国家档案局科技项目（项目编号：2011-X-26）

《网络信息资源归档与利用平台建设的研究》

研　究　报　告

辽 宁 省 档 案 局

目　录

从档案学角度讲，网络信息文件本身就属于电子文件的范畴，许多电子文件管理理念可以应用到网络信息资源管理的过程中。

1.电子文件生命周期理论对网络信息资源归档的借鉴意义

网络信息资源可以按照电子文件的管理方法进行管理，资源文件的生成、发布、捕获、归档、利用是一个完整的运动过程，应当对其全程控制，不仅保存网站文件本身，还要连同元数据、变化日志、插件程序等一起保存，才能保证资源文件的真实性和完整性。

2.档案鉴定理论对网络信息资源归档的指导作用

早在20世纪80年代,法国档案学者罗尔德·瑙格勒提出了电子文件的“双重鉴定论”，一方面要判断电子文件信息的有用程度;另一方面要判断电子文件有用程度实现的可能。网站上的信息以多种格式存在，并且具有多种表现形式，需要从技术上判断其有用程度实现的可行性。对于网络资源文件内容上的鉴定，可以借鉴加拿大档案学者特里·库克（Terry Cook）的“宏观鉴定战略”，从能否反映该机构的职能，能否反映当时的社会环境，能否满足人们的社会期望等角度进行鉴定。

3.文件归档理论对网络信息资源的归档的启示

信息归档与备份是有很大的差别。备份是指数据副本，

一旦原始数据丢失或者损坏且无法修复，副本就可用于恢复原始数据。备份用于保护正在使用的数据。归档是指一份或一组数据记录，专门用于长期保留并供将来参考。

两者之间的一个区别值得注意：备份是数据副本，而归档是原始数据从初始站点移除后，发送到其它站点，用作长期保留。网络信息备份强调信息的可恢复性，在出现意外情况时保证业务的连续进。特别是在线备份，对速度的要求相对较高，但对容量的要求则相对较小。

归档的目的与备份完全不同。一些网站，每天产生的大量数据，是有60%-80%的信息发布完以后的一段时间后，其即时应用的可能性逐渐降低，但不意味着永远不会应用，这些数据必须保留以备查询，保证数据的访问。网络信息资源归档时要最大限度地保留信息资源的原始性，尽量在归档后实现网页上的全部功能，保留网站上的所有的结构、内容、形式和链接等表现特征，以及网络资源的发布者、发布日期、责任人等相关重要的元数据。因此其方案选择也和备份有很大区别。

（二）网络信息资源归档原则

1.保证网络信息资源的完整性与原始性

档案的完整性与原始性是档案的重要属性，网络信息归档也必须保证这一基本属性。网络信息资源依托网站存在，保存其完整性与原始性最好的方式就是以网站为单位进行

归档，同一个网站的所有信息资源保存在一起构成该网站的“全宗”。网页与网页之间的链接关系和网页与程序文件的依附关系也不能被破坏。并且在归档时补录相关元数据，如归档日期、负责单位、联系人等。这就要求在网络信息采集归档时尽量保持信息资源的原貌，体现档案的归档的要求，既：完整性、原始性、可追索性。

2.保持信息的长期可读性

在保障存储环境安全与存储介质安全的前提下，需要其存储格式也“安全”，在若干年后仍能够正常打开。虽然大部分数据重复读取的可能性降低，但归档平台并不是一个“死”仓库，必须保证在需要的时候能方便地读取数据，并且能够脱离网络环境、硬件环境、软件环境，这也是归档系统的必备原则之一。

3.遵从相关法律和法规

网络信息具有交流快捷、获取容易、共享面广等特性，超出了时间性与地域性，使得网络信息的相关管理模式与传统信息管理模式有很大不同，不能沿用传统档案的管理思维进行网络信息资源归档工作。但也不能任意开展相关工作，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。例如：我国2001年10月27日通过的《中华人民共和国著作权法修正案》，信息资源发表权、署名权、修改权、保

护作品完整权、复制权、发行权、改编权等等都有规定。中华人们共和国国务院令《信息网络传播权保护条例》（简称著作权法）2006年7月1日颁布实施，对信息的著作权人与发布人、传播人、利用人的权益都有具体的规定。从事相关网络信息管理也必须完全符合相关法律法规的要求。

