“大数据时代”环境下数字档案信息挖掘与传播实践

研 究 报 告

《“大数据时代”环境下数字档案信息挖掘与传播实践》课题组

2014 年 12 月

目　录

第一部分：档案大数据概念与涵义 3
一、大数据 3
二、档案大数据 4
三、大数据时代 4
四、Hadoop 框架 4
五、数据仓库 5
六、数据挖掘 6
七、智慧档案 6
第二部分：档案大数据信息挖掘的意义与作用 7
一、有利于提高档案信息知识转化能力 7
二、有利于提升档案信息知识服务水平 8
三、有利于推动智慧档案馆信息资源建设 8
四、有利于推动大数据在档案领域的应用 8
第三部分：档案大数据发展环境与背景 10
一、国内发展现状 10
1、首个大数据研究服务基地揭牌 10
2、信息技术公司是发展大数据应用技术的主力 11
3、大数据在公共服务领域的实践 11
4、图书档案界现状 11
二、国外发展现状 12
1、英国——大数据的积极拥趸者 13
2、美国——大数据的策源地和创新引领者 15
3、澳大利亚——利用大数据推动公共服务建设 17
第四部分：档案大数据信息挖掘模式研究 19
一、确定档案大数据挖掘目标 19
1、以存储为目的数据挖掘 19
2、以利用为目的数据挖掘 20
二、遵循档案大数据挖掘原则 21
1、合规性原则 21
2、可扩展性原则 21
3、选择性原则 21
4、准确性原则 22
三、 建立档案大数据“数据仓库” 22
1、互联网档案数据采集 23
2、政务网档案数据采集 26
3、 增量数字档案的接收 28
四、档案大数据信息挖掘流程 28
1、档案大数据信息挖掘前期规划 28
2、建立档案大数据信息存储架构 29
3、进行档案大数据信息预整理 31

4、建立档案大数据分析模型 32
5、建设档案“大数据”数据挖掘平台 34
6、进行档案大数据挖掘结果评价 35
第五部分：档案大数据传播模式研究 37
一、数字档案馆信息服务指导思想 37
1、以信息安全为前提 37
2、以用户需求为指引 38
3、以自主创新为方法 38
4、以人才培养为重点 38
5、以资源开放为保障 38
二、大数据环境下数字档案信息服务特征 38
1、应能够直接提供智能化决策信息 38
2、将实现真正的档案信息个性化服务 39
3、档案信息资源达到全面整合 39
三、大数据环境下档案资源服务的新途径 39
1、云计算 40
2、移动互联网 40
3、社交媒体 40
第六部分：档案大数据信息挖掘实证研究 42
一、档案信息网一般框架及信息资源构成分析 45
二、沈阳市档案网站主体框架及信息资源构成分析 46
三、抚顺市档案网站主体框架及信息资源构成分析 47
第七部分：档案大数据信息传播实证研究 51
一、系统设计目标 52
二、系统设计原则 52
三、系统建设总体内容 53
四、系统区域设置 53
五、系统结构设计 54
六、系统功能模块设计 55
七、数据库设计 56
八、客户端界面设计 56
九、“档案史料”系统的开发与实现 59
1、相关技术 59
2、开发运行环境描述 64
3、 WEB 页面与管理平台实现 65
4、数据库连接实现 65
5、 WEB 网站子系统的实现 66
6、管理平台子系统的实现 67
十、APP 手机客户端子系统的实现 68
1、导航切换具体实现代码 70
2、翻页动画效果具体实现代码 71
3、客户端与后台数据接口实现 72
第八部分：档案大数据信息挖掘与传播面临的挑战 75
一、海量数据筛选的挑战 75

二、资金来源的挑战 75
三、传统信息传播方式的挑战 76
四、大数据人才队伍建设的挑战 77
五、传统观念与价值体系的挑战 77
六、技术创新与应用的挑战 78
第九部分：大数据时代档案部门建设应对策略 79
一、制度保障 79
二、人才保障 79
三、技术保障 80
四、资金保障 80
五、安全保障 81
第十部分：综述 82
参考文献 83

档案大数据时代信息挖掘与传播

档案是人类历史最重要的载体之一，从闻名于世的商代甲骨文档案，到其后的金文档案、石刻档案、简犊档案、以至现在作为各级档案馆主要保存对象的纸质档案，智慧的人们通过符号刻画对历史进行文化、经济、政治等一系列活动的记载。随着社会的进步，人们记录生产生活的手段越来越多样化，从而使档案有了新的载体——电子档案。伴随这一新型载体的出现，传统档案提供利用方式已越来越难适应当今社会发展需要。特别是当今“大数据”的概念已经渗透到每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。在这种大环境下，档案利用服务也开始由提供数据信息向提供知识转变。然而，知识不是自然生成的，知识也不是简单地存在于信息集合中，特别是蕴含在档案中的知识，需要经过抽取和挖掘才能展示出来。依靠人工抽取、挖掘知识，在传统纸质档案时代，档案数量不多的情况下还可实行，但在大数据时代海量档案存在、非结构化数据盛行的今天则会变得心有余而力不足。如何盘活这些珍贵的档案数据资源，使其为国家治理、政府决策乃至个人生活服务，是大数据在档案领域应用的核心议题。

课题主要从档案大数据“数据仓库”建立、大数据信息挖掘与传播模式、档案数据信息采集原则、以及大数据信息挖掘和传播实证研究等几个方面全面论述现阶段档案大数据在档案信息资源建设方面能够发挥的作用，同时为处在大数据时代档案事业的发展方向，以及

如何获取大数据带来的巨大价值提供了可靠的科学依据。特别是在大数据实证研究中为全国档案部门示范了现阶段大数据档案信息挖掘可以采用的方式方法，具有一定的现实意义和指导作用，且带有前瞻性。

第一部分：档案大数据概念与涵义

当今世界快速发展将我们带入了一个全新的时代。物联网、云计算、移动互联网、手机、平板电脑、PC 以及遍布地球各个角落的各种各样的传感器，成为巨大的数据来源。伴随信息处理的快速提升，人类社会的“大数据”时代正扑面而来。大数据已经成为当前学术界和产业界的研究热点，正影响着人们日常生活方式、工作习惯及思考模式。

纵观档案界，智慧档案这一理念的提出得益于大数据技术的迅猛发展，档案数据仓库在大数据风暴中的作用日益凸显。纵观全国，各级档案部门在数字档案馆和档案数字化建设中不断取得成果，其发展方向已经自觉或不自觉地朝着大数据迈近。“档案大数据”这一概念随着智慧档案的提出应运而生，它与未来档案现代化建设是相辅相成、相互促进的。数字档案馆和档案数字化的建设将为档案大数据提供数据基础和实践经验，反过来，档案大数据的思路也将指引着档案工作的繁荣和发展，使档案工作更趋科学合理，更具预见性和可持续性。

一、大数据

大数据又称海量数据，数据量级超越传统理念达到前所未见得程度，一般指的是所涉及的数据量规模巨大到无法通过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成有效的能够用于决策的信息。面对这种超大规模的数据需要更适合的、更高效的、更科学的、更被普遍应用的工具来处理。与传统数据相比，大数据的特点主要体现在数据量体量庞大、数据类型丰富繁多、数据来源广泛等几个方面。根据业界分析调研机构 IDC 的定义，大数据具有 VOLUME（容量）、VARIETY（多样）、VELOCITY（速度）、VALUE（价值）四大特征。

1、 容量(Volume)

既数据体量庞大，起步达到 PB 级别，具权威机构预测，到 2020 年，全球数据量将扩大至少 50 倍以上，而且这种趋势本身也在呈现上升式的变化，单一数据集的规模范围已经突破 TB 达到 PB 以上量级。大容量的同时数据类别也呈现出非结构化的特点，非结构化数据的增长速度比结构化数据快 10 倍以上。

2、多样(Variety)

多样性体现在结构化数据、半结构化数据和非结构化数据。数据多样性的产生是由于新型多结构数据，具体体现在文字、音频、视频、图片、网络日志、地理信息、手机通话信息以及各种传感装置采集的各式各样的分析数据。

3、速度(Velocity)

速度特性体现在大数据在被创建、复制、移动及删除的表现反应。面对海量的、多结构的数据，大数据本身必须具有速度性作为保障，从而降低数据的管理，提高用户的操作体验，降低设备的消耗成本，以高速的效率进行深度的数据挖掘。

4、价值（Value）

价值性决定了海量的多结构数据的意义，从大量不相关的数据中检索弥足珍贵的信息是大数据管理的目的。可以理解为沙里淘金大海捞针，有价值信息的存在是海量数据聚集的目的，价值性决定了大数据具备作为信息预测的可能。

二、档案大数据

档案大数据是在大数据背景下衍生出的一个概念，一般可以理解为数据符合档案特别是电子档案的的特点，如信息的非人工识读性、信息存储的高密度性、信息与载体之间的可分离性、多种信息媒体的继承性，同时又具备大数据具有的海量、多样、高速、价值的4个特点，符合这些特征的档案数据既可以理解为档案大数据。档案大数据存在的意义也同时符合档案利用体系、资源体系、安全体系建设的要求，是档案行业发展到一定程度，特别是在大数据技术日益成熟的当下发展产生的。

