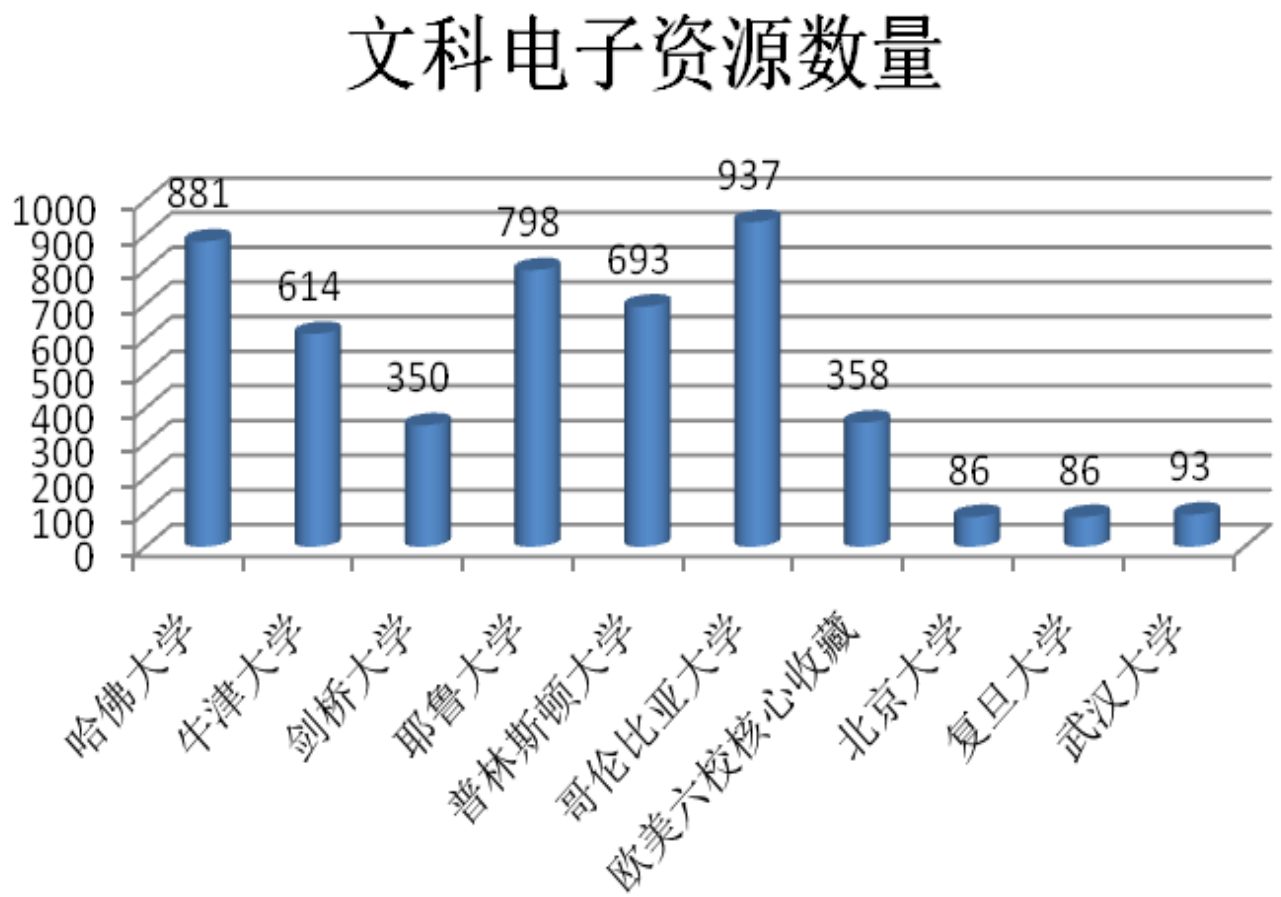
数据分析结果：普通西文图书，缺藏

	国内高校图书馆收藏量（种）	哈佛、耶鲁和牛津大学收藏量（种）	国内高校收藏率	国内高校缺藏率
1950-2000年的常用图书	728371	3811928	19.1%	80.9%
1950-2000年的核心图书	98468	609890	16.1%	83.9%
1950-2000年的英文核心图书	94727	456799	20.7%	79.3%
1950-2000年的德文核心图书	2097	87990	2.4%	97.6%
1950-2000年的法文核心图书	1644	65101	2.5%	97.5%
2004-2007年的常用英文图书	57543	191387	30.07%	69.93%
2004-2007年的学科平均收藏	2988	9850.7	30.33%	69.67%
2000-2007年的俄德法文图书	10704	174486	6.1%	93.9%