4. 网络信息资源管理平台可扩展性

数据是无时无刻不在扩展的，扩展速度超乎想象。这种情况下，网络信息资源管理平台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，网络信息资源归档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，更要灵活的将网络信息资源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，使网络信息资源归档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。

5. 网络信息归档的适时性

网络信息资源消失有很多种原因，其中主要的有网站注销、域名更换、结构更改、机构调整、网站更新、服务器故障、病毒入侵等。网络信息资源归档工作应在信息资源消失之前进行，因此网络信息资源归档的时机选择也非常重要。

以网站更新为例，网站更新存在以下四种情况：①经常

更新，更新间隔的时间是三个月以内；②不经常更新，更新间隔的时间是三个月以上；③有规律地更新，按照计划有规律地进行变化（例如一周一次，一天两次）；④不规律地更新，没有按照计划进行更新，更新时间比较随意，更新间隔的时间也是不确定的。按照以上四种变化情况，网站可以分成以下四种类型：规律且经常更新的网站、规律但不常更新的网站、不规律但经常更新的网站、既不规律也不经常更新的网站。捕获网站文件的时间一般由网站的变化情况来决定的，但往往跟踪不规律更新的网站难度很大，因此需经常性（每天归档或每周）归档。

6. 选择性原则

网络信息数量巨大，预将其全部归档理论上是可以实现的，但需多种因素齐备，如政策支持、经费支持、技术支持、管理支持等多种因素，目前将网络信息资源全部归档只是一种美好愿望。另一个更重要的原因是网络信息资源来源庞杂，混有大量毫无用处的垃圾信息或有害的信息。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。有选择性地归档保存不仅可以节省人力、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。

有选择性归档就必然涉及信息鉴定，可以借鉴档案学的

文件鉴定理论来判断网络信息的价值，确定网络信息归档的对象，应搜集需要的、具有特色的网络资源；来源可靠、内容新、背景明确、发布规范的精品资源；收集信息资源具有完整性；并尽可能地利用开放的免费的资源。可以概况为四个字，即：特、精、全、省。

（三）网络信息资源组织机制构成

网络档案信息组织机制构成是一项复杂的系统工程，单纯依靠科学的组织方法难以保证组织实施的进程和质量。有必要综合分析影响网络档案信息归档、利用的相关因素，认真正视现存问题，为组织顺利实施保驾护航。同时，可以为网络档案信息资源建设标准与规范提供建设性意见。所谓网络档案信息组织机制构成，是指在网络档案信息资源在归档、管理、利用、长期保存等整个生命周期中，分析影响其过程和功能的相关要素，理清各要素之间的关系，使之协调配合，最终形成参与、指导、规范、管理组织工作的基础性体系。从系统论的角度出发，网络档案信息资源组织机制构成涉及管理规范、技术支撑、组织管理、软硬件设施设备和人才队伍建设五个方面，整体框架见图5。

