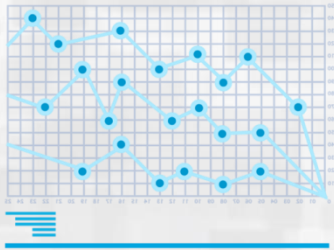
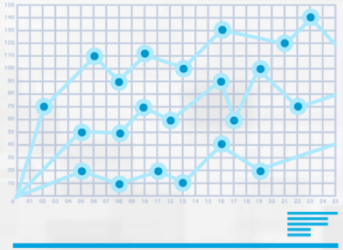
图书馆大数据利用与研究

国家图书馆 信息技术部 刘金哲

2018.06

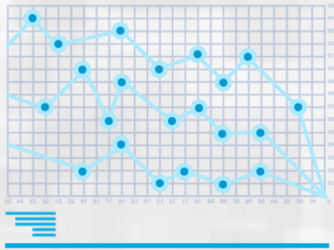


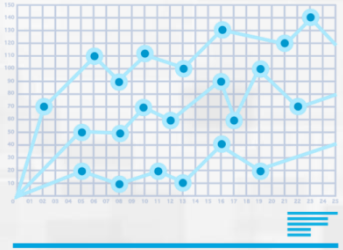


图书馆大数据利用与研究

国家图书馆 信息技术部 刘金哲

2018.06





大数据基础知识

1.1大数据的兴起和发展

2001 Gartner公司的研究报告首次出现“大数据”概念的提法

1980年，著名未来学家阿尔文·托夫勒便在《第三次浪潮》一书中，将大数据热情地赞颂为“第三次浪潮的华彩乐章”

2009“大数据”开始成为互联网信息技术行业的流行词汇。

2008年末，美国计算社区联盟发表了白皮书《大数据计算：在商务、科学和社会领域创建革命性突破》

2012年全世界开始高度重视大数据
世界各国政府制定大数据战略，互联网公司、各个行业也积极应用大数据。

2011年麦肯锡全球研究院发布报告《大数据：创新、竞争和生产力的下一个新领域》，第一次全方面介绍和展望大数据

2013年大数据元年

大数据的发展

国际数据公司（ IDC ）监测，全球数据量大约每18个月翻一番，意味着人类18个月产生的数据是此前所有数据的和。

根据联合国的研究报告，全球数据量到2020年要翻44倍，达到40ZB。

（存储容量单位：TB、PB、EB、ZB、YB、NB、DB）

1EB相当于13亿中国人人手一本500页书加起来的信息量

- **现存网页的总量已超过75亿，假定一秒钟浏览一页，那么看完现存的全部网页，需要至少230年时间，**
- **YouTube每分钟接收超过72小时时长的新增视频内容**
- **维基百科词条每日产生400万字信息，**
- **推特的推文每天产生4亿条，**
- **电子邮件每日2940亿收发量**

……

1.2大数据的产生因素

信息基础设施持续完善，为大数据的存储和传播准备了必要的物质基础

互联网公司领先的技术引领大数据的发展趋势。

互联网、物联网与智能移动终端是海量大数据的重要来源

云计算的集中管理和分布式访问使数据处理速度达到实用程度（<1秒），使大数据服务从概念变成现实。

人工智能算法和技术的发展是大数据的转化器和增值器，使得数据从一盘散沙转变为人们可理解的知识关联图谱。

1.3大数据的定义和特征

比较定义：McKinsey 将大数据定义为“超过了典型数据库软件工具捕获、存储、管理和分析数据能力的数据集”。

体系定义：美国国家标准和技术研究院认为：“大数据是指数据的容量、数据的获取速度或者数据的表示限制了使用传统关系方法对数据的分析处理能力，需要使用水平扩展的机制以提高处理效率”。

属性定义：IDC“大数据技术描述了一个技术和体系的新时代，被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值”。

目前，业内从数据的规模、变化频度、种类和价值密度等几个维度来对大数据的特征进行描述，认为大数据的特征主要体现在“3V”、“4V”。

3V

Garner、IBM提出

- 大容量（Volume）
- 多样式（Velocity）
- 高速率（Variety）

4V

国际数据公司（IDC）提出

- 大容量（Volume）
- 多样式（Velocity）
- 高速率（Variety）
- 价值（Value）

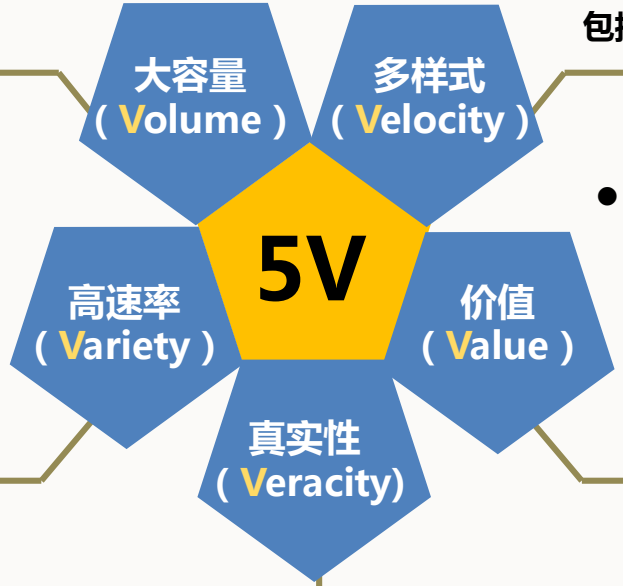
5V

中国电子技术标准化研究院发布的《大数据标准化白皮书（2014年）》认为

- 大容量（Volume）
- 多样式（Velocity）
- 高速率（Variety）
- 价值（Value）
- 真实性（Veracity)

- 指数据体量巨大，供分析的数据规模十分庞大，而这里的“大”只是一个相对的概念，对于不同的数据库或数据分析软件而言，其规模量级会有比较大的差别。

- 从数据类型的角度来看，数据的存在形式从过去的结构化数据为主转变成半结构化数据和更多的非结构化数据，从数据格式上来分，包括文本数据、图片、音频和视频数据等。



- 一方面是指数据增长的速度很快，
- 另一方面指数据以非常高的速率到达系统内部，需要系统做出快速反应，包括数据获取、数据访问、数据处理和数据交付等。

- 数据的价值密度较低，即在大数据范畴中，真正具有价值的数据仅占所拥有数据的很小一部分，需要从低价值密度的原始数据中进行深度挖掘和计算，总结出具备高价值的数据。

- 一方面，对虚拟网络环境下大量的数据需要采取措施确保其真实性、客观性，是大数据技术与业务发展的迫切需求；

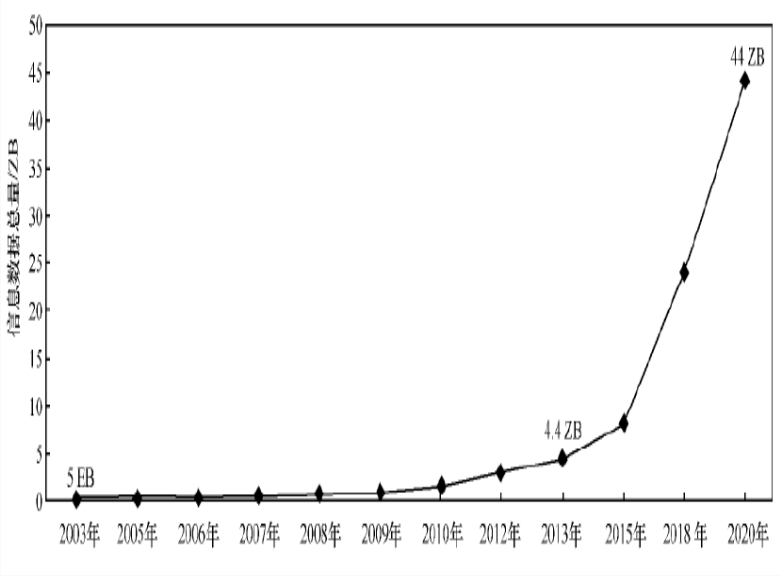
- 另一方面，通过大数据分析，真实地还原和预测事物的本来面目也是大数据未来发展的趋势。