数据分析结果：普通西文期刊，核心期刊齐全

	国外出版情况（目录）		英美六校收藏情况		台湾地区收藏情况		国内高校收藏情况	
学科类别	品种数	学科比例(%)	品种数	学科比例(%)	品种数	学科比例(%)	品种数	学科比例(%)
哲学	6296	7.54	3582	6.61	686	9.43	727	8.65
社科总论	6868	8.23	4057	7.48	921	12.66	714	8.49
政治法律	15992	19.15	13032	24.04	867	11.92	1916	22.79
军事	404	0.48	0	0	0	0	0	0
经济	20449	24.49	10319	19.03	1274	17.51	1732	20.6
文教	13714	16.43	7748	14.29	1215	16.7	1265	15.05
语言	2078	2.49	1261	2.33	750	10.31	345	4.11
文学	4924	5.9	4036	7.44	0	0	430	5.11
艺术	5949	7.13	4661	8.6	822	11.3	587	6.98
历史	5591	6.7	4864	8.97	351	4.82	595	7.08
工具书	1227	1.47	654	1.21	390	5.36	96	1.14
合计	83492	100	54214	100	7276	100	8407	100

数据分析结果：电子资源，专业类资源匮乏



CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

数据分析结果：周边国家小语种图书，尤其不及日本

地区	语种	馆藏目录	图书出版年代							合计种数
			1950—1959	1960—1969	1970—1979	1980—1989	1990—1999	2000—2009	2010—2012	
合计		CALIS	39508	60728	112178	170986	120616	86255	14187	604458
		CINII	190729	307553	432814	583378	712101	810669	208282	3245526
		哈佛大学	47679	108758	164813	179702	159322	185431	49426	895131

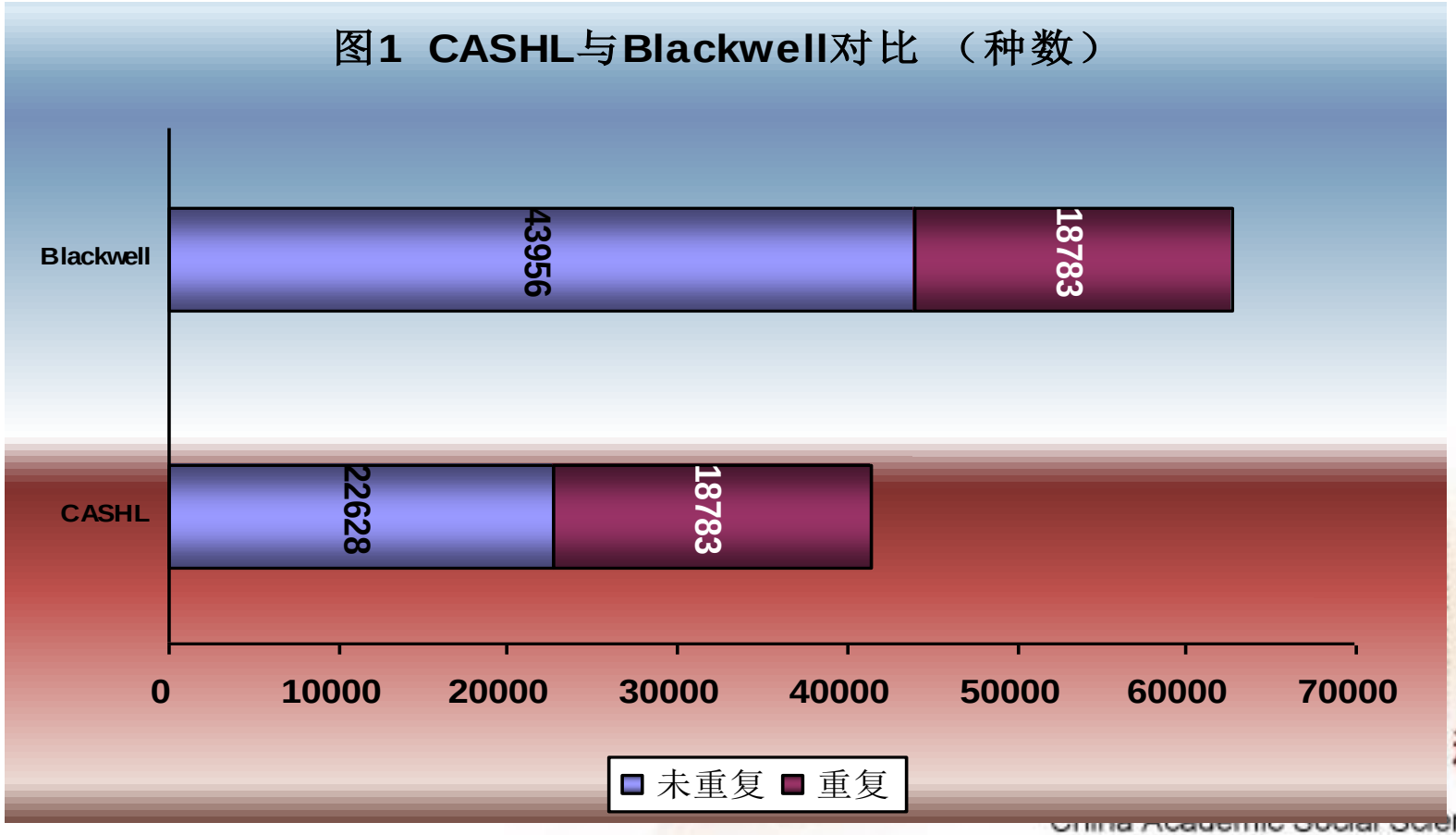
CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