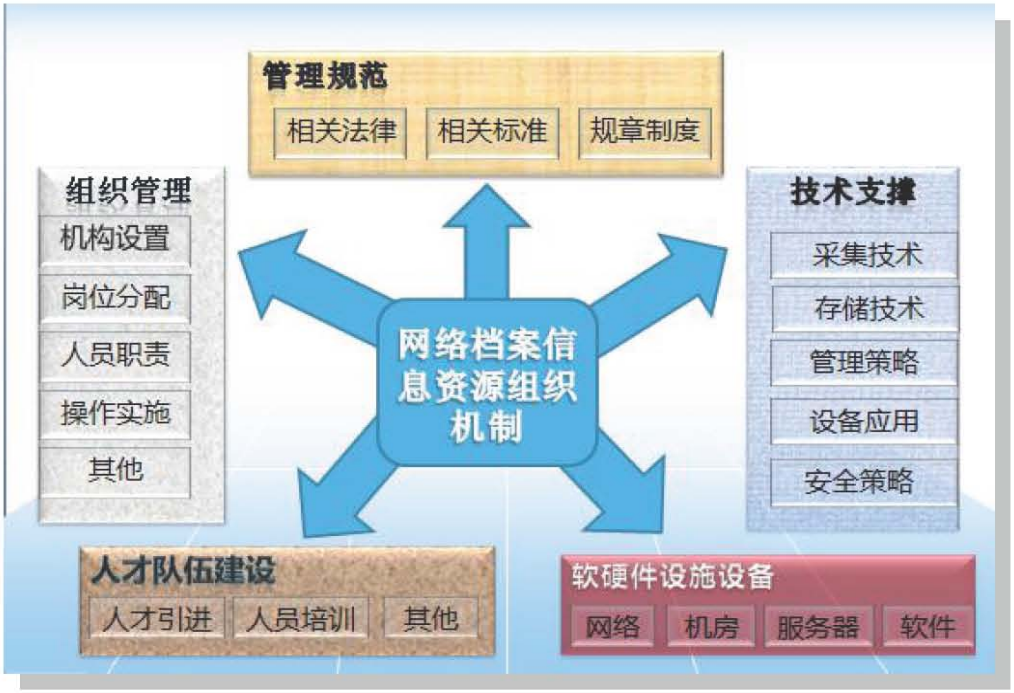


图 5　网络档案信息资源组织机制

网络档案信息资源组织机制构成的根本宗旨：通过建立科学合理的组织机制，服务于网络档案信息资源组织活动全过程，为组织工作顺利实施提供体制、技术、制度、标准、人才保证，以机制的先进性确保工作的先进性。网络信息归档、利用整体构成理论和实践研究成果以其专业性、科学性、先进性，是指导网络档案信息资源组织的重要内容。

管理规范是开展网络信息资源归档与利用工作所必须遵循或参照的一系列准则（包括相关法律，国家、行业、地方、企业标准以及规章制度）。管理规范统领网络信息资源科学归档与利用的前提，管理规范中相关法律、标准的缺失、滞后、不执行性、不实用性等，会导致开展的相关工作缺乏科学和理性的指导，影响信息资源的归档和利用质量。研究完善的管理标准能够使档案信息资源归档与利用工作置于科

学轨道之上，便于协调简化网络信息资源建设程序，统筹安排资源，节约成本，避免重复开发等。

组织管理是指围绕网络信息资源信息归档与利用整个流程工作的机构、部门、岗位、人员等的合理配置，是组织支撑，是对各个部门及人员的总体协调与安排。组织管理是一个较为复杂的有机体系。一个健全的组织管理体系能够明确相关部门及人员所担负的的责任，包括执行或维护网络资源归档与利用中相关环节的责任，保护特殊档案信息资源的责任以及执行管理程序的责任等。如果每一个人、每一个部门都能坚守自己的岗位，各司其职，又能彼此协调沟通，会使整体工作有序、科学、均衡、合理发展。

技术支撑是达到网络信息资源科学合理的归档与利用的一系列手段或活动。包括网络信息资源采集技术、管理技术、存储保护技术、基础设施保障及设备应用技术、安全管理技术等。技术在网络信息资源归档与利用工作中发挥至关重要的作用。

软硬件设备设施建设是维持正常开展此项工作的配套设施与设备的整体。包括机房建设、网络建设、载体保护设备、软件平台建设等。是正常开展网络信息资源归档与利用工作的基本条件，并决定了信息资源所处环境，是物质基础。

人才队伍建设是开展网络信息资源归档的智力支撑。对于从这项前沿性工作的人员既需要有爱岗敬业的精神，又需

要掌握一定信息管理技能，并且能够及时发现问题，勇于克难求进。此外应注重技术人才的引进和人才培训工作。

各构成要素内部之间和其他构成要素之间相互补充、互为基础，缺一不可。管理规范是技术保障和组织管理保障的支撑系统；技术保障是管理规范和组织管理实施的重要手段；组织管理保障是落实法律、标准要求的关键环节；基础设施是技术实施、业务开展的物质基础和必要条件。

随着国内外网络信息资源组织理论的日趋发展成熟，其先进的理论、技术和方法为网络档案信息资源的高效组织提供了可能。以便符合档案特点的网络信息组织理论和技术引入网络档案信息资源组织研究，坚持理论联系实际，形成该领域的组织特色。当然，网络档案信息资源组织机制是档案信息化建设过程中十分前沿的研究课题，涉及的知识面非常广泛，许多问题尚没有很成熟的解决方法，也不可能存在一劳永逸的解决方案，还有待今后不断地深入研究。

（四）网络信息资源保存的系统模型

1.OAIS模型

（1）OAIS模型简介

在数字信息长期保存领域，一直以来都面临着一个难题，就是如何保证数字信息的长期保存与持续利用。OAIS参考模型的出现，为数字档案信息进行长期保存和持续利用提供了一个最佳途径，目前OAIS模型在国内外许多数字信息长