1.4大数据的意义

大变革

大数据是一种新的价值观和方法论，人们思维和工作方式必须转变以适应大数据时代。。

----《大数据时代》

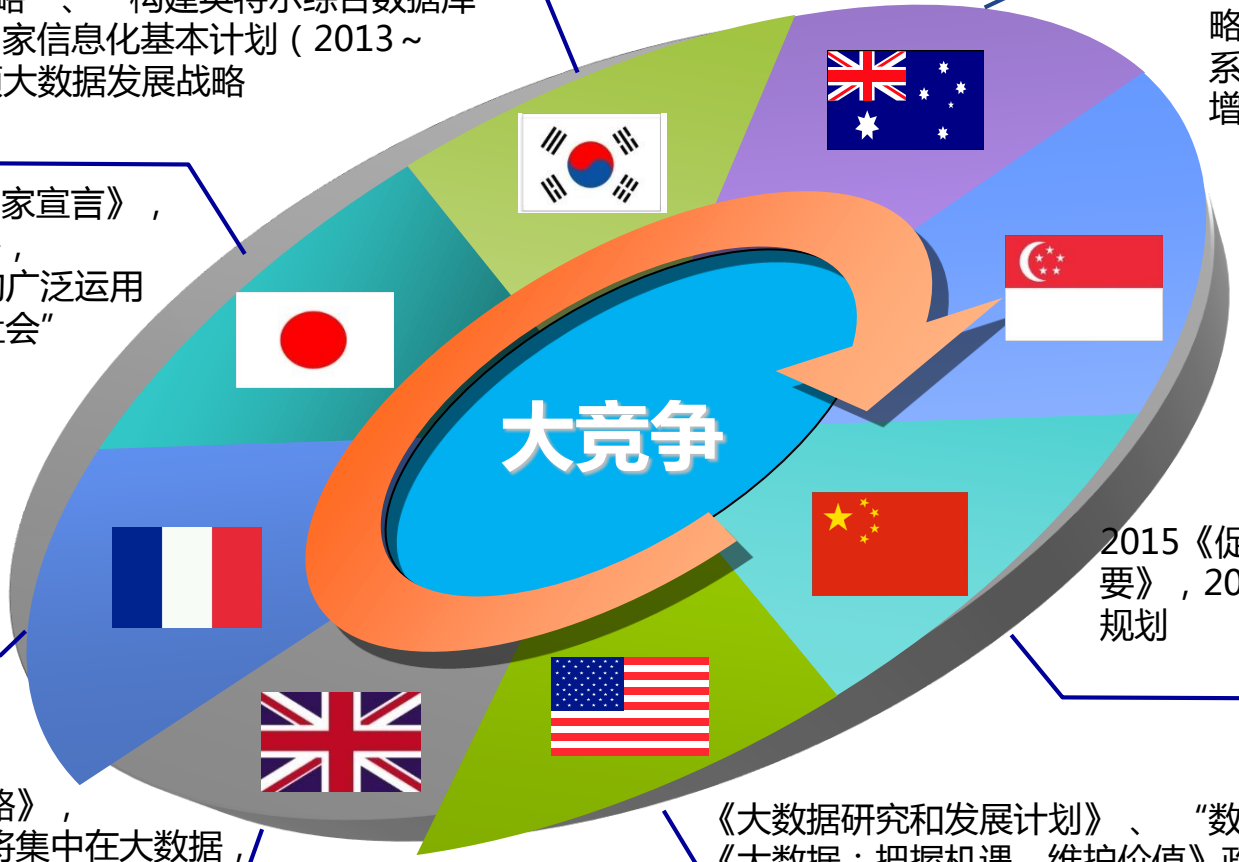


除了上帝，每个人都必须用数据说话。

----爱德华·戴明曾

不仅是人，整个世界都越来越数据化。信息革命深入发展，如潮的数据澎湃而至，数量之巨，种类之杂，来势之快，前所未有。

大数据的本质不在于“大”，而是以崭新的思维和技术去分析海量数据，揭示隐藏的人类行为等模式，由此创造新产品和服务，或是预测未来趋势。



“大数据中心战略”、“构建英特尔综合数据库”以及《第五次国家信息化基本计划（2013～2017）》等多项大数据发展战略

“公共服务大数据战略”，通过大数据分析系统提升公共服务质量，增加服务种类

《创建最尖端IT国家宣言》，日本新IT国家战略，“世界最高水准的广泛运用信息产业技术的社会”

“智慧国家2015年”计划，支持企业采用大数据技术，推动大数据生态的完善

“投资未来计划”，投入专项资金推动大数据技术发展

2015《促进大数据发展行动纲要》，2016推出大数据产业发展规划

《英国农业技术战略》，对农业技术的投资将集中在大数据，成立 “农业技术创新中心”

《大数据研究和发展计划》、“数据-知识-行动”计划、《大数据：把握机遇，维护价值》政策报告

IBM

开发DB2、Informix与InfoSphere数据库平台、Cognos与SPSS分析应用等知名产品。同时为Hadoop开源数据分析平台提供支持。

提供与大数据相关的硬件、软件以及服务，其最为知名的当数Vertica分析平台。“青岛-惠普软件全球大数据应用研究及产业示范基地”落户青岛。

Google

提供客户情绪分析、交易风险、产品推荐、等智能服务， 推出大数据产品BigQuery。

收购3家数据相关企业：Revolution Analytics、DataZen和VoloMetrix，实现结构化数据和非结构化数据的互操作。

Microsoft

未来的时代将不是IT时代，而是DT时代。对阿里云战略增资60亿元，用于国际业务拓展，云计算大数据领域基础和前瞻技术研发，以及DT生态体系建设。

推出百度大数据引擎：百度大脑，数据工场和开放云。

大挑战

大数据对保障国家信息安全和保护个人隐私都提出了极其严峻的挑战。

保护“数据主权”

“棱镜门”的曝光，让大家看到大数据时代维护国家信息安全、保护个人隐私所面临的严峻挑战。

保护个人隐私

研究发现，只要有4个时间点和位置的数据就能确定一个人身份，准确率高达95%。

1.4大数据的应用—公共服务

- 交通部门通过大数据分析实时路况，14年首次播报了春运迁徙实况
- 滴滴打车借助大数据团队的技术支撑，提升了城市出租车调度效率

大数据技术在医疗健康行业的用途：面向医生的临床辅助，面向居民的健康监测，面向药品研发的统计学分析、就诊行为分析等。

- 早在2009年谷歌就利用用户搜索记录预测H1N1在全美范围的传播，在时间上比美国疾控中心的预测还快两周。
- 现在流行的智能手表或者智能手环等，根据自身热量的消耗以及睡眠模式来追踪身体是否健康。