三、大数据时代

此概念最早由全球知名咨询公司麦肯锡提及，它指出“数据已经渗透到当今每一个行业和业务领域，是必备的生产要素之一，人们对于海量数据的挖掘和运用预示着新一波生产率增长和消费者盈余浪潮的到来”。普遍的理解为大数据理念及技术在各行各业得到了普遍的应用，并且成为生产要素的主要构成部分，数据管理及应用具备这样特征的时期可以称作大数据时代。

四、Hadoop 框架

Hadoop 是大数据开发的基础框架，能够满足大数据管理特别是档案大数据管理的需要。Hadoop 是由 Apache 基金会开发的分布式系统基础架构及分布式计算平台。可以让用户在不了解分布式底层技术细节的情况下，开发分布式程序，充分发挥空闲主机进行及存储对大数据进行管理与运算。普通用户可以轻松地在Hadoop 上开发和运行处理海量数据的应用程序，Hadoop 在大数据接收、存储、加载、计算方面具有突出的优势。Hadoop 框架的核心是 HDFS 和 MapReduce。

HDFS 为海量数据提供了存储，MapReduce 为海量数据提供了计算。

1、HDFS

HDFS 具备高度容错性，适合部署在廉价的设备上。HDFS 具有很高的数据访问吞吐量，可以实现流式读取文件，非常适合在大规模数据集上应用，完全适合超大数据集的应用程序。

2、MapReduce

MapReduce 是一种编程模型，用于大规模数据集的并行运算，要求数量级达到 1TB 以上。它借鉴了函数式编程语言和矢量编程语言的特性，可以让实现分布式并行计算的主体不用关心它的实现细节。MapReduce 技术实现了把单个任务细化，并将细化的任务映射(Map)到多个节点上，之后再以单个数据集的形式规约(Reduce)到具备 HDFS 特征的数据仓库里。

五、数据仓库

数据仓库是为企业所有级别的决策制定过程提供支持的所有类型数据的战略集合。数据仓库是在数据库已经大量存在的情况下，为了进一步挖掘数据资源及决策需要而产生。因此它具有以下特点：

1、面向主题

操作型数据库的数据组织面向事务处理任务，各个业务系统独自分离，但数据仓库中的数据可以实现按照一定主题区域进行有效组织。

2、效率高容量大

可以实现自定义的数据分析周期，因此具有效率高的特点，并且可以支持海量数据的管理与应用。

3、集成性强

数据仓库具有集成性强的特点，其数据来自分散的数据库，它可以将分散的无关联的大量数据库进行高度集成。

4、扩展性好

数据仓库的设计可以考虑到未来的一段时期，因为它具有很好的课扩展性，能够节省一次性投入的成本，并且保证扩展后的系统具有很好的稳定性。

六、数据挖掘

数据挖掘又称作资料探勘，数据挖掘一般是指从海量的数据中通过有效的算法搜索藏匿于其中信息的计算过程。数据挖掘最早应用于计算机科学，后来逐渐拓展到其他领域，挖掘过程通常通过信息统计、数据在线分析及处理、有效情报检索、机器学习、专家系统和模式识别等诸多方法来实现整个计算过程。

七、智慧档案

智慧型档案的概念来自于智慧档案馆，它们所聚焦的都是数据的智能分析和主动推送，同时是大数据预测的一种表现，是档案大数据在深度分析挖掘后的科学结论。智慧档案是在数字档案基础上发展起来的，是数字档案未来的发展方向，要想达到智慧档案的标准，那就必须将大数据的概念引入到数字档案的管理方式上来，并且将智能化数据分析引入到与数字档案相关的网络建设、资源管理、建筑设施、服务创新等方面，加强大数据与智慧档案的融合，同时要注意两方面的建设。

1、树立智慧档案的理念

围绕大数据的应用将激发前所未有的档案服务理念。社交网络的流行和物联网的建设使得公共部门对个体和群体的实时观察了解正在逐渐密切，有了这种理念的准备，将为预测利用档案的偏好趋势提供了强有力的工具。未来智慧城市的管理，智慧档案馆的建设都应该基于智慧档案的理念，智慧档案的开放与挖掘将极大地提升社会的公开透明度和提高政策制定的效率。一方面，多种类型档案数据的公开大大提升了政府办公决策的公开程度，同时通过公众的监督推进社会的民主化进程。另一方面，通过为大众提供创新平台，充分汲取群体的智慧，有效榨取档案数据的可利用价值，才能提升社会意识和政府效率。智慧档案理念的树立具有紧迫性、普遍性，跟不上时代的步伐就将被时代所抛弃，达不到普遍的认同就不具有生命力，智慧档案的理念是使得全社会发展进步，使得全行业革新进取。

2、加强智慧档案的决策能力

随着大数据时代的来临和深化，档案资政决策行为将日益凸显它的优势，让档案具有智慧是大数据预测的一种表现。伴随着海量数据的有序累积和数据处理能力的不断提升，利用档案数据来进行判断和预测的作用将会得到无限的放大，

大量的数字档案信息将为社会的前进提供助力，将为决策提供依据。海量数字档案信息的出现对档案馆产生了冲击，使它的收集管理对象发生了改变，进而管理方式也将服务与决策的需要而调整，维护手段也应与时俱进的丰富与创新。档案决策功能背景下尽管会出现许多新问题，给档案系统带来一定程度的挑战，但大数据时代已经迫近，档案行业自身应该根据实际情况，科学合理有序地应对这些挑战。明确档案管理的目的，明确数据管理，信息挖掘的规范，解决好这些关键环节，整个档案行业才能在大数据时代将服务决策功能发挥的淋漓尽致，才能适应未来智慧档案发展的需要，才能在面对新观念冲击、新环境影响、新技术革新、新资源变化的大背景下，让档案系统变被动为主动，变落后为进步，高度发达的智慧档案的决策能必将引领行业在大数据时代的发展。

第二部分：档案大数据信息挖掘的意义与作用

一、有利于提高档案信息知识转化能力

档案馆是人类社会的记忆中心，是智慧的结晶 ，是档案数据汇聚和传播的重要场所。帮助利用者运用前人经验解决新问题，必须依靠档案数据所提供的知识服务。大数据与云计算的应用可以提供一种基于“数据仓库”的，能够对档案数据进行搜寻、组织、分析、重组的数据利用方式 ，形成实用性的信息产品，有效支持档案利用者知识创新，并将知识转化为生产力服务。同时通过对特定数据内容的加工、分析、挖掘，形成专业的知识产品，为政府、企业、社会团体的决策提供智力支持和科研信息素材。

二、有利于提升档案信息知识服务水平

档案馆个性化服务是基于信息用户的信息使用行为、习惯、偏好、特点及用户特定的需求，向用户提供满足其个性化需求的信息内容和系统功能的一种服务。首先，它应该是一种能够满足数字档案馆用户的个体信息需求的服务，即根据用户提出的明确要求提供信息服务，或通过对用户个性使用习惯的分析而主动地向用户提供其可能需要的信息服务；其次，应该是一种培养个性、引导需求的服务，这样可以帮助个体培养个性、发现个性、引导需求，促进社会的多样性和多元化发展。通过统计分析档案利用信息（例如利用登记信息、反馈建议、利用行为等），可以确定档案信息或服务受欢迎的程度；分析档案用户的类型和个人爱好，发现用户访问模式和用户需求的趋势，从而发掘潜在档案用户，提供档案信息推送服务。而对用户进行的跟踪分析，则从不同侧面来研究用户的信息需求及其行为规律，从而为进一步优化数字档案馆的档案信息资源建设，改进数字档案馆的资源组织和管理模式，以及改善数字档案馆服务方式，提供科学的依据。

三、有利于推动智慧档案馆信息资源建设

在智慧档案馆的资源建设方面，档案大数据信息挖掘能够发现馆藏资源的缺漏，有针对性地补充和丰富档案信息资源或其他文献资料；分析档案资源的利用率，预先发现用户群体兴趣的变迁；通过对用户每次利用的档案信息进行关联分析，挖掘各类档案信息之间的关联规则和比例关系，进一步优化馆藏；利用路径分析模式挖掘捕捉用户频繁浏览访问的路径，改进数字档案馆站点结构设计。

除此之外，建设具有主题鲜明、超大容量、稳定安全的“数据仓库”也是大数据档案信息挖掘的重要组成部分。档案工作者可以运用关联、分类、聚类等方法，打破原有的整理体系，从海量档案信息中按照相关专题进行挖掘、分类、加工、整理和有序化重组，构建特色档案信息库及各类专题档案信息库。为弥补现有档案数据库信息量的不足的缺点，档案管理者可利用 Web 挖掘方法从海量网络信息资源中检索出与专题相关的文献信息资料，分类、整合后形成知识性、系统性的二次、三次数字文献信息资源，并建立信息资源主题指南。以上办法都可以为数字档案馆向智慧型档案馆转变起到强力的助推作用。