期保存项目中得到应用。通过实践证明，该模型已成为一种开放性数字信息长期保存的基本框架和所应遵循的原则。

OAIS (Reference Model for an Open Archival Information System，即开放式数字信息系统模型)是1995年在国际标准化组织(ISO)的请求下，美国国家航空和航天局的空间数据系统咨询委员会(CCSDS)开始开发的一个规定概念和参考框架。参考模型几经修改扩充，于2002年1月最终通过审核，正式成为一项新的国际标准(ISO：14721)。

OAIS信息模型提出了信息包的概念，它将信息在系统中输入、运转和输出的结构概念化，分为提交信息包(SubmissionInformation Package，SIP)、存档信息包(Archival InformationPackage，AIP) 和分发信息包(Dissemination InformationPackage，DIP)。提交信息包是信息生产者提供给OAIS的信息包。一个或多个提交信息包需要被转换成一个或多个存档信息包进行保存。存档信息包有一系列完整的保存描述信息和相关的内容信息。最后，根据利用者的请求，OAIS需要以分发信息包的方式提供一个存档信息包的所有或者部分内容给利用者。OAIS还有六项功能，分别是摄入功能(Ingest)、档案存储功能(Archival Storage)、数据管理功能(Data Management)、系统管理功(Administration)、保存规划功能(Preservation Planning)和存取功能(Access)。这六项功能为OAIS数字存档系统的总

体功能框架的设计和实现提供了较为完整的高层概念框架模型。

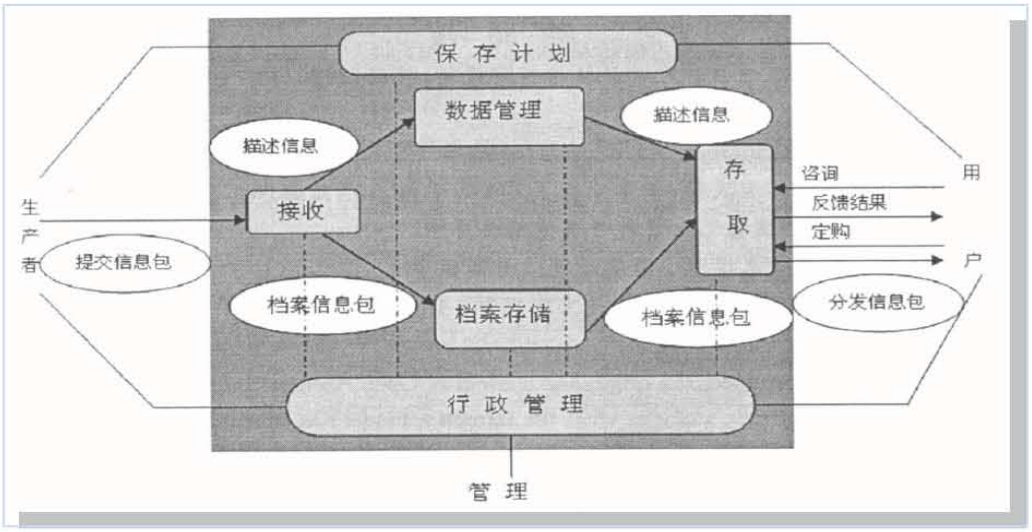


图6 OAIS 功能模型

（2）OAIS参考模型的适用性

OAIS参考模型不定义任何实施这些概念的特殊方法。具体实施者参考OAIS参考模型为开发提供特定服务和内容的指导，但模型不假设或局限于任何特定计算机平台、系统环境、系统设计范例、系统开发方法、数据库管理系统、数据库设计范例、数据定义语言、命令语言、系统界面、用户界面、技术、所需媒体。因此，真正对数字资源长期保存系统设计或实施需要根据实际情况将功能组合或者分解。

（3）OAIS模型在网络信息资源保存中的应用

国外基于OAIS的研究项目具有代表性的有：美国国会图书馆领导实施NDIIPP项目；加利福尼亚大学(DPR项目)数字保存仓储是加利福尼亚大学图书馆数字保存计划的基础；荷兰国家图书馆针对长期存取荷兰电子出版物的需要而提出

的专注于长期存储和大规模存档的e-Depot系统；由欧洲国家图书馆联合会常设委员会发起的， (NEDLIB Networked European Deposit Library)网络化的欧洲存储图书馆项目；英国Cedars的分布式数字档案原型系统(The Distributed Digital Archiving Prototype)项目；以及美国的ERA项目，澳大利亚的ADRI项目等。