以京东为例

1、提供千人千面的精准营销

以京东为例

2、优化供应链中库存、配送的管理

运用自动补货系统，

设计出最优配送路线

以京东为例

3、提供个性化的智能服务

在网站搜索引擎和推荐引擎的服务当中利用大数据提供智能化的服务，帮助用户从海量信息中筛取所需信息。

分析该商品用户两次购买的平均时间，在下次接近这个平均时间时，给用户推荐相应商品。

大数据与互联网 —— 互联网信息获取、互联网交流沟通

如何从互联网产生的海量数据中找到需要的信，同时利用大数据理论和技术对网民搜索内容、习惯、爱好、行为、关键词等深入分析，为网站建设、搜索引擎技术改进提供依据。

比如，百度每天响应来自138个国家和地区的数十亿次搜索请求

国外社交网站Facebook、Twitter国内微博、微信等社交工具的不断壮大，每天都会产生大量的数据。通过对社交网络中的大数据进行分析，可以了解用户的思维习惯及其对社会的认知。

1.6大数据系统架构

价值链观点

数据生成	数据获取	数据存储	数据分析
关心数据如何产生	获取信息的过程 ●数据采集 ●数据传输 ●数据预处理	解决大规模数据的持久存储和管理 ●硬件基础设施 ●数据管理软件	数据检查、变换和建模并提取价值

层次观点

应用层

利用编程模型提供的接口实现不同的数据分析功能, 包括查询、统计分析、数据的聚类和分类等

计算层

将多种数据工具封装于运行在原始ICT 硬件资源之上的中间件中, 典型的工具包括数据集成、数据管理和编程模型等

基础设施层

由ICT 资源池构成, 可利用虚拟技术组织为云计算基础设施。

建设方式

从下而上	**从现有的数据资源出发，分析所能获得的数据资源，归纳整理出数据中心的数据体系结构，再往上推出数据中心的信息体系结构**
从上自下	**从应用需求出发，分析大数据平台要实现的目标，归纳出需要从平台中获取哪方面的信息，倒推出平台的信息体系结构和数据体系结构**
上下结合	**既考虑所能获得的数据资源，又考虑实际业务对数据中心的应用需求**

图书馆一般建议采用上下结合的方式

图书馆与大数据

2.1 图书馆中的数据

图书馆有什么样的数据呢？

1. 数字化资源

大量的由纸质图书转换的数字资源、电子书资源、数据库资源、各种声、图、视频影像资源 。

数量非常大，增长速度比较快，图书馆大数据主要组成部分。

2. 社交网络时代的非结构化数据

读者使用图书馆服务的过程中产生的来馆频次、活动范围、浏览历史、书籍借阅数据、网站点击数据、馆藏使用情况统计、读者的地理位置、搜索历史、搜索时间等这些非结构化和半结构化的大数据信息。

所占比重非常低，并且缺乏大数据的分析，数字图书馆很难融入企业等用户群体的细节服务

图书馆的数据是大数据吗？

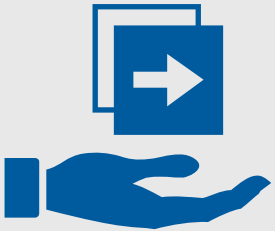
从数据体量看

以国家图书馆为例，截止到2018年4月

数字资源总量达
1649TB

外购数据库达
254个

各类元数据
3.6亿条



读者的注册总量达
392万人

推广工程统一实名用户库

用户达**1025.73万人**

电子图书
378万种

电子期刊
5.58万种

电子报纸
3164种

学位论文
706万篇

会议论文
657万篇

音频资料
113万首

视频资料
16万小时

特色数字化资源
2.9亿页

从数据的产生速度来看

图书馆在资源建设量和读者注册量不断增加的同时，读者行为数据的增长也非常迅猛，需要对这些数据快速处理，才能达到对这些数据实时应用的目的。

从数据的价值来看

图书馆拥有资源数据、书目数据以及读者数据，这些数据本身在读者服务中就会发挥很大的价值，各种读者行为数据和资源访问日志如合理的利用，也能产生巨大的价值。

图书馆的大数据能被利用吗？

随着“大物移云”技术的应用，图书馆产生了大量数据日志、服务记录等海量信息资源，加之前面所说的服务记录等，**半结构化、非结构化数据的增长远远超过了结构化数据的增长速度。**但这些数据的存储、处理和利用还处于相对缺失的状态。

通过对半结构化、非结构化数据的分析，可揭示以前很难或无法确定的重要相互关系，可以利用当前无法利用的数据，从而使数字图书馆资源的结构更加全面,更加顺应时代的发展,更加适应读者的需求，可以深入理解读者并给予智慧型的解决方案，最终提高数字图书馆的核心竞争力。

2.2大数据对图书馆的影响和挑战

（1）资源的挑战	
数据量大	如何收集？如何组织？如何存储？ 如何提高运算效率（在线/离线）？
非结构化	如何描述？如何关联？如何分析？
异源异构	如何整合？如何检索？如何发现？

（2）传播的挑战	
渠道终端多样化	如何保持产品与应用的开发？
内容服务网络化	如何描述？如何关联？如何分析？
数据环境开放化	数据如何开放？隐私信息的安全？版权保护？
应用体验新颖化	内容与科技的融合？功能的创新？

（3）服务的挑战	
服务内容	从信息服务向深度加工服务、智能处理服务、专题知识服务的转变
服务方式	更强调个人化、去中心化、实时化、虚拟化、智能化

（4）用户的挑战	
用户需求个性化	用户需求分析？内容精准推送？个性化服务？
内容需求知识化	知识体系的建立？知识价值的发掘？
信息检索实时化	文化资源和资讯的整合发布？信息快速定位？

（5）管理的挑战	
信息化设施	IT基础架构的整合与优化？统一数据环境和统一数据架构？
数据管理	语义网技术、搜索引擎技术、智能分类技术等，自动提取非结构化数据的检索信息
人员管理	如何提升数据收集能力、分析能力和创新理念等。

2.3国内外图书馆大数据典型案例

国外

美国各类公共图书馆、行业协会开展“数据无边界运动”

大英图书馆成立阿兰•图灵研究所，主攻大数据分析及应用研究

德国数字图书馆以1842家机构为支撑开放560万件资源

韩国体育观光部推动建设大数据平台，38家公共图书馆参与

国内外图书馆大数据典型案例

国内

1.发布年度分析报告
2.打造大数据展示墙
如上海图书馆年度悦读账单，浙江省公共阅读报告，深圳“图书馆之城”

中国知网提出“大数据出版”的概念、构建了机构知识管理与服务系统，实现信息服务到知识服务的提升

2.4 大数据与图书馆的融合点

1.实现跨系统的数据融合

随着图书馆业务的不断发展和读者服务方式的不断丰富，图书馆的数据生产方式也变得越来越多样化。

图书馆的数字资源

有自建的、外购的，
也有从互联网上获取的

书目元数据

有自行编目的，
也有整合外部提供数据的

读者数据

有到馆办证产生的，
也有网上注册产生的

读者行为数据

来自于提供读者服务的各个系统：借还书系统、门禁系统、资源发现与获取系统、读者门户网站、各种移动服务系统、RFID读者行为采集系统、网络监控管理系统，甚至是可穿戴设备管理系统。

2.海量数据的存储与处理

海量数据存储和处理是图书馆业务发展、也是数字图书馆发展的重要基础。

图书馆根据业务需求和数据情况，可以选择采用“大数据分布式架构”来实现。

在数据存储方面，采用“大数据的分布式存储技术”

将数据存储在分布式的存储集群中，就能够有效解决数据容量和性能均衡分布问题。利用数据分片和数据路由技术，实现数据的分布式存储，不但可以提高系统的可用性和存取效率，而且可以增强系统的可靠性，还易于扩展。

在数据处理方面，采用“大数据的分布式数据处理技术”