数据分析结果：西文图书新书采购情况，仅收录1/3

Blackwell学术新书2004至2006年新书报导书目总量为62,739种，
CASHL未收录 的图书为43,956种，未收录率达70.06%。



数据分析结果：西文文献保障不足

	用户已发表文献的引文数量（篇）	引文涉及出版物品种数量（种）	本馆收藏量	本馆保障率	CASHL收藏量	CASHL保障率
经济学图书	2035	1525	606			
法学图书	2498	2349	391			
哲学图书	2917	2098	1089			
历史学图书	1754	1689	317			
图书平均保障率						
经济学期刊	5166	730	4614			
法学期刊	1280	663	834			
哲学期刊	786	362	616			
历史学期刊	87	73	54			
期刊平均保障率						

数量 / 年代	外文文献引用		CASHL收藏	
	数量（引文/涉及出版物）	占引用外文文献总量百分比	数量（引文/涉及出版物）	保障率（即CASHL收藏占该时期引用外文文献总量百分比）
1900以前	135	0.82%	57	42.22%
	115	1.21%	43	37.39%
1900-1949	588	3.56%	348	59.18%
	448	4.72%	238	53.13%
1950-1959	504	3.05%	310	61.51%
	348	3.67%	167	47.99%
1960-1969	892	5.40%	590	66.14%
	606	6.39%	333	54.95%
1970-1979	1597	9.67%	1068	66.88%
	1001	10.55%	556	55.54%
1980-1989	2764	16.73%	2085	75.43%
	1616	17.03%	1019	63.06%
1990-1999	5478	33.15%	3838	70.06%
	3185	33.57%	1733	54.41%
2000-2010	4090	24.75%	2634	64.40%
	2518	26.54%	1230	48.85%

CASHL

CASHL的实际措施

- 总体：年投入增加
- 年代：部分经费用于回溯补藏
  - 零散与集中采购结合
  - ALIBRIS二手书
  - HUP电子书与POD方式结合
- 语种：设立专项资金购买周边国家小语种文献
- 专业化：设立专项资金购买大型特藏

主要内容

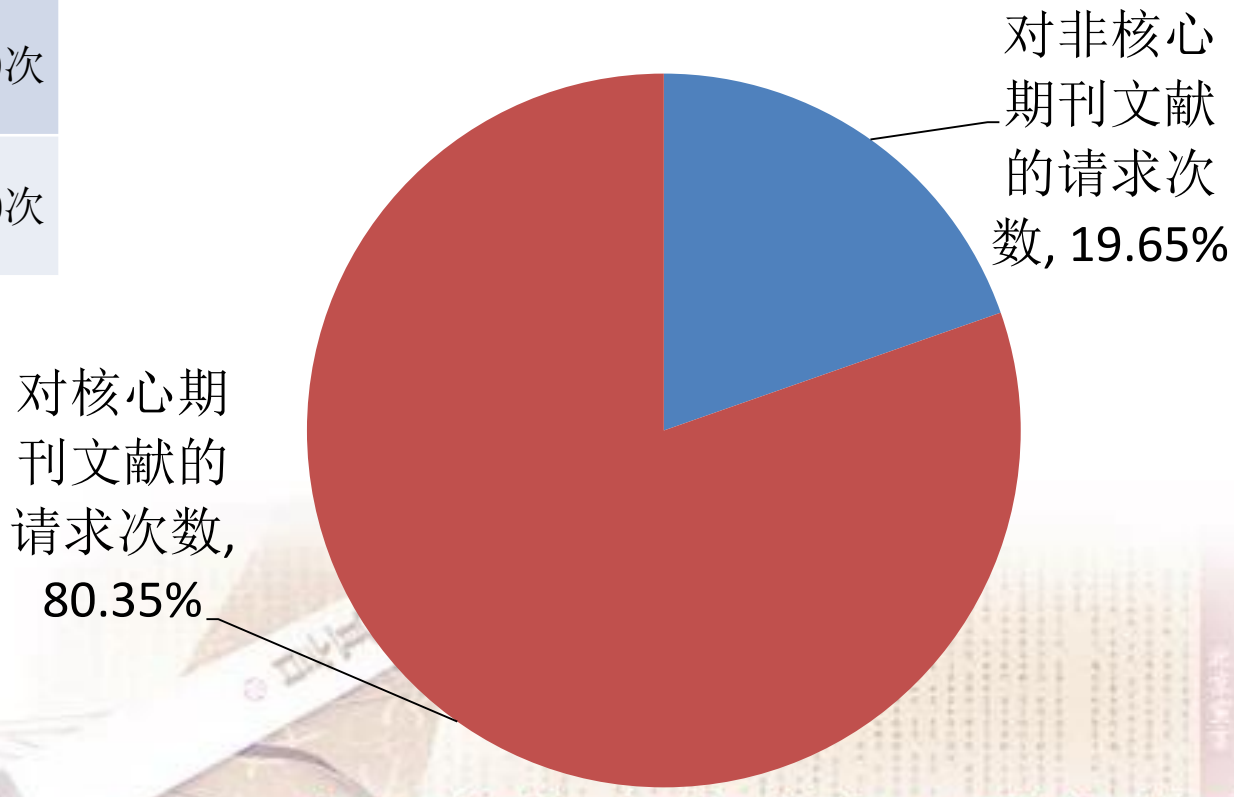
1. 背景：CASHL与大数据
2. 基于馆藏分析理论的大数据应用
3. 基于需求驱动采购的大数据分析
4. 基于科研支持的大数据挖掘
5. 结语：面对大数据