我国基于OAIS的研究成果，主要应用在数字档案与数字图书方面的管理工作。如：中科院档案馆—中科院档案馆数字档案馆；国家图书馆—中文元数据方案(CMDS)；各省开展的数字档案馆建设项目等。

2.其他系统框架

在众多网络信息资源采集保存的项目中，并非都采用OAIS模式，一些模式也非常实用和简便。例如Internet Archive项目，该项目是开始比较早的网络信息资源保存项目，今天已经有比较大的规模，并开始和很多机构合作。Internet Archive项目建设并没有采用OAIS模型见图7。

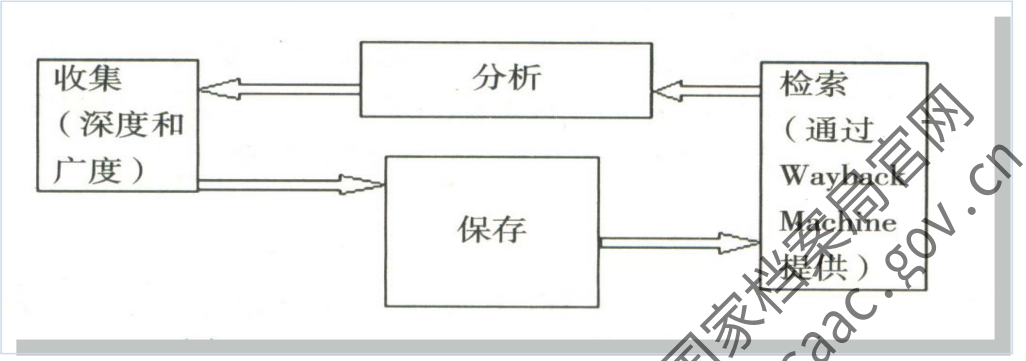


图 7 Internet Archive 系统模型

Internet Archive系统主要由四个部分组成：收集、分析、保存和检索、收集主要包括两个方面，一个是宽范围收集，另一个是窄范围收集。宽范围收集即对多个网站进行全面收集，窄范围收集是对选定站点或选定主题来收集。在该模型中存储和保存是合并在一起的。收集到的网络信息资源采用专门的数据格式ARC存储在磁盘中，针对不同类型的网络信息资源采用不同的存储载体，并初步决定对存储在数字线性数字线性磁带上的数据定期进行迁移。然后通过检索工具向用户提供检索。

除此之外还有美国国会图书馆从2000年开始的网络信息资源保存项目MINERVA,以及澳大利亚的PANDORA项目,在建设过程中，都形成了具有自身实际情况保存与利用模式，也是非常实用和成功的。从中可以看出，无论采用OAIS或是其他模型构建的网络信息资源保存系统，需具备这样一些关键条件，能够支持机构之间互相合作，具备保存和检索分开处理的功能，采用被广泛接受的协议，还应该具备可扩展性，能够进行局部更新，而不是必须要进行系统重新。

（五）网络信息资源归档方式的研究

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归档方法,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:

选择式归档、全面式归档和综合式归档。

1.选择式归档

选择式归档是根据某种标准,对互联网资源进行选择后再归档的方法。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归档。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:

（1）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;

（2）对每个被选定的主题可以制定单独的归档程序表,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;

（3）对归档后的各个条目可以进行完全著录,进而并整个归档数据库中;

（4）通过与出版者事先的协议，解决版权问题,这样,每个归档后的条目可以立刻被读者通过互联网进行检索利用;

（5）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;

（6）不能由智能程序进行归档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，其缺陷也是显而易见的，其缺点是：

首先在做出选择时，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。在这里，移植了纸质文献收藏的方法，而印刷版的环境是相对明确和稳定的，甚至是可预测的。但网上资源却迥别于传统的纸质文献，而且其产生、发布和利用仍然处于初期阶段，含有极多的变化因素。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，一种说法是有很多信息即便在未来也是无用的，自然不在选择之列。这固然是个理由，但缺乏足够的力证。

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，尤其是博客、微薄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。因此，许多国家的相关项目采取了目标选择

性保存方式也就是可以理解的了。

2. 全面式归档

全面式归档是尽可能将所有互联网资源进行归档的方法。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，全面式归档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。

但全面式归档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，全面式归档中近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. 综合式归档