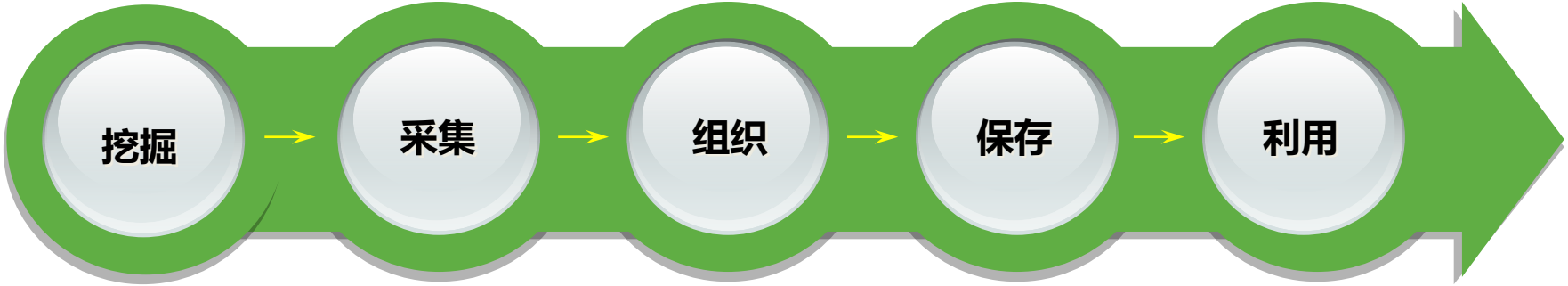
在对图书馆的海量数据进行数据处理时具有很大优势，能够突破集中数据处理方式的瓶颈，采用并行数据处理的方式，将数据处理任务分布到多台廉价服务器上运行，可提高数据处理速度，可以随着服务器的扩充线性增长。

3.创新资源建设新模式

建立资源整合管理平台支持多种异质文档及其元数据的管理，支持多媒体文档的存储、保管、检索和管理，支持结构化数据与非结构化数据的统一管理。

借鉴搜索引擎搜集网络信息的优势，开放集成网络环境下的各类数字内容，使自己真正成为信息社会的知识服务枢纽。

拓展对原生大数据和特藏大数据的挖掘、采集、组织、保存与利用建设基于大数据的特色数据库。比如美国国会图书馆开发的24个不同专题特色库、中国拓片专题。



4.提升图书馆的服务水平

对于图书馆来说，提升图书馆的服务水平，应该从**理解用户**和**满足用户需求**两方面入手，在这两方面，大数据都能够起到至关重要的作用。

一方面

利用大数据技术，分析挖掘读者行为信息，以便了解用户以及用户的行为特征和偏好，包括的群体特征和个性化特征

另一方面

在充分了解用户需求的基础上，利用大数据技术从图书馆丰富的馆藏中定位和寻找符合用户需求的资源，进而根据用户的阅读水平和偏好，对资源进行重新整理和组织，以便向用户提供更有针对性的个性化、精细化服务。

5.优化图书馆的业务流程

传统业务流程“以资源为中心”，以资源采购为起点，以资源借阅或利用为终点

图书馆“以资源为中心”的业务流程正在向“以读者为中心”的业务流程转变

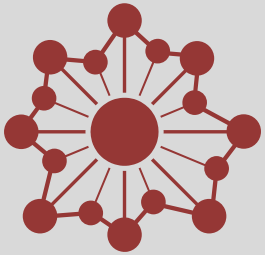
以“读者需求”为起点，以“读者服务”为终点

在大数据环境下，**用户行为分析和预测**已经成为可能，这使得图书馆员能够利用大数据技术**分析用户、了解用户**，从大量结构化、非结构化数据中**寻找读者的隐性诉求**，**对用户的阅读行为和偏好做出准确预测**，用以**优化图书馆的业务流程，改善传统文献的采、编、阅、藏政策**，同时**使数字资源的采访、加工、保存、发布与服务政策更加合理**，更大限度地发挥图书馆在服务读者方面的作用。

6.提供图书馆的决策支撑

首先，应用大数据对全量数据进行分析，能够有效的规避样本数据中细微错误被放大的可能性，能够**保证在决策过程中数据基础的准确性**

其次，大数据思维强调的是对海量数据的分析和挖掘，适当的忽略数据的精确度，而转向数据的宏观趋势，保证了**数据分析结果的全面性和宏观性**



最后，大数据能够分析和挖掘出数据背后隐藏的信息，不需要知道事务的因果关系，而是直接给出结论，因而能够**直接获得决策建议**

7.开展深层次的知识服务

时代发展，人们对图书馆的需求也逐渐从信息服务发展到知识服务。大数据在图书馆的知识服务领域能够发挥重要的作用。图书馆知识服务可以从以下几方面入手：

①借助大数据技术，开展“专题数字馆藏”、“虚拟参考咨询”、“个性化知识服务”、“学科知识导航”以及“定题知识服务”

②可以利用大数据技术获取互联网上的相关资源，扩充图书馆资源，作为知识服务的补充

③可以通过分析挖掘各类资源间的关联关系，形成知识网络，为读者提供可视化的知识网络服务

④可以加强对用户检索行为数据的挖掘利用，提升知识发现服务的针对性和有效性

2.5图书馆与大数据融合需要考虑的问题

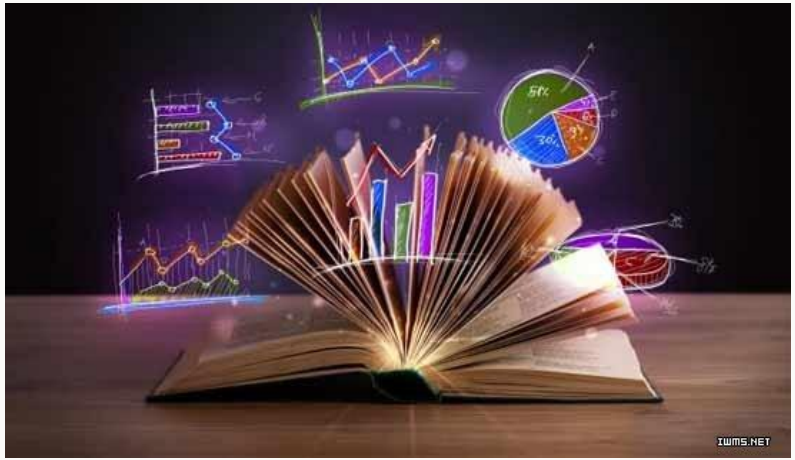
（1）数据收集

要对数据进行培养，即将过去传统的“被动式”的收集数据，变成“主动式”的数据收集，不仅对已有数据“单纯的”进行收集，而是要“生产”或“寻找”数据。这需要对业务的更深入的理解，也需要更高层次的决策。

数据收集不应仅考虑有什么数据就收集什么数据，而是要决定收集哪些数据，从解决问题的角度出发，去了解需要哪些数据，缺少哪些数据，哪些数据的精度还不符合我们的需求，从而主动收集、补充和生产这些数据。

（2）数据估值

针对海量数据数据千差万别、形式各异，在现有技术条件下几乎不可能完全收集、整理和处理完成的情况，**需要制定图书馆数据价值评估标准，对数据价值进行量化评判与评估，按照数据重要性进行排序与分类，并建立数据价值信息库。**不但有利于收集核心数据，更有利于了解数据价值分布情况，方便数据的后续收集保存与使用。