四、有利于推动大数据在档案领域的应用

课题理论研究成果可为全国综合档案馆大数据应用提供理论与现实参考，促

进档案馆利用新理念和新技术提高档案信息化管理水平，促进档案信息化发展进度。研究实践成果将用于辽宁档案工作实际，推动大数据在档案建设中的应用。

第三部分：档案大数据发展环境与背景

当前，大数据成为国内外的研究热点，研究成果和成功案例层出不穷，国内的互联网公司都试图获取和整合更多的用户行为数据，把碎片化的数据用种种手段整合起来并加以利用。此外由于数据服务本身对数据收集、存储、分析和加工等方面的需求，一些硬件层面的变革也在产生，国际IT巨头如IBM、EMC、惠普等都开始在这些领域推出了针对性的产品。国内IT企业也开始在原有的业务和产品基础上加大数据领域的研发和投入。在这样的时代背景下，档案界也悄然揭开了大数据研究与应用的序幕。与其他以盈利和价值为导向的行业如零售业、体育业、金融业相比，国内外档案界对大数据的相关研究还处于概念引入与内涵拓展阶段。

一、国内发展现状

工信部发布的物联网“十二五”规划中，信息处理技术作为4项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。而另外3项信息感知技术、信息传输技术、信息安全技术，也与“大数据”密切相关。国内大数据市场有市场需求广阔、后续增长潜力大、投资前景好等优点，后续发展值得看好。但庞大的人群和应用市场，复杂性高、充满变化的行业条件，以及在政策、理念和历史因素等方面的约束，使得我国成为世界上最复杂的大数据国家。

1、首个大数据研究服务基地揭牌

2014年6月28日，中国统计信息服务中心与厦门市信息化局共建的首个“大数据研究服务基地”在厦门揭牌，同时，“大数据研究实验室”也落户厦门市一家科技公司。中国统计信息服务中心与厦门市信息化局共同携手，将以“大数据研究服务基地”和“大数据研究实验室”为依托，搭建大数据产业发展平台，整合产学研各方资源，加快厦门乃至我国大数据产业的发展。目前，厦门市信息化局正加紧研究大数据产业发展行动规划，以推进大数据应用、发展大数据产业、保障大数据安全。厦门市还将建设涵盖市民、法人、地理空间的全市大数据资源库，同时强化政务信息资源整合与共享，并在此基础上发展智慧交通、智慧医疗，让市民共享大数据成果。

2、信息技术公司是发展大数据应用技术的主力

在大数据时代爆发的背景下，无数互联网企业开始重新审视自己的行业定位，将数据资源和数据价值提升到自己的核心战略中，并衍生出一系列新型服务和产品，而这种趋势还在继续迅速爆发中。例如，2011 年凡客成立了数据中心，希望实现互联网的系统化和数字化的管理，包括库存、进货周期、周转、订单等方面的数据研究，以及研究新产品的上架与新用户增长的关系，每上线一个新品与它能够带来的用户二次购买的关系等；新浪微博则成立数据部，在海量微博用户和信息的基础上开展了一系列数据挖掘和分析的研究和实验，同时开放数据给第三方创业公司，初步形成了依托其上的一个小型社交媒体数据分析挖掘生态；携程网则将自己多年来在 OTA 服务领域积累的数据和用户习惯深度挖掘出来，并在此基础上形成了一套新型服务模式和服务体系；今年四月份百度则推出了一款特点鲜明的大数据产品“百度预测”，目前包括了世界杯预测、城市预测、景点预测、高考预测、疾病预测，陆续还将推出房地产预测，票房预测，就业预测，金融预测等等。

3、大数据在公共服务领域的实践

大数据技术已经在一些地方政府主导的“智慧城市”项目中有了实施，以“平安城市”视频监控应用为例，该项目是通过构建一个覆盖整个城市的集成化、多功能、综合性治安防控网络，帮助公安部门更高效、更精确的控制和打击犯罪，从而保证城市环境的和谐与稳定发展。其视频接入规模从成千上万到十几万、甚至几十万都有，其中涉及了治安监控、指挥通信、侦查破案、规范执法、社会服务等多个分区领域，视频质量也从标清进步到了高清时代，所产生的数据无论在规模还是结构上，都符合大数据的定义。对监控大数据的智能处理能从大量非结构化的视频数据中提取出有价值的信息，从过去和目前以事后查看为主，慢慢转变为事前预警，这就可以为公安、交通等各行各业提供更为有效的业务信息支持。”

4、图书档案界现状

为考察国内图书档案界对大数据技术的关注情况，笔者以“大数据”为关键词，在 CNKI 全国学术期刊全文数据库档案学分类下进行检索，2012 年之前没有相关文章，2012 年 1 篇，2013 年便达到 25 篇之多。可见档案人对新技术也保

持着极高的关注度，也从侧面反映了大数据技术与档案的高度相关性。

同样，数据挖掘与分析也是档案大数据处理的核心所在。与传统数据挖掘相比，大数据背景下要求从数量更为巨大、结构繁多的档案信息数据中挖掘出隐藏在背后的规律，揭示数据的真实价值，发挥数据的最大化价值。如何充分有效地利用知识挖掘方法对档案信息资源进行开发分析，是大数据时代知识服务深入发展的重要研究方向，也是各类档案机构服务创新的关键途径与支撑服务技术。

国内数字档案馆也已开始重视以用户为导向的档案信息服务，基于分析平台对档案信息资源进行深度、动态、广泛的知识挖掘，揭示档案信息交流的各个要素以及其相互之间的联系，促进档案信息的横向交流以实现信息资源共享，满足用户网络交流需求以提高档案利用效率，从而实现档案信息增值服务和提高档案信息服务的竞争力。如福建省的“网上公共档案馆”项目，共有64个公共档案专题全文数据库，内容涉及闽台关系、历史文化、经济生活、人文景观、档案珍品、民生热点问题以及与重大时事和重要节庆有关的档案专题。福建网上公共档案馆试点单位还建立了一个反映当地爱国主义教育特色的爱国主义宣传档案全文专题数据库，充分挖掘馆藏档案信息资源，最大程度发挥利用价值，实现档案信息资源社会共享。

此外还有青岛档案馆的网络社区、上海档案信息网的“档案博客”、南昌市档案局的微博等都是档案信息知识服务的亮点，同时也收到了良好的社会效果。

二、国外发展现状

社会进步和信息技术发展不断改变着图书情报和档案界的工作方式，随着档案信息化的快速发展，互联网上各类档案信息源，如档案网站、论坛和博客等提供的大数据信息，正成为档案开发和利用的新型资源。而移动网络和社交媒体技术的发展使得将这些信息资源组合成整体并延伸至更大的范围的集成服务成为可能。美、英等国对网络环境下档案资源模式的研究比较成熟，运用高科技手段是其提高信息整合效果的关键。如英国国家档案馆网站设计的“新千年档案展览”，直观展示每一世纪的第一天；澳大利亚国家档案馆在线网展的“Faces of Australia”（澳洲面孔）、“Pic of Week”（每周图片）等栏目；日本共享档案馆通过协议合作在一个检索工具上实现资源的最大限度利用；美国国家档案与文件署则加入了奥巴马政府的“大数据”国家战略，联合其他高校与研究机构成立相关

项目组，开展大数据应用研究。

1、英国——大数据的积极拥趸者

早在2011年11月，英国政府就发布了对公开数据进行研究的战略政策，英国内阁部长弗朗西斯·莫德说，其实英国政府早有意带头建立“英国数据银行”，政府想算清楚究竟这个国家或政府创造了什么；英国不只是要成为世界首个完全公布政府数据的国家，还应该成为一个国际榜样，去探索那些公开数据在商业创新和刺激经济增长方面的潜力。可以说，英国是大数据的积极拥抱者。无论是政府、研究机构，还是企业，都已经开始行动，抢占“数据革命”先机。

2013年1月，英国注资6亿英镑（约9.12亿美元）发展8类高新技术，其中，1.89亿英镑用来发展大数据技术。英国政府专门建立了“数据英国”（data.gov.uk）网站，将公众关心的政府开支、财务报告等数据整理汇总并发布在互联网上，并对其中的热点议题和重要开支进行进一步阐释，并对公众意见进行反馈。

除却简单的开发政府数据之外，英国政府于2012年5月注资十万英镑，支持建立了世界上首个开放式数据研究所ODI(The Open Data Institute)。ODI是非营利性组织，它将把人们感兴趣的所有数据融会贯通在一起，每个行业的各个领域一面产生各种数据而另一方面又可以来利用这些数据。英国政府通过利用和挖掘公开数据的商业潜力，为英国公共部门、学术机构等方面的创新发展提供“孵化环境”，同时为国家可持续发展政策提供进一步的帮助。

据报告显示，英国政府通过高效使用公共大数据技术每年可节省约330亿英镑；通过数据使用，优化政府部门的日常运行和刺激公共机构的生产力，可以为英国政府节省130亿至220亿英镑；减少福利系统中的诈骗行为和错误数量将为英国政府节省10亿至30亿英镑；有效地追收逃税漏税将为英国政府节省20亿至80亿英镑。