综合式归档是两种归档方式综合运用。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。

分析各个归档方法，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归档,再辅之以与发布单位签订的归档协议。

（六）网络信息资源归档与利用的策略与流程

1.收集方案

（1）收集策略

在开展网络信息资源长期保存工作前，第一个需要解决的问题就是如何确定保存资源的范围。事实上，保存所有数字资源不仅没有必要而且也很难做到，因此如何做到有的放矢恰当地选择合适的资源作为长期保存的对象就成为了首要问题。

网络信息收集与传统档案收集方式不同，纸质档案电子档案的归档与征集有法可依，归档单位、移交单位、归档时间、归档范围、征集范围、征集对象、征集方法等都是较明确的，而网络信息资源归档对于档案部门来说尚属全新领

域，但档案部门开展此项业务，必然不能脱离和必然依据档案部门的业务特点、职能范围、业务领域等，制定具有档案特色的一整套业务流程和管理策略。

从档案学理论看，依据传统的档案收集办法，相应的网络信息资源收集有两种方式。一种是呈缴式，即网络资源所有者遵从法律法规向归档部门呈缴本单位的网络信息资源，归档单位负责保管、开发和提供利用。另一种是征集式，即归档部门与网络信息资源责任单位达成协议或默契，主动向社会征集网络信息资源，同时提供保管、开发和利用。第一种方式需要相应的法律法规支持，并具有一定的强制性，归档单位和信息责任人都在法律约束下开展工作，此种方式相对征集式较稳定。第二种方式需要归档部门主动承担责任较多，工作难度相对较大，持续长期开展此项工作，受内部机构、职责的变动、稳定的资金支持影响较大，但在尚无相应的法律法规支持下，采用第二种方案开展此项科研工作是档案部门的必由之路。

（2）收集技术

网络信息资源归档从技术角度来讲，就是运用采集软件自动完成从网络下载到本地的过程，一般大型网站建站的时候，固定的内容 如新闻、下载等都是使用固定模板、自动生成静态页面的方式，这样就使得在源码中表格等设置都是一致的。采集就是利用这样的一个原理，搜索页面中与采集

设置相同的部分，然后搜集网站内容进入数据库。 比如：某站的新闻在源码中是这样的

<table class="news"><tr><td>新闻内容</td></tr></table>

很容易可以看出，上面就是一个表格，然后包含新闻内容，设置采集方式的时候，就可以 从遇到页面的<table class="news">这个标记开始，到下一个</table>标记结束。运行采集后，就会将该站所有的新闻全部采集下来了。

深层次的采集原理主要是通过Web页面之间的链接关系，从Web上自动的获取页面信息,并且随着链接不断向所需要的Web页面扩展的过程。粗略的说，它主要是指这样一个程序，从一个初始的URL集出发，将这些URL全部放入到一个有序的待采集队列里。而采集器从这个队列里按顺序取出URL,通过Web上的协议，获取URL所指向的页面，然后从这些已获取的页面中提取出新的URL,并将他们继续放入到待采集队列里，然后重复上面的过程，直到采集器根据自己的策略停止采集。上面只是简单的一个采集的举例而已，实际应用中会比这复杂的多。

网络信息采集并非新技术，目前网络上网站采集软件也有很多，常用的有TeleportUltra、WebZip、Mihov Picture Downloader、WinHTTrack HTTrack、MaxprogWebDumper等，这些软件功能各异，有的可以设置采集时段，有的对多媒体

下载支持好、有的对脚本语言采集支持强。但基本原理一样，自动获取或手动添加网络资源网址，快速浏览网页，将网站以静态页面的型式保存在硬盘中。有一些软件是开源的，可以在网络上免费获取。采用这些软件可以实现离线浏览某个网页、将网站镜像、或把别人网站的内容搬到自己的网站内。但与档案部门网络信息资源归档的目的不相符。网络信息资源归档并不是简单的进行资源备份或存储。网络信息资源归档与OA系统中流转的电子文件归档一样，需要将网络信息各个时期变化，信息原貌、及其相关元数据完整的采集进行归档。对目前现有的软件采用技术可以参考，建立符合归档要求的一套适合档案部门要求的软件平台。