（3）数据安全

要考虑到数据的存储和访问安全。避免数据在保存、使用和传输的过程中被非法删除、修改和复制，从而造成数据的丢失、篡改和泄漏。在数据提供访问时要做好访问权限控制，保证数据在许可范围内被利用。

要考虑到隐私保护问题。在大数据应用过程中要区分数据的用途，规范数据的使用者，避免隐私泄露的潜在风险。需研究数据的保密问题，规范大数据开发利用行为，建立大数据使用规范和安全标准，以确保数据在内部流转、系统流转乃至外部流转的过程中是安全可控的。

国家图书馆在大数据方面的实践

建设文津搜索系统

（1）建设需求

国图建设文津搜索系统，没有采用商业化的成品软件，而是采用自主开发的方式建设，这是由建设需求决定的。根据国图对文津搜索系统的定位，该系统需要具备资源发现系统应具备的需求：

分类检索	高级检索	全文检索
二次检索	中英文通检	简繁体通检
同义词扩展	检索结果分类	检索结果排序
检索结果过滤	检索结果聚合	检索结果导引

个性化需求

数据整合与挖掘方面

- 需对国家图书馆通过自建、购买、征集、采集等方式所获取的各种类型的数字资源和传统文献资源元数据进行整合，实现元数据的本地存储，以实现高效的元数据检索服务；
- 需要基于统计分析和数据挖掘技术，向用户提供高质量的检索结果排名、相关检索和检索建议等。

系统性能与架构方面

- 需支持每分钟10万次并发和亚秒级响应，满足大并发、低功耗以及动态可伸缩的要求，可随着资源数据量和用户访问量的增加方便地进行扩展。

个性化需求

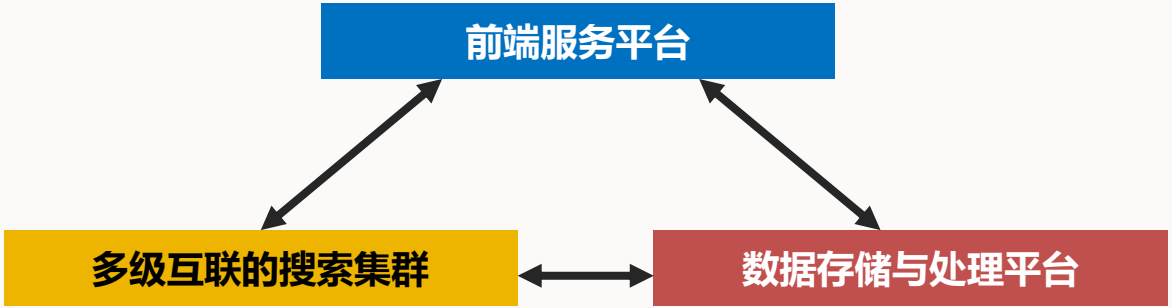
与其它系统的衔接方面

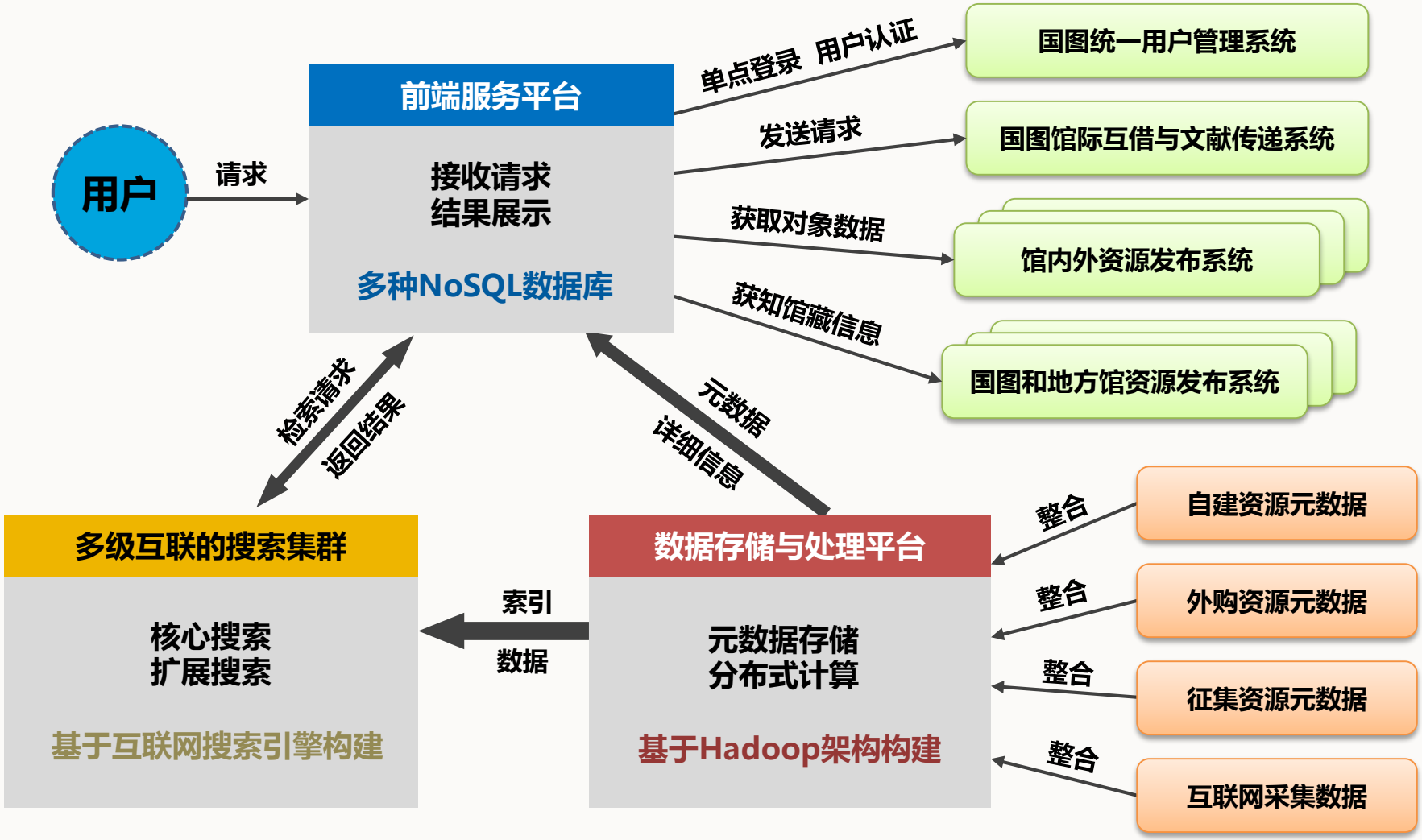
- 需与国家图书馆统一用户管理系统集成，实现**与国图其它读者服务系统之间的单点登录以及资源访问权限控制**；
- 需提供到馆内外资源发布系统的链接，实现**自动认证并能够直接跳转到资源的详细显示页面，方便读者获取对象数据**；
- 对实体文献，需提供**到各馆OPAC的馆藏链接**以及**到国图馆际互借与文献传递系统的接口**。

文津搜索系统没有采用传统关系型数据库技术构建，而是采用了**Hadoop分布式系统架构**和**各类NoSQL数据库**等大数据技术来实现。

（2）构建策略

采用大数据架构，该架构支持分布式服务和大规模数据处理，系统分成三个核心组件：前端服务平台、数据存储与处理平台以及多级互联的搜索集群。三个核心组件协同运转，来实现系统的预定目标。