英国国家档案馆由英国公共档案馆和皇家历史手稿委员会合并而成，是世界上最大的档案馆之一，已有一千多年的历史。不仅保管政府部门的文件，而且向公众提供利用可公开的政府文件，同时也提供与英国历史有关的私人档案信息。英国国家档案馆非常重视网络与信息化建设，随着信息技术的发展不断调整管理策略和发展方向。早在1998年，英国国家档案委员会在一份名为《英国档案馆

的发展道路》的报告中，就提出了建立英联合王国网络档案馆的计划，并明确给出了15条具体建议。该计划旨在采取联机目录方式进行档案信息一体化网络的组织与实现，即国家档案馆作为全国档案联机目录中心，根据标准的著录规则和数据交换格式，对入网档案馆和文献信息机构提供的档案信息进行统一编目，通过互联网将档案目录数据进行实时传送和交换，形成逻辑上的目录库，并按地区、类型或载体对这些目录进行组织，供网上用户查询使用。

2006年10月，英国国家档案馆发布了新的发展规划——《2006-2011英国国家档案馆新视野》，提出三个方面的发展目标：领导和变革信息管理；为未来保存今天的信息；将历史引入每个人的生活。英国国家档案馆认为在线服务是实现这些目标的重要措施。为此，2008年7月1日，英国国家档案馆专门发布一个新的在线策略——《提供（信息）和赋予能力：英国国家档案馆的在线（网站）策略》，展示了在未来三年英国国家档案馆的在线（网站）策略，主要包括如何利用网站实现国家档案馆的发展目标，如何适应不断变化的网络环境，如何继续提供更优质的信息服务等。该策略对影响在线服务的一系列变化作出回应，提出要充分利用新技术推动在线服务变革，服务变革的核心理念是“基于信息技术的服务方法必须围绕公众设计”，主要任务是“通过在线服务，使公众了解并接触政府文件、法案以及官方信息，了解从国家档案馆能够获得和利用哪些信息；通过在线服务，为政府和公共部门的档案工作提供指导”等。

这一系列政策的实施，为当今大数据时代背景下档案信息资源的挖掘与传播打下了良好的基础。英国国家档案馆多年来坚持不懈地建设与完善目录数据库建设，使英国国家档案馆网站具有了超强的信息检索能力，可以便捷地提供多种途径的档案信息检索服务，用户根据个人需要，或系统浏览档案目录，或按主题、时间、档号等特征进行目录查询。例如，国家数字档案数据集（NADA）提供英国政府部门形成的数字化档案，这批档案可以满足关注政府部门历史、计算机应用历史、信息技术对社会的影响等方面用户的需求。PROCAT联机目录（online catalogue）就是一个覆盖950万卷档案目录信息的多级式目录数据库，提供了多种本地和远程检索途径，如自由检索、引导检索、档号检索、简单查询、熟练查询、高级查询等，并给出档案的详细出处。国家档案目录信息网由英国国家档案馆参与的“利用档案”项目（the Access to Archives, A2A）及其姐妹项目共同构

成。A2A 可提供英国国家档案馆以外的英格兰各地各类文献部门保存的自公元 900 年至今 1100 多年间的主要档案信息资源的目录数据，方便了社会各界对这些文献遗产的利用。

2、美国——大数据的策源地和创新引领者

从计算机革命开始以来，美国拥有一大批的领军企业，包括谷歌、微软、EMC 这样的业界巨头，也有 Facebook、Splunk、Teradata 这些后起之秀，这是任何其他国家短期都无法复制和匹敌的巨大力量。

美国企业也拥有对于数据重视和应用的历史传统，IT 基础设施的完善，以及各种精准营销理论和实践也都是走在世界前列，比如基于消费数据、信用卡数据挖掘的精准营销等，还有电话、DM 印刷品和邮件营销在美国都很兴盛，随着互联网兴起，谷歌、IBM、YAHOO 等美国企业对基于网络的精准营销又是走在全球的前列。因而大数据最典型案例中，就包括传统企业沃尔玛“啤酒+尿布”案例，以及谷歌公司通过大数据分析成功地预测流感爆发等。

更加重要的是美国政府数据开放和支持力量。美国政府的数据开放一直是走在全球前列的， 2012 年 5 月美国数字政府战略发布，更是提出要通过协调化的方式，以信息和客户为中心，改变联邦政府工作方式，为美国民众提供更优公共服务。

在开放数据、创新驱动以及技术研发支持下，美国大数据的研究和应用已是走在全球前列。2013 年 5 月，奥巴马政府宣布了“大数据的研究和发展计划”，提出“通过提高我们从大型复杂的数字数据集中提取知识和观点的能力，承诺帮助加快在科学与工程中的步伐，加强国家安全，并改变教学研究”。在斯坦福这样的大学里也开始开设诸如机器学习这样全新的课程，培养下一代的“数据科学家”。伯克利加州大学、迪肯大学等大学也专门开设了研究大数据的相关课程。如今，美国不仅是全球首个将大数据从商业行为上升到国家意志和国家战略的国家，也是数据科学家和面向未来的大数据人才储备启动最早的国家。

美国国家档案馆和文档管理署(The National Archives and Records Administration，以下简称 NARA)既是接收、记录美国联邦政府重要文件的官方机构，也负责保管大量的国家历史档案资源. NARA 有数量巨大、检索功能齐全的数据库资源，通过美国国家档案馆网站可以充分利用美国国家档案馆馆藏超过

5000万份的历史资料，为了方便用户利用，网站开发了一系列的网络数据库，如检索非电子文件的档案研究目录系统（Archives Research Catalog，ARC）、检索电子文件的档案数据库检索系统 （Access to Archives Database，AAD）、国家档案馆图书馆目录 （NARA Library Catalog ）、检索缩微资料的缩微出版物检索系统（Microfilm Publications Search）以及肯尼迪总统暗杀记录收藏参考系统（The president John F.Kennedy Assassination Records Collection Reference System），丰富的数据库资源极大的方便了用户的检索和利用。

为响应奥巴马政府的“大数据的研究和发展计划”国家战略，NARA与美国国家科学基金会NSF (National Science Foundation)、北卡罗莱纳大学教堂山分校联合启动了“十亿电子文件信息架构”项目(CyberInfrastructure of Billions of Electronic Records，CI-BER)，为数十亿联邦政府电子文件建立母版，并实现不同方式的可视化呈现等，后来又加入了杜克大学、阿什维尔大学、阿什维尔市等新的合作伙伴,形成了一个分别代表计算机科学,政治学,人文科学、工程学、信息和图书馆学等领域的合作团队。

“想象一下你要如何在40TB，大概七千万件档案中查找与一个特定地理位置相关的电子文件？现在你只需要一个平板电脑，点击地图上你感兴趣的地点，就会在旁边出现一张相关档案文件的列表，当你点进列表中，甚至能看到每一条文件的元数据”，这是CI-BER项目组在2011年大数据分析与可视化专题讨论会上演示的阶段性成果，目前已实现的工具集包括：

- 大数据集的检索
- 在大量记录中识别出包含特定地理位置信息的文件
- 定位能够打开这些文件的应用软件
- 打开文件
- 从文件中抽取元数据
- 确定文件有关的地理范围
- 为索引附加文件元数据和所涉及地理位置的经纬度信息

这些工具都是针对NARA馆藏的联邦政府电子文件量身定做的，随着研究的不断深入，其功能将愈加完善。

在利用大数据分析技术深入挖掘信息资源的同时,NARA对信息的宣传与传

播也非常重视，不遗余力的扩大档案资源的影响力。NARA 在很早之前就开始了对“新媒体与档案管理”这一课题的研究。他们将“新媒体”定义为：以 Web 2.0 和社交媒体等网络新技术为支撑的信息交流平台，其中牵涉到社会参与和内容共享等一系列活动，政府机构和组织能够通过这一平台与广大民众紧密联系在一起。在通常情况下，这一新技术平台由非政府的第三方组织（网络服务公司）运作，以其异常高效灵活的特点，日渐融入到人们的日常生活。档案管理机关如要跟上时代前进的脚步，对“新媒体平台”的利用将非常重要。

新媒体以沟通互动为基础，也常被称为社交媒体。Nara 将其进一步细分为三类：一是如微博、博客、维基网站那样鼓励创作并发布原创内容的网络空间；一是社交网络工具，如 Facebook、LinkedIn 等；还有一种是网上文件存储与共享空间，如 Flickr、Picasa 等。NARA 于 2010 年 12 月制定了一份详尽的社交媒体战略规划书，这一战略目有六大核心理念：合作、领导、发起、多元、聚合、开放。并进一步细分出三大目标服务群体：内部员工、政府部门及社会公众。关于内部员工，NARA 相信新媒体技术可以帮助雇员们更有效率和活力地完成工作，网上信息共享与协作可以激发个人潜力，为解决问题提供帮助；对于政府部门，NARA 希望通过新媒体，将不同政府部门的档案管理者、从业者们联合起来，提高政府档案管理的效率,并为新媒体平台上所产生的大量数字信息的记录保存寻找最佳解决方案；而服务社会公众则是新媒体平台最重要也最根本的目标，同时也是呼吁公众为档案历史挖掘、档案文化传播贡献更多力量。