DDA理论

- Demand Driven Acquisition，需求驱动采购，基于用户需求和用户使用资源的图书馆资源采购，也可以更泛义地指基于用户直接或间接反馈的图书馆采访，包括基于教师请求和对图书馆资源使用的数据分析。
  - 向用户提供对大范围图书的立即访问，一旦有需求即可采购；
  - 以比传统采购模式更切实的方式呈现更多用户可能使用和购买的图书书目；
  - 如果实施恰当，DDA可使只按需要购买成为可能，图书馆得以节省经费，或以目前同样的图书经费总额，取得更高的图书使用率。
- 已成为国外高校图书馆馆藏发展战略中的常见组成部分
- 与传统的图书馆主导模式相结合，目前主要用于电子资源采购
- 例如：美国国家标准《图书的需求驱动采购》（**Demand Driven Acquisition of Monographs**，NISO RP-20-2014）

核心期刊利用率

对非核心期刊文献的请求次数	167831次
对核心期刊文献的请求次数	686479次
总计请求次数	854310次



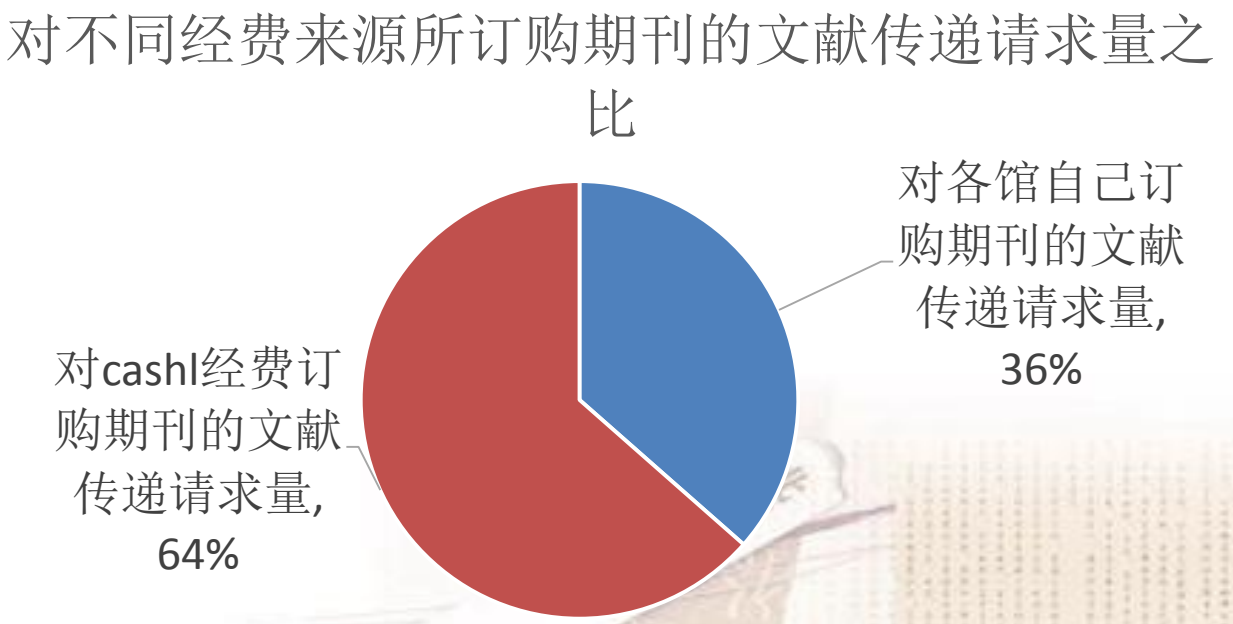
CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

CASHL经费购买期刊的利用率

CASHL经费购买的期刊，占全部服务期刊的20%
CASHL期刊的文献传递请求，占全部服务总量的64%



CASHL

中国高校人文社会科学文献中心
China Academic Social Sciences and Humanities Library

结论与讨论中的措施

- 是保障重点还是保障全面？
- 是查漏补缺还是优先核心资源？
- 是新刊还是旧刊？
- 是印本期刊还是电子期刊？
- 如何分学科开展期刊建设？
- 长期保存问题如何解决？