用户
请求
前端服务平台
接收请求
结果展示
多种NoSQL数据库
单点登录 用户认证
国图统一用户管理系统
发送请求
国图馆际互借与文献传递系统
获取对象数据
馆内外资源发布系统
获知馆藏信息
国图和地方馆资源发布系统
检索请求
返回结果
元数据
详细信息
多级互联的搜索集群
核心搜索
扩展搜索
基于互联网搜索引擎构建
索引
数据
数据存储与处理平台
元数据存储
分布式计算
基于Hadoop架构构建
整合
自建资源元数据
整合
外购资源元数据
整合
征集资源元数据
整合
互联网采集数据

前端服务平台

由多台服务器组成的集群

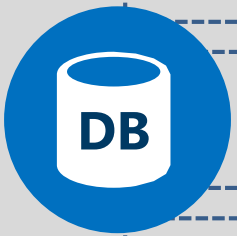
前端代理服务器和**缓存服务器**提供负载均衡以及静态文件的缓存

Web Server服务器分发检索请求并返回检索结果

检索辅助服务器提供相关检索和检索推荐服务

日志收集服务器负责收集日志

采用三种NoSQL数据库保存与读者交互的数据信息



Redis数据库保存读者的检索历史信息和翻译词库的信息

MongoDB数据库保存检索热词信息、标签云信息以及读者个人设置信息

Cassandra数据库预先存储完整的元数据信息，包含书封图片数据、到对象资源的URL信息及书评信息

基于互联网成熟的搜索引擎构建，是一个分布式搜索引擎架构

每一部分包含若干组结构相同的服务器，每一组服务器又分为检索根节点、检索父节点和检索子节点三个层级。**通过这种多级互联的方式实现负载均衡，保证高并发情况下的检索效率**。

数据存储与处理平台

采用HDFS分布式文件系统

- **存储网络采集信息、日志文件、元数据中间文件和生成的引文件**

采用MapReduce程序

完成数据处理功能

① **对元数据进行一系列数据处理，包括数据清洗、查重、转换、合并、挂接书封目次信息以及建立数据关联等；**

② **完成资源重要性计算，以提高检索质量；**

③ **构建核心索引和全文索引；**

④ **进行日志数据的统计分析，生成统计数据；**

⑤ **对用户的行为数据进行分析挖掘，用于向读者推荐相关资源。**

采用HBase数据库

- 存储需要整合以及整合后的元数据

（3）应用成效

2012年上线运行以来，目前已经拥有了超过3.6亿条元数据，绝大部分的数字资源都能提供到对象数据的访问链接，提供一站式的资源发现与获取服务。

系统的设计目标是具备每分钟十万次以及峰值每秒钟一万次的检索请求处理能力，在实际服务中基本能够做到即搜即得。

实现了对用户行为数据的分析挖掘，实现个性化推荐。

分布式架构，动态可伸缩设计，服务器冗余，系统和数据安全可靠性高

3.2 数据管理与分析平台

项目建设目标

采集国图重点系统的数据，利用大数据技术构建大数据分析的应用模型，从不同维度和精度对采集的数据进行统计、分析和挖掘，构建用户和资源的“数据画像”，并通过对资源利用状况、用户行为等信息进行揭示，为领导层提供决策支持，为资源采购、服务布局等业务优化提供参考，为读者的个性化、精细化服务提供数据支持。

- 汇集电子资源及传统文献的元数据，可作为描述“资源”属性的数据基础
- 保存了读者检索、查看、在线阅读等行为数据

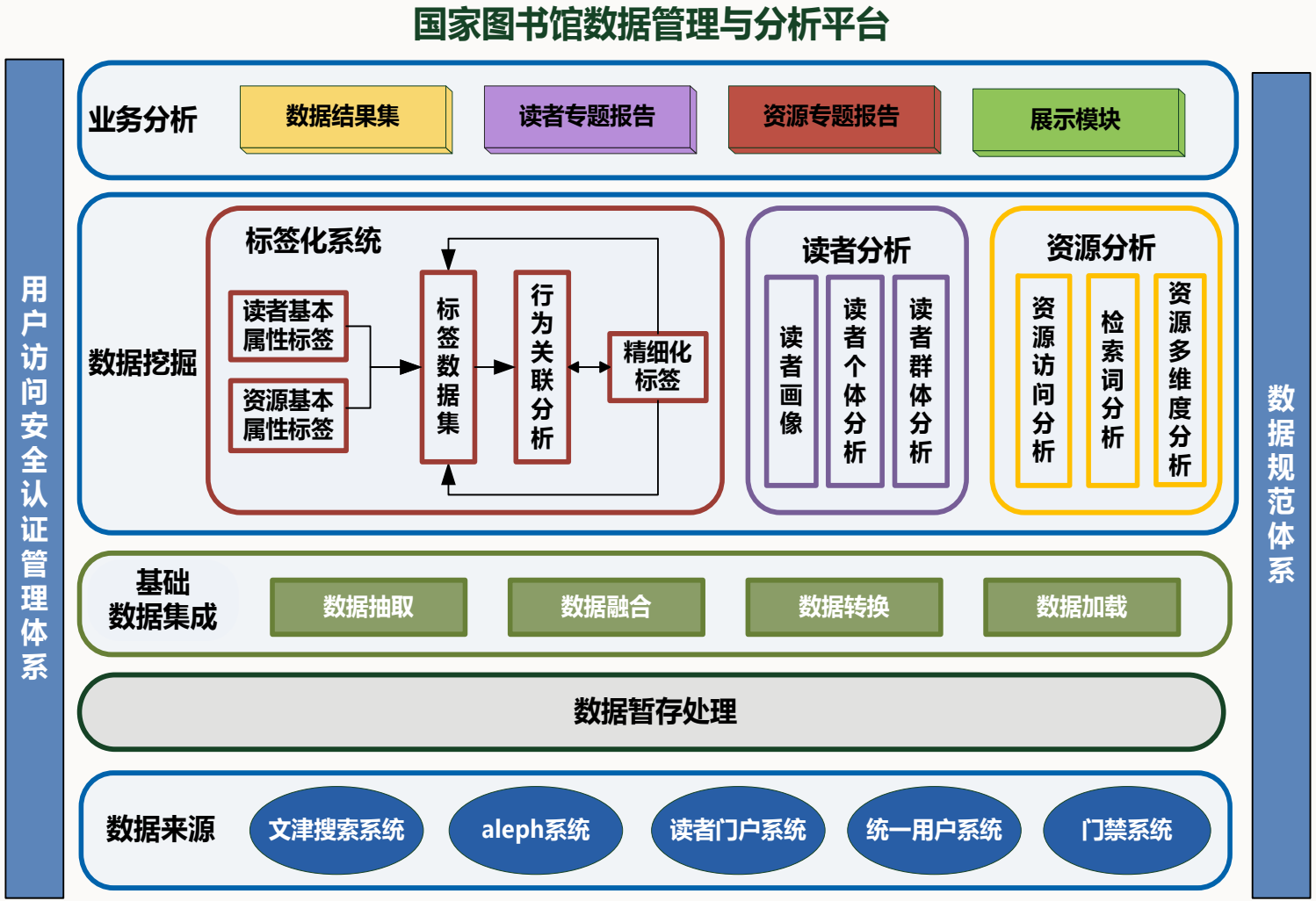
- 保存有最全的读者信息，可作为描述“读者”属性的数据基础
- 系统中还保存了读者注册、登录等行为信息