3、澳大利亚——利用大数据推动公共服务建设

2013 年 8 月，澳大利亚政府信息管理办公室（AGIMO）发布了公共服务大数据战略。该战略将以“数据属于国有资产，从设计着手保护隐私，数据完整性与程序透明度，技巧、资源共享，与业界和学界合作，强化开放数据”等六条大数据原则为支撑，旨在推动公共行业利用大数据分析进行服务改革，制定更好的公共政策，保护公民隐私，使澳大利亚在该领域跻身全球领先水平。预计这六条原则将极大地提高生产力及创新收益，并协助政府解决各种难题。

为将六大原则落实到实处，该战略将还制定了一个具体的行动安排：2014 年 3 月推出大数据实践指南，2014 年 7 月前出台一份关于大数据分析中所遇难题的报告；然后推动 ICT 行业和教育行业提供大数据分析中的必要技巧，制定

一份数据分析指南和两份在建项目指南；开发一个信息资产登记系统；记录大数据分析中的技术演进。

澳大利亚公共服务大数据战略最早由政府ICT监管委员会于2013年6月提出方案，澳大利亚信息管理办公室(AGIMO)八月份发布的正式版本又咨询了相关机构和业内人士意见

澳大利亚联邦政府首席信息官Glenn Arche表示，“政府希望通过大数据分析系统提升公共服务质量，增加服务种类，并为公共服务提供更好的政策指导。澳大利亚政府希望，在大数据分析的运用、提高效率、与其他政策和技术协同以及为公共服务领域带来变革等方面，澳大利亚能领先全球其他国家。”

澳大利亚政府还积极应用开放数据的理念和行动践行开放政府的愿景和目标。Data.gov.au是政府信息目录的开放数据平台，用户可以在该网站上简便地搜索、浏览和利用澳政府国家、地区政府的公共数据，政府鼓励所有用户通过更新工具和应用从信息中得到实惠。该网站包括114个部门的1103个数据库和18个应用软件，分为首页、数据、目录、应用软件、资源、更好的实践、建议和关于网站八项主题。网站上的数据来自澳政府多个部门，提供数据下载，并提供其他数据目录或资源的链接。澳大利亚政府数据开放通过5个阶段将数据开放流程化，这5个阶段依次是：发现数据（Discover）、过程处理（Process）、授权许可（License）、数据发布（Publish）、数据完善（Refine）。

在数据中心建设方面，2013年8月，隶属于澳大利亚财政与解除管制部门的ICT采购部发布了《数据中心结构最佳实践指南》草案，供公众审议。该《指南》旨在为澳大利亚政府机构提供优化数据中心结构相关运营活动的建议，是澳大利亚政府数据中心战略2010-2025的一部分，目的是在将来为数据中心节省10亿美元的成本。据分析公司Gartner的数据显示，2012年，澳大利亚数据中心的总数达到49577家，达到巅峰阶段，随后会缓慢下降，到2015年将降至45545家。为了满足业务需求，许多大型企业和研究机构OCLC、微软、甲骨文、Rackspace等都在澳大利亚投资建设了数据中心。

公共服务的理念渗透至澳大利亚的各个行业，档案行业自然也不例外。“你的故事，我们的历史”（Your story, our history）是澳大利亚国家档案馆的服务宗旨和宣传口号，档案馆不是高高在上的政务机关，而是与澳大利亚民众生活息息

相关的记忆宝库。

家谱档案是民众最为关注的档案门类之一，澳大利亚档案馆对馆藏资源进行分类和深入挖掘，整理出专门的家谱档案，并在官方网站中开辟了“家族历史”(Family history)专题版块，帮助澳大利亚民众查找祖先，了解家族历史。当然，这些档案不能涉及所有澳洲公民，主要为曾在澳大利亚军队服役或20世纪以来的移民档案。

检索家族历史的检索选项也非常详细，包括家庭成员是否曾在政府任职;是否申请过护照、是否领取过养老金或其他福利、是否曾申请来澳探亲、是否注册过专利或商标、是否为澳正式公民、是否接受过政府拨款或奖学金、是否行使过选举权等等。

澳大利亚档案馆还设立了案例研究专栏。在这个栏目中，展示了5位澳大利亚知名人士或普通民众的家族历史，其中就有一个中国人在澳大利亚的家族历史。资料的详尽程度令人惊讶，不仅有用文字梳理出的家族成员和发展脉络，还有珍贵的照片和实物档案。这些档案均已完成著录，只要点击图片或链接，就可以看到详细的信息，如：题名、内容、时间、范围、系列号、标签、档案条形码、储存地、是否可获取、载体格式等。人们还可以通过点击查看电子版复印件(View digital copy)，看到清晰的档案材料。这些历史全都是通过澳大利亚档案馆收藏的档案中研究得来，充分展示了挖掘档案潜在信息与利用价值的重要性。

第四部分：档案大数据信息挖掘模式研究

一、确定档案大数据挖掘目标

档案包含着大量关于自然界与人类社会各种事物的存在方式和运动状态信息，具有历史价值和使用价值。档案工作是社会主义各项建设事业不可缺少的一个重要环节，是一项重要的综合性管理工作。档案大数据的挖掘必须紧紧围绕中心工作进行，根据工作计划提出数据挖掘方案，将各种有价值或潜在价值的档案数据收集起来， 为什么挖掘数据资源，挖掘哪一部分数据资源是开展此项工作首先需要思考的问题。目前档案工作发展状况来看，数据挖掘主要分为以数据存储为目的，和以服务利用为目的两种。

1、以存储为目的数据挖掘

档案是一种社会信息的传承，从传统意义上的纸质实体档案，到建议实行“双轨制”保存的电子档案，再到互联网上即时更新的各种“孤本式”网络信息

资源，它是以一定的形式存在，并且其中有一部分要永久保存下去为子孙后代造福。过去，档案资源的积累主要靠移交接收单一渠道，而且这一渠道常常是封闭的、被动的，具有很大的局限性。档案检索则主要依靠手工著录、卡片检索。

在大数据时代，各单位在日常工作中产生的数据和信息呈爆炸性增长，最终作为档案保存下来的文件也相应的增长。数据时代的到来为档案资源建设带来了新机遇、新发展，资源积累的渠道是多种多样，如货币征购、无偿捐赠、协议寄存、复制收集等，这些渠道都是全开放的，不受地域、时间的限制，为档案资源的积累产生了极大的效益。同时，随着档案信息化的快速发展，互联网上各类档案信息源，如政府官方网站、新闻门户网站、论坛和博客等社交媒体网站等提供的大数据信息，通过网络采集技术，正成为档案保管、开发和利用的新型资源。另外，信息技术的进步和数据库技术的发展，使档案管理变得更为快捷和方便。

确定以存储为目的的档案数据挖掘范围必须要做好以下几点：

一是要树立“大档案”意识。所谓“大档案”是一种适应时代要求的新理念，就是跳出档案看档案，跳出档案抓档案。档案资源体系建设要有前瞻性、预见性，要有统一规划和要求，不能零打碎敲、盲目行事。要将档案资源体系打造成为便民服务的体系，最大限度满足社会需要，使人民群众共享建设成果。

二是要建设数字档案馆管理体系。档案资源数字化必然产生相应的数字管理系统，建设数字档案馆是实现档案信息管理现代化的必由之路，而数字档案馆的建设必须在统一标准规范的指导下进行。

三是要实现档案信息资源共享。现代档案资源体系建设的聚焦点和最终目标应该是全方位、全覆盖的共享体系，应当打破档案资源信息“孤岛”，将众多档案资源由“单体”实现“统筹”，形成“互联互通”的大档案数据仓库。

2、以利用为目的数据挖掘

档案的保存和管理是为了利用，我们保存档案的目的，是为社会主义事业服务，提供档案给各方面使用，充分发挥档案的作用。档案提供利用工作是档案工作诸环节中最富有活力的一个环节，是档案工作联系群众，服务群众的纽带。一方面，通过提供利用工作把馆藏的大量档案材料提供给利用者，满足多方面的需要，充分发挥档案的作用；另一方面，是对档案工作最实际有效的宣传，能扩大档案工作在社会上的影响，争取各方面的重视与支持。在大数据时代，用户对档案资源的需求已不仅仅局限于原始的档案，需求变得更加个性化、多样化。档案提供利用工作的基本内容，就是熟悉“数据仓库”中所存档案的内容和成份，了解客观需要，通过各种方式迅速、准确地将有关档案提供给党和国家各项工作使用。这就需要通过对档案信息资源进行二次开发，制作多种形式的编研产品，并主动提供给用户或推送给潜在用户。

档案信息二次开发是指利用各种计算机技术、多媒体技术、通信技术等对档案及相关信息进行再次开发，它是档案资源信息化开发的高级内容，其所蕴含的信息量和信息价值也是无比巨大的。大数据时代下，档案数量急剧增加、多种多样档案类型以及非结构化数据的大量存在，给档案信息资源的二次开发带来了机遇与挑战，如何在海量数据中选择有价值的信息并找出它们之间的关联，编研开发非结构化档案数据信息等，都是开发、利用档案资源中必须面对和思考的问题。