复旦大学图书馆正在协调组织CASHL期刊建设新方案

主要内容

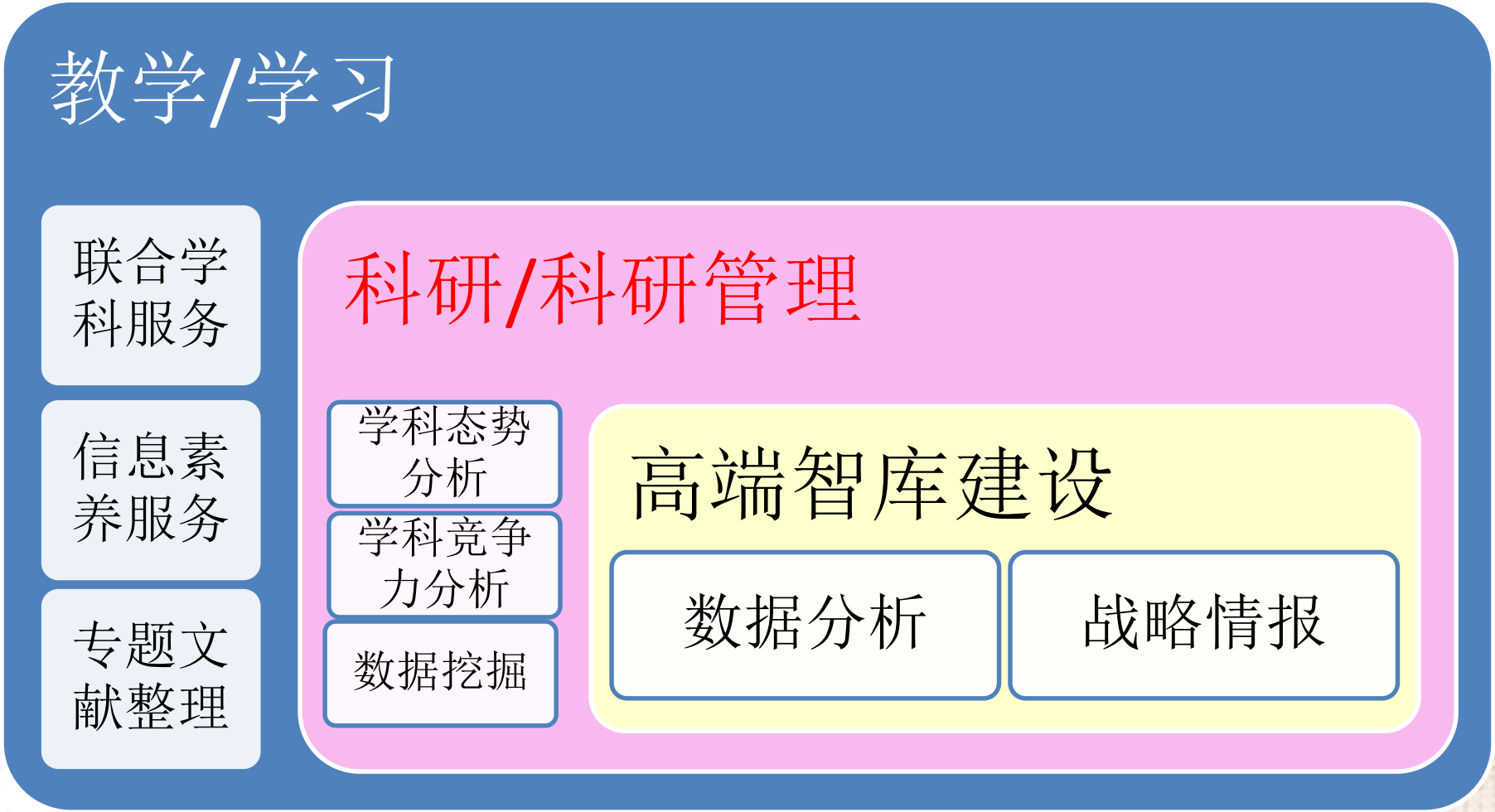
1. 背景：CASHL与大数据
2. 基于馆藏分析理论的大数据应用
3. 基于需求驱动采购的大数据分析
4. 基于科研支持的大数据挖掘
5. 结语：面对大数据

CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

人文社科知识发现服务架构



CASHL大型特藏文献

具备以下特点的大型文献：

- 学科集中，有相对完整的专题；
- 在国内（至少是高校范围内）具备相对唯一性；即，也是没有必要在国内买多个复份的；
- 系统性和完整性，需要在一个地方收藏的，无法拆分的；
- 平时经费很难采购的文献；
- 能够成为文专图书建设的标志性收藏的；
- 能够揭示、报道并为全国服务；
- 学者荐购的。