- 支撑着采访、编目、opac检索和流通等传统业务
- 存有传统文献书目元数据
- 保存着读者借还书、预约和续借等行为数据

- 提供各类中外文外购资源库和自建特色资源库的资源访问服务，
- 存有资源书目元数据、自建资源的全文对象数据和读者阅读历史等信息。

- 包含读者在各个阅览室的刷卡记录，
- 是读者到各阅览室阅览书刊行为的体现。

系统整体架构


国家图书馆数据管理与分析平台
用户访问安全认证管理体系
业务分析
数据结果集
读者专题报告
资源专题报告
展示模块
数据挖掘
标签化系统
读者基本属性标签
资源基本属性标签
标签数据集
行为关联分析
精细化标签
读者分析
读者画像
读者个体分析
读者群体分析
资源分析
资源访问分析
检索词分析
资源多维度分析
基础数据集成
数据抽取
数据融合
数据转换
数据加载
数据暂存处理
数据来源
文津搜索系统
aleph系统
读者门户系统
统一用户系统
门禁系统
数据规范体系

平台在建设过程中，用到了各种分析方法来进行建模

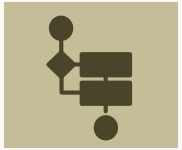
序列分析法

采用时间序列进行分析，将读者行为数据按照时间的顺序排列，从而研究随机数据序列所遵从的统计规律，形成国家图书馆服务洞察分析，分析内容主要包括读者注册量、到馆借阅、在线登录和在线阅读等行为。

聚类分析法

按照读者和资源两大类的基本属性进行聚类。其中读者聚类是根据读者借阅行为，将借阅某一学科或专业的读者按照读者基本属性进行聚类，而资源聚类则是根据被借阅情况，将某一类读者喜欢借阅的资源按照基本属性进行聚类，从而分析其中所包含的读者与资源之间互相联系的规律。