因此，档案管理工作的基础是“存储”，实质是“利用”，存而不用等于白存，用而不存等于没用。二者相辅相成，缺一不可。大数据时代我们应充分利用先进科技手段，使档案资源最大限度地发挥其特有作用，这是实现档案工作目的

的直接手段，也是档案管理工作自身价值的体现。

二、遵循档案大数据挖掘原则

1、合规性原则

馆藏档案信息很多内容涉及国家、外交、疆界、民族等方面，敏感、绝密、且未解密档案信息，不宜也不能对外公开，在网络上获取新型档案资源信息来源更加广泛，可能涉及信息商业机密信息、个人隐私信息等。与数据仓库建立相似，进行数据挖掘也必须遵守国家、行业以及本地区的相关法律、法规及各种标准规范。保证维护信息资源的知识产权、著作权和信息的保密性。如《中华人民共和国著作权法修正案》、《信息网络传播权保护条例》等。

2、可扩展性原则

数据是无时无刻不在扩展的，特别是网络信息资源，扩展速度超乎想象，所以数据挖掘管理必须保证自身功能的可扩展性以及容量的可扩展性，以满足数据类型的多变性和迅速增长的数据量的要求。同时，档案信息挖掘也是一个庞大而长期的工程，不能一蹴而就，需要系统规划，循序渐进，不断完善，常抓不懈的工作。不但要依靠新技术来推进，更要灵活的将数据挖掘技术与档案学理论动态结合，掌握好工作重心和档案工作的发展趋势，使档案数据挖掘工作始终处于不断完善发展之中，实现此项工作的可持续发展。

3、选择性原则

针对全部馆藏档案信息资源进行数据分析和挖掘是我们的最终目标，但因馆藏信息资源数量巨大，并且每时每刻都在无限扩展，因此需要逐步逐项开展，特别是网络信息资源来源庞杂,混有大量毫无用处的垃圾信息甚至是有害的信息，鱼龙混杂的网络信息不仅加大了保存的成本，也妨碍了归档信息的再利用。所以必须坚持选择性原则。有选择性地获取和挖掘此部分资源不仅可以节省人力、物力和财务,也可避免这些垃圾或有害信息带来的负面影响。有选择性数据挖掘，信息鉴定确定数据整理和挖掘的对象。具体应有以下几点

（1）特色性挖掘

认真分析社会需求，然后针对自己的馆藏特点，形成档案信息资源特色。以辽宁省档案馆为例，在国内各级档案馆馆藏中辽宁省奉系军阀档案、满铁档案档案界较为完整和权威的档案资源，因此是辽宁省档案馆馆藏特色。

（2）针对性挖掘

挖掘档案信息要有针对性，要密切注视和分析社会动态，通过会议、新闻媒体、上级文件等各种形式，把握社会热点，有针对性地开发社会需要的档案信息产品。

（3）服务性挖掘

档案信息开发必须紧紧围绕信息利用者的活动，随时根据档案信息利用者提出的要求，以最快的速度加工处理档案信息，提供高质量的信息产品。

（4）规模化挖掘

档案信息的开发不能停留在小规模、零散的、单项的挖掘上，虽然这种开发能在一定程度上满足小范围利用者的需要。

4、准确性原则

即对数据挖掘质量的控制问题。数字档案馆数据库中涉及大量的数据信息，在这些海量数据面前，不可避免的会出现冗长，甚至错误的数据，所以在进行数据挖掘时，应根据数据挖掘任务的不同，选择适合的挖掘类型和算法，并对出现的错误数据进行修正、处理、加工，为档案馆提供科学合理的各种分析报告和相关预测信息，指导档案馆工作人员采取正确手段，并为档案馆改进服务、做出决策提供智力支持。

三、 建立档案大数据“数据仓库”

大数据时代，档案馆维护和传承记忆的功能将更加重要，构建一个基于互联网的，以档案数字资源为主体，以文本、图片、音频、视频等为形式，为中华民族集体记忆的建构和传承提供文献支撑的“中国记忆”数据仓库，将成为档案人新的目标与使命。

建立数据仓库首先是要搜集数据，数据越丰富越好，数据量越大越好，只有获得足够的数据，才能获得确定的判断，才能产生认知模型，这是量变到质变的过程。档案数据来源和采集范围主要包括：

- 互联网上具有档案价值和参考的信息
- 传统馆藏档案数字化形成的资源库
- 各立档单位的数字化后的档案文件资料
- 具有档案性质的行业、专题信息资源库

采集的途径主要有通过网络在线采集网络上现有的各种信息资源。或者根据社会需求，采购一些全文光盘数据库补充数字档案馆数字资源的不足，如中国科技文献数据库、中国科学文献数据库等。

1、互联网档案数据采集

大数据时代，档案馆的核心竞争力在于其拥有的档案资源。因此，要实现从传统档案资源观向“大档案观”的转变， 尽可能地收集全面数据、完整数据和综合数据，更多地关注一些底层化、碎片化、复杂化的信息，从而构建一幅反映国家和社会变迁的实时全景图。档案馆在进一步推进纸质档案数字化、加快电子文件接收进馆、选择性采集政府网页信息的基础上，应有意识地收集一些诸如电子邮件、社交媒体、即时通信等价值重大、形式多样的数据资源，从而实现档案资源的全方位保存，真正建立覆盖人民群众的、满足长远需要的档案资源体系。

（1）互联网档案大数据采集方式

互联网虽然提供有价值的信息资源，但也存载着大量的信息垃圾。上世纪90年代中期，一些国家相关组织着手尝试互联信息的采集方法,都是利用传统的网页爬虫技术，将目标网站上的含有信息的网络资源文件采集下来，并以某种方式保存起来，经过多年的实践后,目前成熟的方法可归结为三种：选择式、全面式和综合式采集。

① 选择式采集

选择式采集是根据某种标准,对互联网资源进行选择后再采集的方法。丹麦国家图书馆和加拿大国家图书馆采用这种方法。这种方法将互联网资源与纸质资源同等对待,对于以静态网页发布的信息资源作某种范围的选择后采集。目前公认的选择标准是首先判定资源在未来对于研究人员是否具有利用价值,这种方法可以说是传统纸质文献采购方式的某种变通和延续，其优点是：

- 每个选出的条目其资源质量是有保证的，而且在当代技术能力的水平上它可以被最大程度的利用；
- 对每个被选定的主题可以制定单独的采集程序表，统计它的出版日程和频率,尽量完整的集中某个主题下的内容；
- 对采集后的各个条目可以进行完全著录,进而合并至整个采集数据库中；

- 通过与出版者事先的协议，解决版权问题,这样,每个采集后的条目可以立刻被读者通过互联网进行检索利用;
- 可以对归档后的资源进行二次整合，如其重要性和资源级别的分析与确认等;
- 不能由智能程序进行采集的站点资源可以通过其他方法进行收藏,如与出版商达成协议而采用专门方法等。

当然,这种方法的人为性非常明显,缺点主要有在做出选择时,与主题相关的资源在未来对于研究人员是否具有利用价值,选择判断是主观做出的,难免偏颇。相比于全面式归档,选择式归档是极受限制的,选择必然会漏选某些有价值的资源。其次,选择式归档属于劳动密集型,单位条目的成本相对较高,如果收藏范围逐渐扩大,那么劳动力的增加会呈无限之势,显然不符合网络时代的工作要求。再次,从内容本身而言,选择式方法通常只捕捉到上下文关系连接紧密的资源,而对各种相关的资源却难以准确抓取。

但从另一角度看，网络页面和内容的爆发式增长，尤其是微博、微信等网络形式和多媒体内容的广泛出现，互联网站的数量、内容和信息量的增长已使互联网档案资料的保存面临很大困扰。

②全面式采集

全面式归档是尽可能将所有互联网资源进行采集的方法。这是网络资源归档的理想模式,它试图运用抓取智能程序自动归档所有互联网资源,而极少掺杂人工干预。在理论上,全面式采集突出的是将所有的资源在某个周期内以最少的人工干预归档,而且能够很好的将网络资源依附的网站框架和组织结构保存,对管理人员而言,能够保障信息资源的原始性，对利用人员而言，不仅能够利用上下文关联的资源,而且能够检索到其它的相关资源。

但全面式采集要求计算机全天候的运行以及巨大的存储空间，基础设施费用高。另外虽然人工干预归档量较少，但对系统可靠性有很高的要求，必要时还需要工作人员24小时监控。澳大利亚国家图书馆实践后发现,全面式采集中近40%的抓取是残缺的或是有瑕疵的，突出的有价值资料往往被忽视,而这些问题在归档管理时是不易觉察的。要解决这个问题,需要对抓取软件的智能化改进和质量

检查软件可信赖功能的增强。

③综合式采集

综合式归档是前两种采集方式综合运用。美国国会图书馆采用了综合式归档方法,该馆联合其他合作伙伴,如互联网归档局,建立主题归档联系,即对商定的主题范畴内的资源进行全面归档, 如2002年选举、“9·11”事件这样的主题。