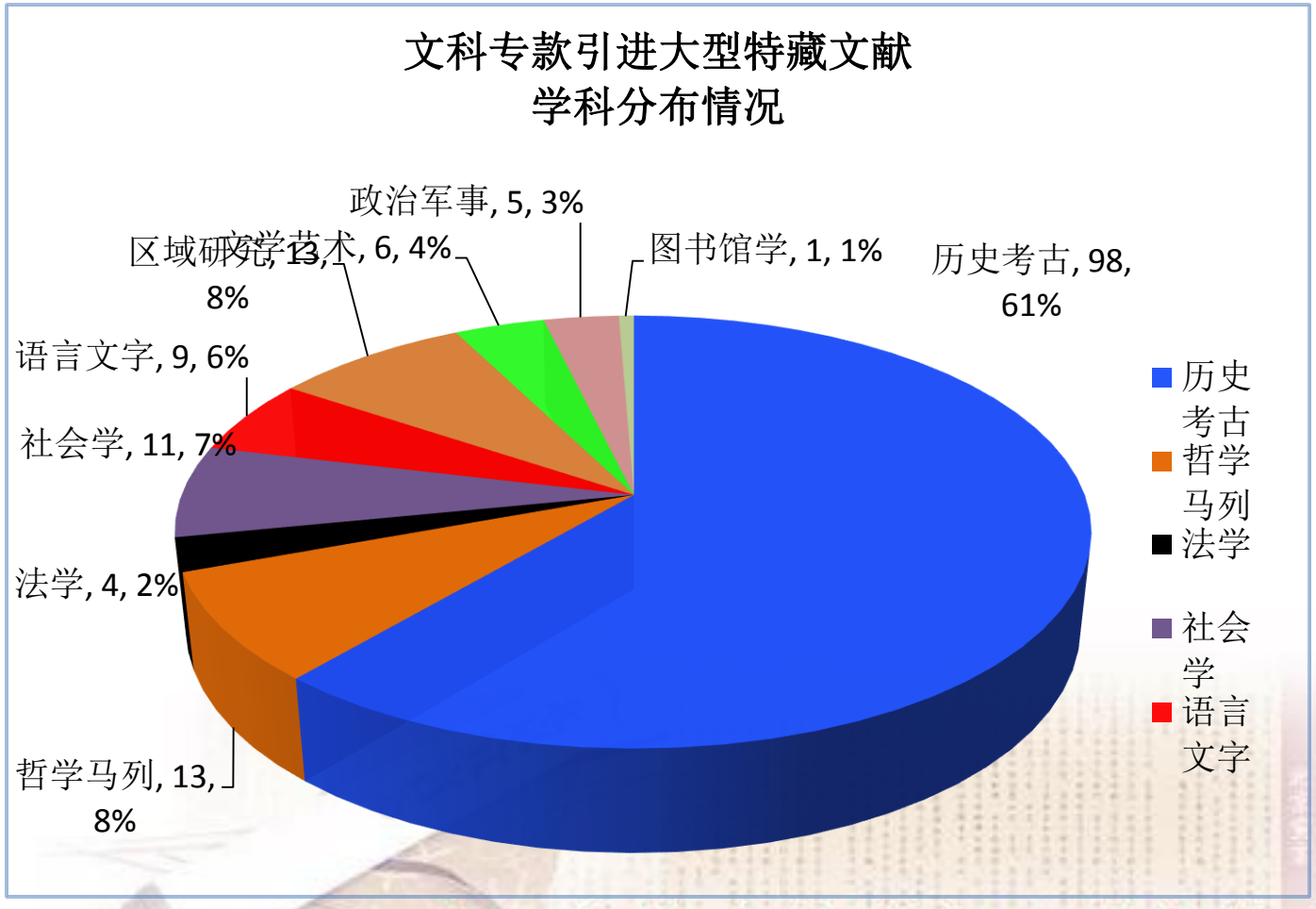
CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

CASHL大型特藏文献

学科	文献数量
历史考古	98
哲学马列	13
法学	4
社会学	11
语言文字	9
区域研究	13
文学艺术	6
政治军事	5
图书馆学	1



CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

大型特藏文献简介

特藏文献被公认为极具科研价值与收藏价值的珍贵文献，但受其价格昂贵的限制，诸多高校图书馆无力购买收藏。为了满足全国人文社科科研人员的研究需求，也为了弥补高校图书馆收藏的空白，CASHL于2008年度开始大批购入特藏文献。首批引进大型特藏文献多为第一手的原始档案资料，涵盖历史、哲学、法学、社会学、语言学、经济学等多个一级重点学科，涉及图书、缩微资料、数据库等不同介质，系北大、复旦、武汉大学等知名学者强力推荐。

历史类：

China inland mission　缩微平片 82卷

Church Missionary Society archive　缩微平片 262卷

China and Protestant Missions　缩微平片 1752张平片

British Intelligence on China in Tibet，1903-1950　缩微平片 576张平片

外務省帝国議会調書　胶卷 114卷

外交通信全覧（正、続）　图书 61卷

国家学会雑誌（復刻版）（图学学会杂志）　期刊（复刻版） 影印本？177卷

China through western eyes　缩微平片 153卷

FOREIGN OFFICE FILES: UNITED STATES OF AMERICA Series Two: Vietnam, 1959-1975　缩微平片 205卷