项目建设成果



1. 针对数据采集、数据抽取、数据融合和数据转换等一系列数据处理流程，形成了统一的数据处理规范

2. 构建了一套读者与资源的标签体系

3. 建设了数据管理与分析平台，对数据统计、分析和挖掘成果进行可视化展示

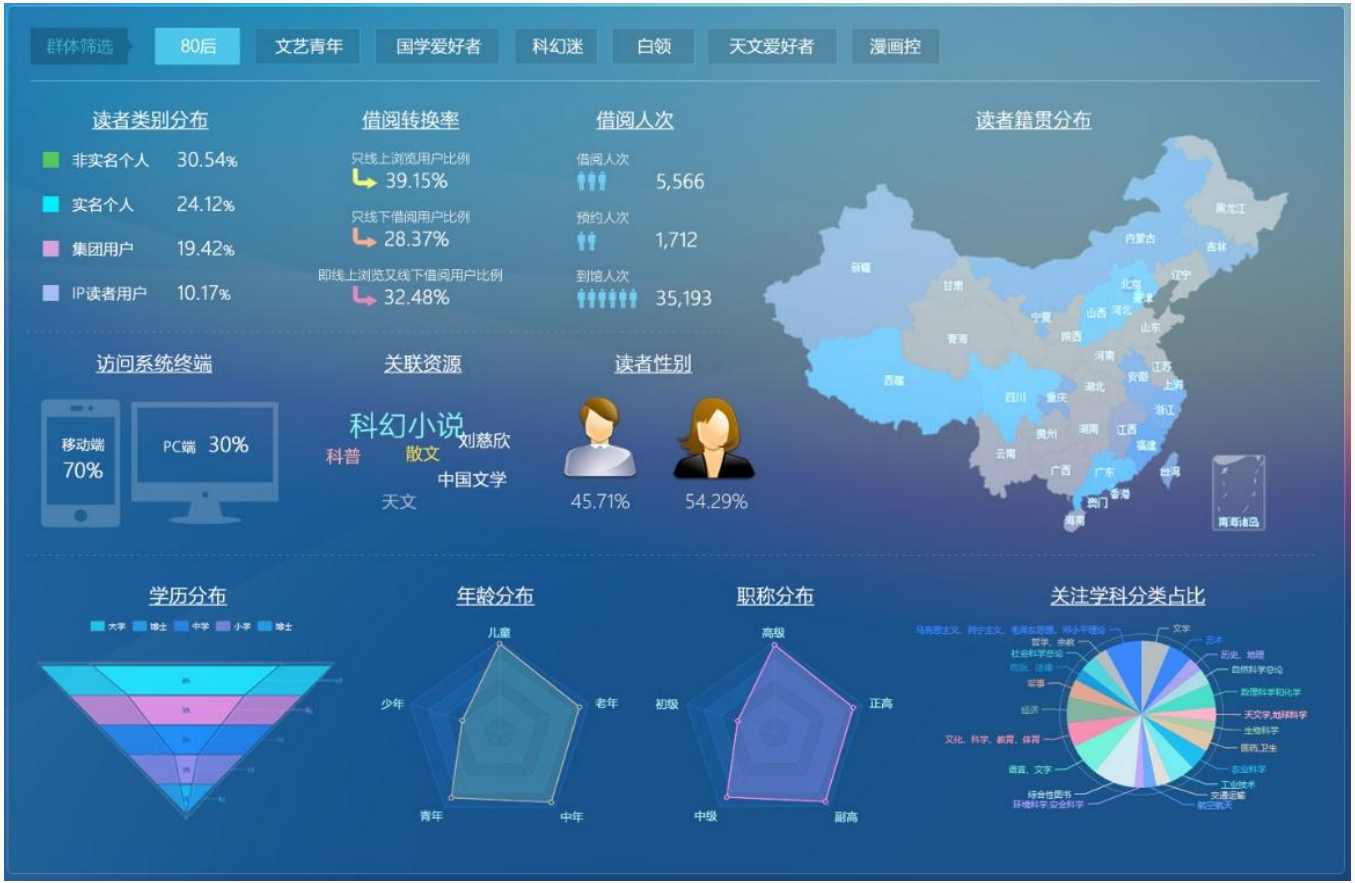
4. 开发了自定义分析和迭代分析工具，可用来开展专题业务分析

平台部分建设效果示意如下：

读者个人画像

平台部分建设效果示意如下：

读者群体画像



平台部分建设效果示意如下：

资源个体画像

平台部分建设效果示意如下：

资源全体画像

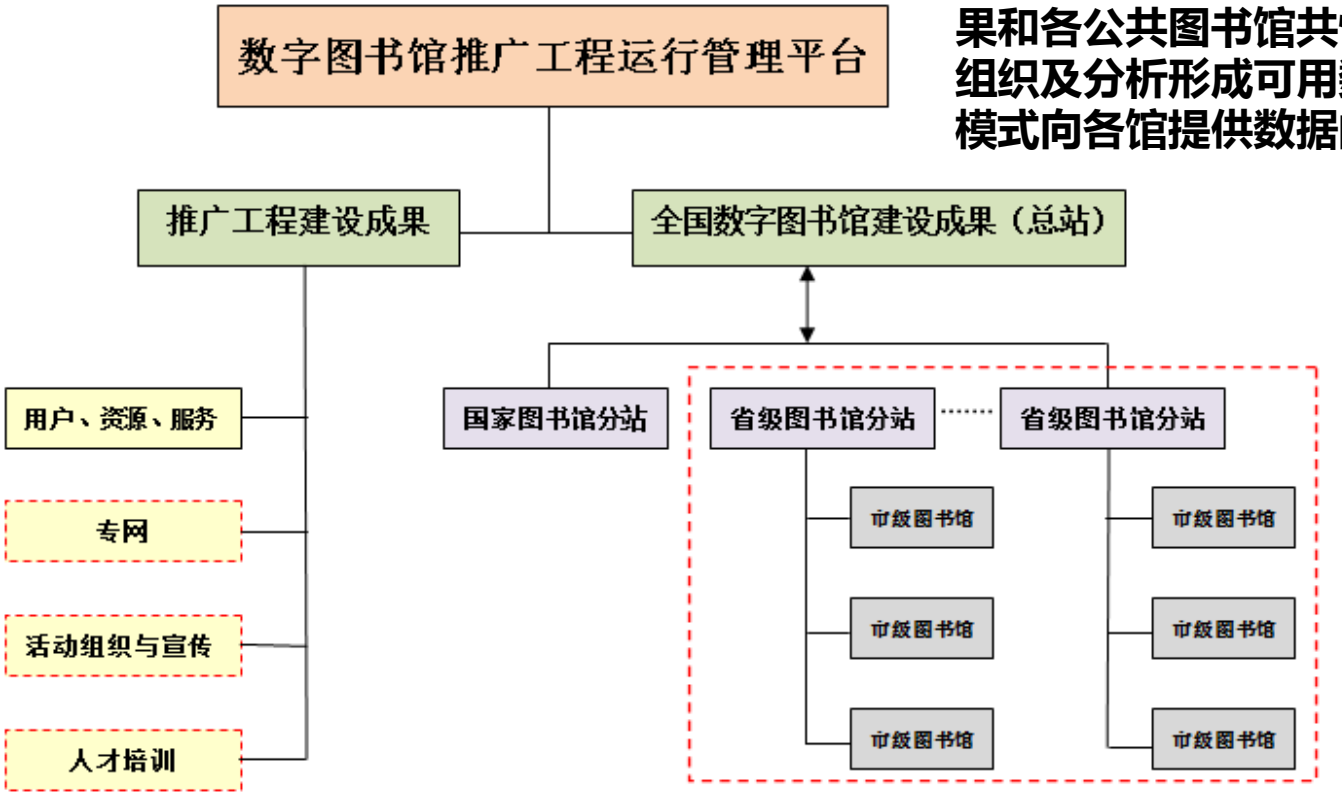
3.3数字图书馆推广工程运行管理平台升级项目

为了带动全国公共图书馆开展推广工程大数据整合，扩大数据采集的范围，深化采集数据的力度，提升采集数据的质量，我们于2018年初启动了推广工程运行管理平台升级项目。

建设目标

采集国家图书馆以及全国范围内省、市、县公共图书馆文献资源建设、用户情况、服务效能、信息化设施、支撑保障等共性数据进行展示，既有固定频率的统计数据，也有实时数据统计，客观地反映数字图书馆推广工程建设与服务的成果，以及在工程的带动下全国范围内各级公共图书馆数字图书馆的建设现状。

系统架构



搭建云数据中心，汇集采集到的推广工程实施成果和各公共图书馆共性的数据，然后进行数据的组织及分析形成可用数据，在应用层面以B/S的模式向各馆提供数据的可视化展示和调用服务

对全国公共图书馆数字图书馆建设与服务成果的展示部分采用总分站建设的形式，在国家图书馆同时建设总站和分站，在有条件的省馆分站，发布区域数字图书馆建设个性化的建设成果和服务情况，最终形成“国家—省—市”三级展示模式。

模块化设计

具有良好的可扩展性和伸缩性，灵活地增加要展示的数据项，方便支持新的分站交换信息/数据的需要；

兼容性强

操作系统、中间件、数据库和应用系统，满足实际工作中的全方位数据采集及展示需求。支持自动采集和手工采集方式，

标准接口

标准的Web Service接口、采集客户端C/S程序或Windows服务来接收第三方系统的数据，

。

建设进展　已经完成推广工程总站、国图分站的建设工作。

以国图分站为例

实现主要手工统计报表的自动导入和基本的数据处理

实现了主要服务类设备和系统的日志和数据自动分析和获取

基础数据	**馆情数据、资产数据、信息化基础设施类的数据**
用户类数据	**办证用户、在线用户、无线网用户、阅览室用户、电子阅览室用户、微博微信QQ等第三方用户、体验区用户、手机用户等**
资源类数据	**实体文献、外购数字资源、自建数字资源、征集和联建资源、移动资源、长期保存和灾备资源等**
服务类数据	**实体文献流通情况、外购数据库访问情况、自建和移动资源的访问情况等**

以国图分站为例

完成了与馆内主要业务系统的对接

- 自动化集成系统
- 统一用户管理系统
- 唯一标识符系统
- 发布与服务系统
- 手机门户和移动阅读平台
- 文津搜索系统
- 国图公开课
- 推广工程培训平台
- 交流平台
- 读者门禁与流量管理系统
- 电子阅览室读者管理系统

等等

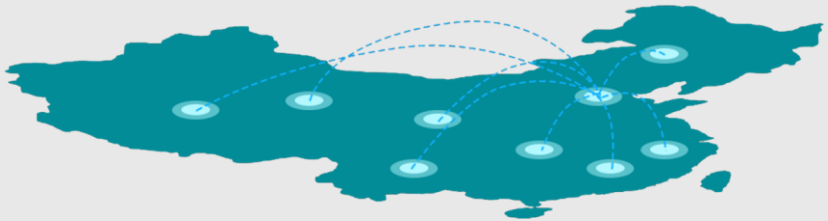
以国图分站为例

平台同时收割了全国2900多家省、市、县级公共图书馆共性的数据

一．**业务类数据：**包括实体文献建设情况，包括编目、入藏等；数字资源建设与发布情况等；

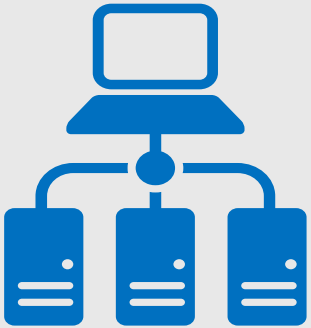
二．**服务类数据：**包括实体文献流通情况、数字资源服务情况、移动服务建设与访问情况；不同渠道、不同平台的用户基本情况、用户访问情况；公共图书馆特色服务的访问情况；宣传和活动情况等。

三．**保障类数据：**信息化实施情况、人才和培训情况等。



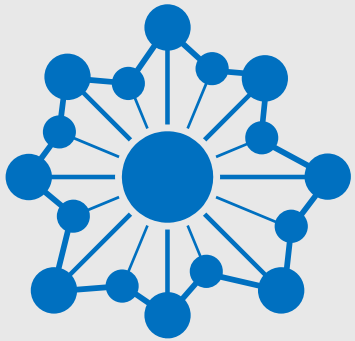
发展思路

与数据管理与分析平台对接，把收集到的数据和日志通过开放的接口提供给数据管理与分析平台使用，实现采集、处理、分析等大数据链条的融汇和贯通



提供数据传递的接口，支持多种方式供其他系统调用，比如体验区、多媒体展示系统、其他网站等

未来系统将在有特色、有条件的公共馆进行复用，建立不同的分站,收集地方馆个性化、特色化的数据进行分析展示，从而能够全面展现全国范围内公共图书馆数字图书馆建设与服务成果。



- 形成一个超大、全面的数据池，在此基础上可以进行大数据整理和分析；
- 既支持单个分站纵向的数据分析，又支持多个分站横向的数据比对；
- 既支持馆内业务数据的分析，又支持对用户行为数据的分析；
- 为图书馆内的业务发展、服务提升提供决策依据，也能全面反映图书馆行业的发展历史和发展趋势。

结束语

凡是过去，皆为序曲。

——莎士比亚

预测未来最好的方法就是去创造未来

——林肯

大数据提供的只是参考答案非最终答案。

无论数据时代还是大数据时代，探索和创新精神都是最珍贵的

谢谢！