分析各个采集方式，目前都存在一定的缺陷，如：选择式方法疏失了也许在将来很有价值的资料;全面式方法则过于宽泛等。目前最为理想的模式是全面评估、综合式方法再辅以某种程度的控制, 根据产生网络信息资源的网站类型，其采用的硬件设备、技术手段、网站代码、呈现方式等等，分析不同网站的信息的特点、类型、复杂程度、以及主权单位等，从多方面因素分析从而制定采集策略，对具有长期研究价值的资源进行全面采集,再辅之以与发布单位签订的采集协议。

（2）互联网档案大数据的采集目标

① 基于既定目标的数据采集

大数据的最终价值在于利用。随着信息环境的变化和社会档案意识的觉醒，用户的档案信息需求层面不断加深，需求领域也不断拓展。

首先，随着社会的进步和档案意识的增强，人们对精品化信息与专业化的知识服务的需求越来越强烈。他们需要的已不再是简单的获取文献，而是如何从繁杂的信息环境中捕获和析取解决所面临问题的信息内容，并将这些信息融化或重为相应的知识或解决方案。

其次，随着用户信息素养的不断提高，档案用户已从信息服务的“被动接受者”转换为“主动选择者”。 他们更希望获得一种为自己量身定做的个性化信息服务。而相对于传统档案的利用方式和管理方式，基于这种“定制服务”的数据采集与分析显然更具有方向性和现实性，更能发挥大数据的威力。下面我们针对这种情况举几个例子，借此说明这类情况的数据采集来源。

② 提升档案馆服务效能的数据采集

档案馆可借助大数据技术，对档案馆用户身份记录、借阅记录等结构化数据

及存储行为、搜索方式、行为轨迹乃至SNS上的言行记录等半结构化数据进行分析，有效发现用户隐性诉求，分析研究本馆档案查阅受众人群与利用程度，从而更好地提升档案馆的服务效能。通过分析用户对馆藏目录的点击率，档案馆还可选取点击率高的档案进行数字化，进而开展深层次的信息服务。同时还可以利用微信、APP等工具进行相关档案的主动推送服务。此类档案数据的收集主要来自于利用大厅视频档案数据、网站馆藏目录点击率、电子调阅单等。

③ 提升馆舍智能管理水平的数据采集

以库房监控与管理为例，档案馆可参照国家档案局《档案库房温湿度标准》将档案馆库房的温度控制在14～24℃±2℃，相对湿度控制在45%～60%±5%，通过收集各个时期温湿度智能控制系统的设置数据，并对数据进行科学有效的分析，找到库房中温度与湿度变化的规律，在中央空调工作期间能够很好地进行温湿度反馈，从而帮助中央空调进行温湿度的调节。达到在最容易忽略湿度的春秋温度适中的季节里或者是连续阴雨天气里，借助加（除）湿机进行库房内的湿度调节的目的，使作为“孤本”保存的纸质档案长期处于一个适宜的温湿度环境中。此类数据采集范围可定向为库房温湿度电子记录。

2、政务网档案数据采集

政务网上的档案数据采集与政务网环境下运行的档案管理系统有关，其中包括电子邮件数据采集、电子档案接收等。

（1）电子邮件数据采集

截止2013年6月，电子邮件用户达24665万，占网民使用率41.8%，手机邮件占手机网民的29.1%，比2012年12月增长1.8%。由此可见，电子邮件已成为电子政务、电子商务中必不可少的信息资源。电子邮件系统作为一种日益普及数据业务的交流手段，每天都承载着大量的数据交换。这些交换的数据中包含大量有用的信息与知识，鉴于目前电子邮件点对点交流的特点，这些信息、知识只能保存在通信的各方，不能得到充分的利用。对于政府机关、企事业单位来说，电子邮件作为日常内部和外部交流的重要手段，很多重要的信息和知识都包含在电子邮件中，而这些信息和知识是政府机关、企事业单位的宝贵财富，分散在个体不能共享，是一种资源的浪费。为了能够方便有效地收集和管理有用的邮件信息，应

加强电子邮件知识管理的研究，将电子邮件加入档案大数据仓库，从而提高电子邮件的效用。

国家档案局于2005年就制定了《公务电子邮件归档与管理规则》。随着办公自动化的深入发展，档案业务部门对电子邮件的档案数据采集进行了有益实践，通过电子邮件数据采集、数据存储管理和搜索等功能模块，有助于更清晰有效地识别、提取和保存电子邮件中的各类重要信息，从而达到共享利用有用电子邮件信息的目的。此外针对电子邮件的数据采集，既要加强对电子邮件采集的内容、背景、结构和元数据等个体的完整研究，又要加强对电子邮件原生次序的有机联系整体的完整研究。

（2） 存量档案数字化

在国家档案局的强力带动下，全国正大力推进馆藏档案数字化工程，积极争取国家资金支持档案部门大力开展馆藏传统载体档案的数字化工作，馆藏数字化档案势必成为档案大数据重要数据来源之一。

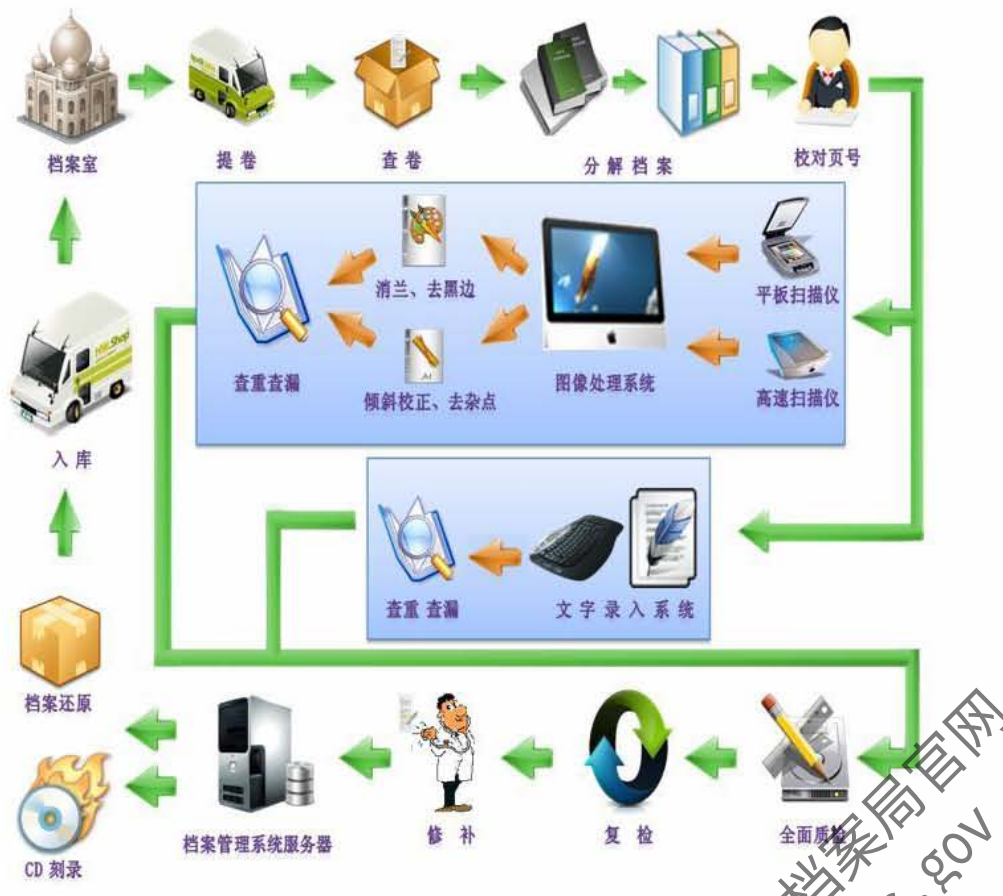

图 1　档案数字化流程图

对于这一类档案数据，档案馆应按照元数据模型，将由传统载体转化的数字化档案数据，整体统一归档，建立结构合理的电子档案数据库及目录数据库，运用各种技术手段将传统的影像档案及实物等档案进行现代化处理形成多媒体数据库。对于传统档案转换的电子档案的结构数据、元数据、数字化版本都要描述清楚，以备来日查找利用。

3、 增量数字档案的接收

增量数字档案即是指对电子文件和电子档案的接收。档案馆数据资源种类繁多。在档案馆的数据资源中，既有数字化的纸质档案、也有接收进馆的电子文件、音视频档案等。在国家电子政务背景下，今后更将有海量的电子档案持续不断进入档案馆。可以预计，数字形态的档案信息将逐步成为档案大数据的重要组成部分。

四、档案大数据信息挖掘流程

简单的说，数据挖掘就是从大量数据中提取或“挖掘”知识的过程，此过程通常包括六个基本步骤：定义问题、准备数据、浏览数据、生成模型、浏览和验证模型、部署和更新模型。从档案信息挖掘角度来讲就是对现有档案数据信息进行分析，将档案信息内在之间及内与外在所包含的信息进行组合提炼，最终将所需要的结果呈现出来。此过程并非现成软件系统自行匹配就能够完成的，需要在对馆藏档案有足够的了解的基础上，精心制定方案，准备整理资源、指导技术实施、部署应用。关键流程应包括以下几点：