Corpus Inscriptionum Latinarum（拉丁铭文集成）　图书

Discoveries in the Judaean Desert（犹大沙漠古卷）　图书

Calendar of State Papers (英国国务档案纪事)　图书 97卷

Shanghai political and economic reports 1842-1943（上海:政治经济报告1842-1943）　图书 18卷

Japan political and economic reports 1906-1970（日本:政治经济报告1906-1970）　图书 14卷

Islam : political impact, 1908-1972（伊斯兰教：政治影响1908年至1972年）　图书 12卷

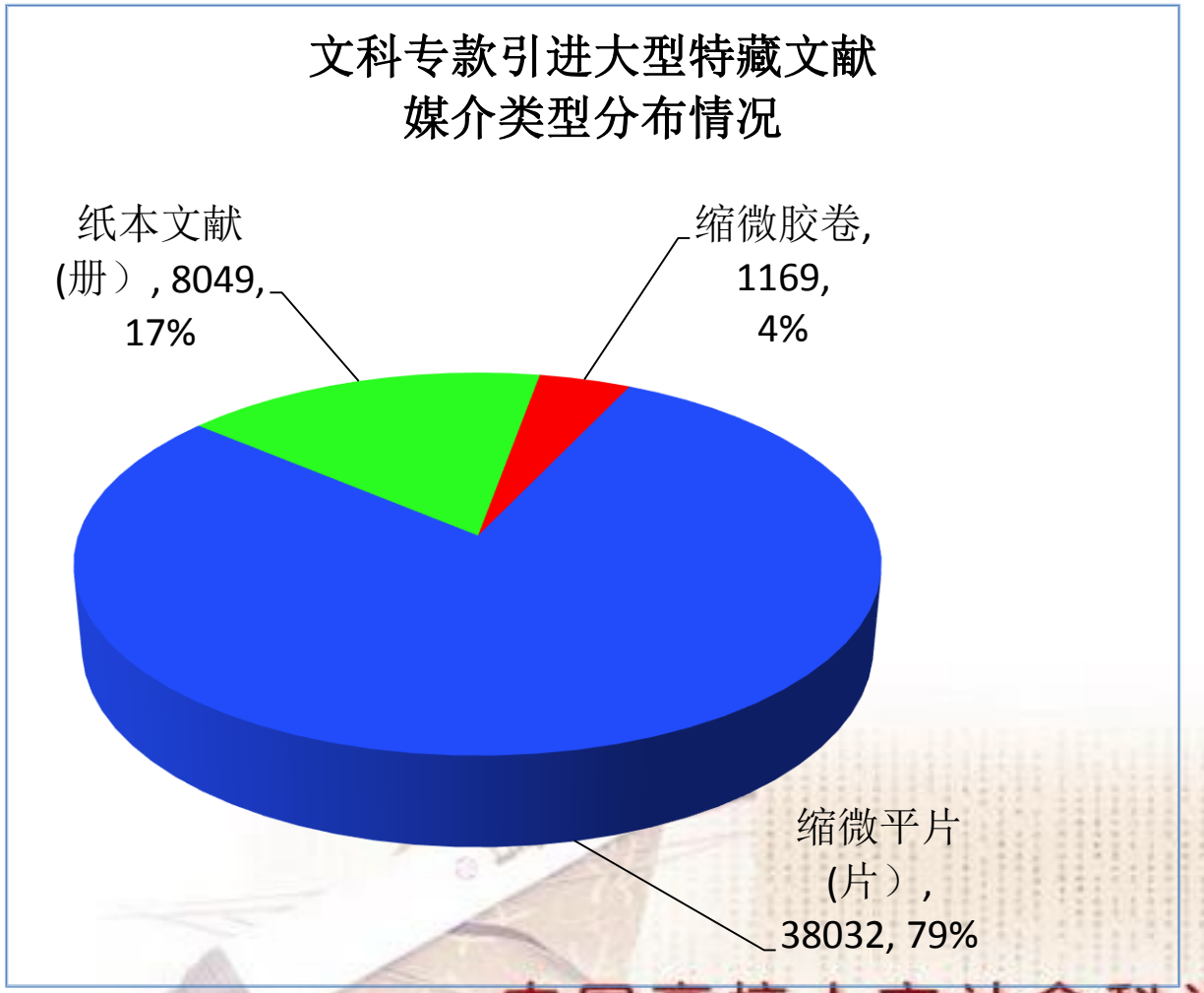
Soviet Union Political Reports1917-1970 (苏联政治报告1917-1970)　图书 12卷

The slave trade into Arabia 1820-1973(阿拉伯奴隶贸易)　图书 9卷

1	2	3
4	5	6
7	8	9
10	11	12

CASHL大型特藏文献

文献类型	推荐数量
纸本文献(册）	8049
缩微胶卷(卷）	2037
缩微平片(片）	38032



CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

CASHL “特藏++” 项目：知识挖掘

- 项目完整、全面揭示特藏内容，并通过问题/需求导向对文献内容进行揭示、报道或评介，从而提供专题服务；
- 项目以文献内容深度服务为主，有大型特藏所在研究领域教师或专家担任顾问，并提供需求；
- 在尊重知识产权前提下，项目成果能够提供在线服务，用户能通过互联网在线获取(包括文献传递)；
- 具备完整规范、深度挖掘内容的元数据（如文本索引目次数据）；

CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

CASHL“特藏++”项目：知识挖掘

首批试点项目名称及简介	学 校
项目名称：《日本外交文书》深度服务 《日本外交文书》153册	武汉大学
项目名称：卫理公会传教士信件缩微胶卷专题资源数字化加工与特色库建设 Missionary Files: Methodist Episcopal Church issionary Correspondence, 1846-1912(Africa, Europe, India, Malaysia)（传教士文件：卫理公会传教士通信，1846 年-1912 年（非洲，欧洲，印度和马来西亚））胶卷 28卷	中山大学

- 服务项目
- 资源建设项目
- 知识挖掘项目
- 积累数据项目

CASHL

中国高校人文社会科学文献中心

China Academic Social Sciences and Humanities Library

主要内容

1. 背景：CASHL与大数据
2. 基于馆藏分析理论的大数据应用
3. 基于需求驱动采购的大数据分析
4. 基于科研支持的大数据挖掘
5. 结语：面对大数据

CASHL

破题：信息服务的趋势：从超市到厨房的一体化渐进？

农场

超市

厨房

个体图书馆时代：
高质量学术出版物的采集收藏

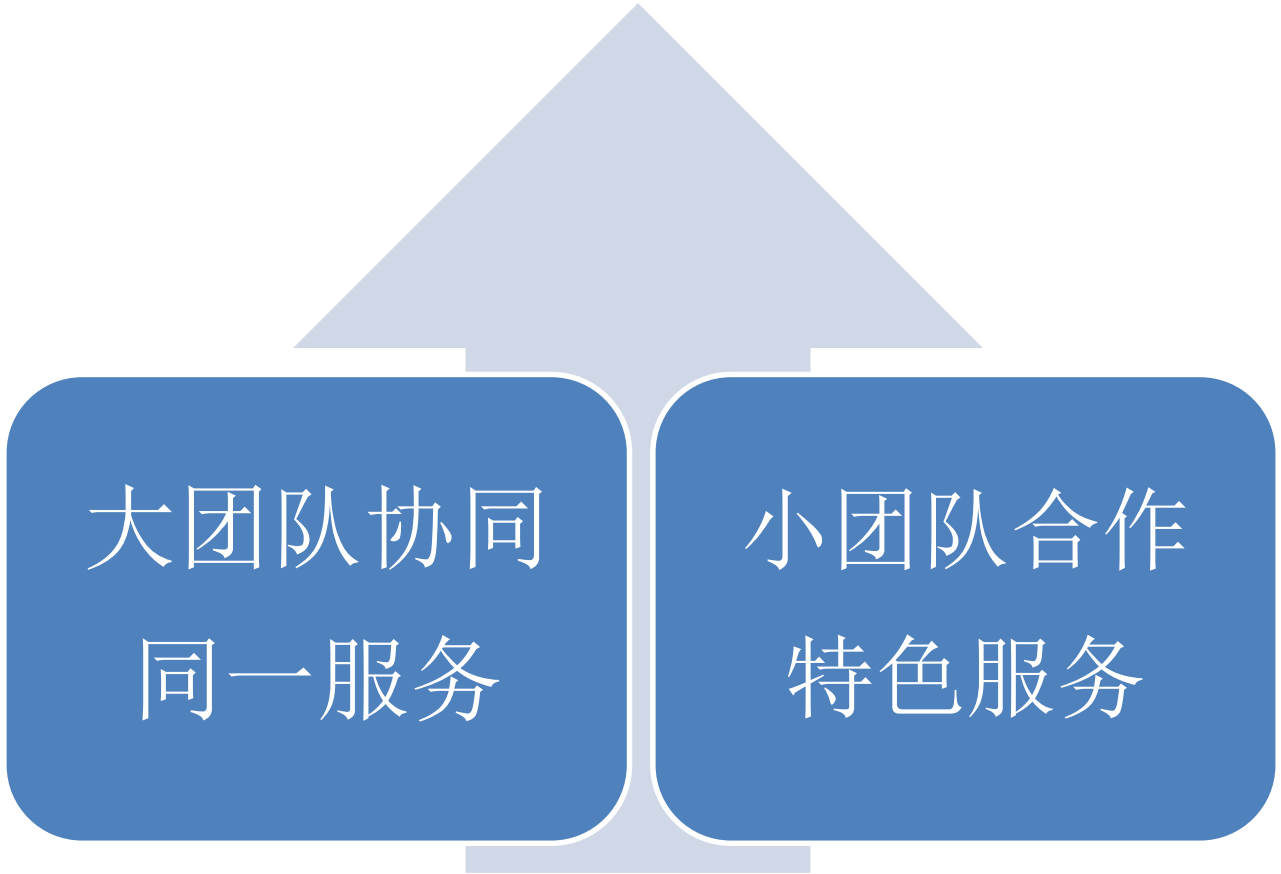
文献保障系统时代：
学术出版物的共建共享

知识发现时代：
为知识创新提供知识产品

馆藏服务 → 文献保障 → 知识发现

（改编自Y S CHI Scholarly publishing in the digital age & 张晓林“从图书馆学到信息发现

建设新的协同创新机制



CASHL

谢谢！