1、档案大数据信息挖掘前期规划

设想数据挖掘的预期目标与效果，确定主题。事实上档案部门开展海量数据挖掘与现有商业性数据挖掘主要目标定位是有所差别的，不能纯粹跟风效仿；数据挖掘技术实施与平台建立需要投入大量的人力物力，筹集大量资金，目标定位不准确，可能导致最终成果毫无利用价值，造成资源浪费的现象。因此档案部门应对数据挖掘目标要有清晰的认识、对预期效果有准确的定位。应与国家信息化事业的战略取向保持一致，应该围绕“社会效益最大化”的目标，充分了解政府、公众及当前档案工作的需求，尤其是网络应用的需求，从而预定哪些资源需要进行数据挖掘，需要到达怎样的质量指标，而不是盲目开展。

制定数据挖掘工作方案。开展的项目总体规划和安排，它既是项目实施依据，

也是续后的监控、协调和管理依据。通常包括以下内容：

- 确定指导思想、任务目标和阶段目标；
- 制定项目详细工作内容；
- 确定资源范围、类别、规模、技术要求、技术路线；
- 设定完成项目工作目标所遵循的标准、使用的设备、操作方法和技术手段；
- 预期成果——说明项目完成总体目标，预期达到的有形或无形成果和社会效益；
- 风险控制——制定安全管理的实施策略、实现方法、实施组织形式；
- 详细说明本部门和承担企业的各自分工的主要内容，确定责任与人员；
- 项目实施预算表——项目实施所需的费用分类汇总；
- 方案论证——对方案的先进性、适用性，资金投入上的合理性、实用性，实施上的可能性、标准及制度的可操作性、风险性进行全面科学的综合分析，为项目决策提供客观依据的一种研究活动；
- 建立工作组织机构——根据确定的项目目标，明确划分分解目标，列出所要进行的工作的内容，制定岗位职责标准与考核要求，使之成为有秩序、高效率、部门合理分工、密切协作的数据挖掘管理组织体系。确定组织领导者、参与者，明确任务、责任与分工；
- 制定预算——规划所需设备与人员，充分考虑潜在的资金投入，编制合理的预算方案，获取政策支持与资金支撑，开展数据挖掘项目。

2、建立档案大数据信息存储架构

该部分主要是根据前期制定的目标和规划，遵循档案数据挖掘的原则，搜索所有具有档案价值的内部和外部数据信息，并从中选择出适用于数据挖掘可用的档案数据建立资源仓库。这些庞大数据资源结构复杂多样，需要有足够的空间和选择适合的存储解决方案，通常现有传统数字档案存储设备、机制及技术手段等方面很难满足大数据挖掘的需求，需要考虑对整个存储架构与数字档案管理模式进行革命性的重构，并且要适当超前考虑，使存储能力的能够满足档案数据量的增长，这与传统方式相比存在着质的不同。

① 基于可扩展性建立数据存储平台

在传统的数据仓库上进行对相似数据集的挖掘操作，一般都在一个单独的存储设备上进行。现在这种方法对大数据处理技术和存储容量的可扩展性来说已经不是最优的选择。在大数据存储中，更多的应用是一次写、多次读，读得更多是大数据存储的一个特点，而在传统的数据存储中，读写是随机的，由于每个应用不同，其读写的比例也是随机的;大数据存储需要具有横向的可扩展性，并可支持多种接口、多种数据访问协议，便于不同数据进入这个大数据平台。

②基于数据分散性建立分布式存储体系

大数据存储通常采用分布式存储体系，分布式存储体系，将大规模海量数据用文件的形式保存在不同的存储节点中，并用分布式系统进行管理。其技术特点是为了解决复杂问题，将大的任务分解为多个小任务，通过让多个处理器或多个计算机节点参与计算来解决问题。分布式文件系统能够支持多台主机通过网络同时访问共享文件和存储目录，使多台计算机上的多个用户共享文件和存储资源。分布式文件系统架构更适用于互联网应用，能够更好地支持海量数据的存储和处理。基于新一代分布式计算的架构很可能成为未来主要的互联网计算架构之一。目前典型的分布式文件系统产品有GFS（Google File System 文件系统）、HDFS（Hadoop 分布式文件系统）等。

③基于数据分析建立网络存储系统

传统的网络存储系统采用集中的存储服务器存放所有数据，存储服务器成为系统性能的瓶颈，也是可靠性和安全性的焦点，不能满足大规模存储应用的需要，在大数据时代，数据分析是需要从数据围着处理器转改变为处理能力围着数据转，将计算推送给数据，而不是将数据推送给计算，分布式网络存储系统采用可扩展的系统结构，利用多台存储服务器分担存储负荷，利用位置服务器定位存储信息，它不但提高了系统的可靠性、可用性和存取效率，还易于扩展。

④基于非结构性建立存储方案

传统档案数据以结构化或半结构化数据为主，但档案大数据资源通常是非结构化数据，包括文本、音视频、动画、图像各类文件格式纷繁复杂，特别是在网络中采集的档案信息资源，格式类型更为复杂，包括公务邮件、网页、博客、微博等，格式类型有 XML、HTML、各类报表等。因此应改变以结构化为主体的单一存储方案，采用分而治之的思想，使用分布式文件系统进行存储，方便增加

节点实现大数据稳步处理。

3、进行档案大数据信息预整理

档案大数据信息预整理主要指在数据挖掘处理以前对数据进行的一些前期整理。现实中档案数据有些是不完整的或冗余的，或与数据挖掘目的不一致的，或有些数据是影响挖掘结果正确性的，导致挖掘结果差强人意，甚至有些信息是有害的信息，如网络中的不当言论、反动信息等。为了提高数据挖掘的质量和效率，需对这些档案大数据资源进行预整理，为资源挖掘做准备，包括根据既定主题对现有资源进行分类、剔除冗余数据、填补关键信息、数据格式转换等。具体如下：

① 档案资源分类

档案大数据挖掘整理分类与档案业务管理中的分类有所不同，一般有固定模式，如文书档案、人事档案、会计档案、科技档案、声像档案等。进行档案大数据分析整理时的分类可以更为广泛、多角度、多维度进行多重分类，如重大事件、统计资料、人文、地理、历史等等，分类方式可以更为丰富多样。档案大数据资源分类的依据主要取决于分类对象的属性或特征。

②数据清理

通过填写缺失关键信息值、识别或删除偏离目标信息并解决不一致性来“清理”数据。主要是达到如下目标：格式标准化、异常数据清除、错误纠正、重复数据的清除等。

③数据集成

把不同来源、格式、特点性质的档案数据在逻辑上或物理上有机地集中，多个数据源中的数据结合起来并统一存储，建立档案数据仓库的过程实际上就是档案数据集成。

④数据变换

规范化档案数据使其适用于数据挖掘的形式，让信息数据能够快速、高效、准确地被计算机所识别，从而使得采集上来的档案数据能够提供更好的应用服务。

⑤数据归约

在进行档案数据挖掘时往往数据量非常大，在少量数据上进行挖掘分析需要

很长的时间。数据归约技术可以用来得到数据集的归约表示，它小得多，但仍然接近于保持原数据的完整性，并且结果与归约前结果相同或几乎相同。

4、建立档案大数据分析模型

数据挖掘是多门学科、多个领域的综合性技术，包括统计学、机器学习和数据库等领域内的研究成果，其它还包含了可视化、信息科学等内容。数据挖掘纳入了统计学中的回归分析、判别分析、聚类分析以及置信区间等技术，机器学习中的决策树、神经网络等技术，数据库中的关联分析、序列分析等技术。建立适合挖掘算法的模型是档案数据挖掘成功的关键，但具体采用那些技术、算法需要根据资源类型、目的、实现方式等方面进行综合考虑。

挖掘算法建立是对档案隐性和显性知识的内在和彼此关联因素的分析基础上，通常是复杂的非线性关系。主要解决要从哪些方面或角度开展档案数据分析，各方面包含什么内容或者指标，要建立怎样的数据关联等。数据分析模式建立方法也有多种，常用的分析方法有回归分析、聚类、关联规则、特征、变化和偏差分析、Web 页挖掘等，它们分别从不同的角度对数据进行挖掘。

①分类

分类是找出数据库中一组数据对象的共同特点并按照分类模式将其划分为不同的类，其目的是通过分类模型，将数据库中的数据项映射到某个给定的类别。它可以应用到档案的分类、档案的属性和特征分析、公众需求热度或兴趣分析及公众利用档案资源趋势预测等，如档案馆可根据利用人群利用档案情况进行分析，据此进行档案分类定向提供服务，更能提高档案利用率以及利用者的利用兴趣。

②回归分析

回归分析方法反映的是事务数据库中属性值在时间上的特征，产生一个将数据项映射到一个实值预测变量的函数，发现变量或属性间的依赖关系，其主要研究问题包括数据序列的趋势特征、数据序列的预测以及数据间的相关关系等，它可以应用到档案管理的各个方面，如根据档案资源寻找和探索历史事件产生原因、发展过程、及发展趋势等，档案保存环境对档案存储介质的影响，档案载体在生命周期内的重要变化，针对某类档案什么时间段利用最频繁等等。

③聚类