2. 网络信息资源长期保存的机制

网络信息较一般电子文件更为多样化、复杂化、和多变性，在长期保存策略中，用于OA系统或原生电子文件归档保存策略，不能完全适用与网络数字信息归档管理。例如网络信息资源之间关联性较强、文件格式类型丰富、表现形式多样。强制统一网络信息格式、会丧失网页本来面貌，降低利用效果；又如同普通电子文件一样，网络信息资源对软硬件具有依赖特性，保存软硬件环境，网络信息资源达到一定规模后，其保存难度逐渐增大，档案工作者的工作量和资金负担会远远超出自身的能力。因此需制定科学、稳妥的适用于网络信息资源的长期保存机制。

（1）规范网络信息资源管理途径。制定相关标准和工作规范，使网络信息资源归档管理有章可循，保证网络信息资源从采集到长期保存管理上的连续性和规范性。

（2）网络信息资源归档保持原有结构、面貌，并且脱离原来的软、硬件环境。采取相应措施，使电子文件脱离原来的软、硬件环境，并且保持原有的网络信息文件结构和本来面貌，从而达到长久保存的要求。

（3）确保信息的长期可读性。保证网络信息的长期可读，是对信息进行利用和研究的出发点。研究解决网络信息长期保存问题，例如网络信息元数据问题，永久标识符问题、归档格式问题、可迁移性等问题等。同时在管理过程中，应特别重视防范病毒侵袭、控制非法存取、合理进行数据备份、数据及时迁移。制定完善的管理制度，数据载体应经常进行检测，判定被检测载体是否需要重写或更新，对检测出错的载体进行有效的修正或更新，以防突然损坏造成数据丢失。定期进行信息的有效性验证，设备环境更新时应确认库存载体与新设备的兼容性，不兼容时应进行的载体转换工作，保证信息的长期可读性。

3. 利用策略

在保证知识产权情况下，重点开展网络信息开发利用试点项目，分步实施，分类实施，探索适合我国国情的网络信

息资源开发利用工作的经验。以“需求主导、效益明显、适宜推广”为原则，发现和培育信息资源开发利用的典型，及时总结成功经验。信息资源开发利用工作是一项的长期任务，必须发挥各方面的积极性，各尽其职，齐心协力，常抓不懈。

研究网络信息提供利用的模式，包括简单利用和综合利用，前者通过再现网络信息的原貌满足用户要求，后者通过知识挖掘和知识发现，为用户提供新的知识。

七、辽宁省档案管理模式下网页归档与利用实证研究

（一）网络信息资源归档利用平台建设总体构思

1. 建设目标、规模

针对目前国内网络信息资源长期保存问题上缺乏实质性工作的现状，结合辽宁省档案馆的职能，将档案管理理念纳入到网络信息资源管理的收、管、用全过程，研究建立符合我省实际的网络档案信息资源归档与利用机制和技术手段。首先以辽宁省档案信息网为试点，开发一套适合国内网络信息特点和相关人员工作流程的网络信息归档利用平台。继而在全省乃至全国范围内对存在流失风险档案网站及其他的重要网络信息资源进行归档、整合，并适时对外发布，以维护网络档案资源的原貌。

这套系统以政府、科研、教育、文化等重要网站为主要目标，归档范围包括的各种电子政务信息（电子公文、电子

邮件以及在政务网上发布的各类信息)、互联网上其它具有档案价值的信息。

利用网络信息采集技术阶段性地进行这些网站的信息资源采集，并保存为Web归档文件。Web归档文件中包含有这些网站上当时发布的所有信息资源文件，如网页、图片、视频等。同时，系统还提供针对Web存档文件的回放与检索系统。用户通过回放系统，可以浏览、查看所存档网站昔日的旧貌，读取该网站上以前曾经发布的，但在现在的网站上已经找不到的信息。用户通过检索系统，可以像使用百度等互联网搜索引擎那样，用关键词搜索所有存档网站中的网页。所不同的是，像百度这样的搜索引擎，为用户找到的是现在互联网上依然存在的相关网页。而该平台的搜索引擎找到的却是在现在的互联网上已经不再存在，而只有在网站存档文件中存在的昔日的相关网页。通过网络信息资源归档与利用平台，可以对反映我国政治、经济、文化等诸方面的重大事件的信息进行长期保存，使中华民族的文化遗产得到应有妥善保存和保护。

2.归档平台项目建设策略

（1）网络信息资源长期保存利用逻辑框架

采用《ISO14721 开放档案信息系统参考模型（OAIS.模型）》作为平台建设的逻辑架构，在数字信息保存方面，OAIS参考模型已经成为公认的标准模式，网站信息资源的保存也

可以借鉴 OAIS 参考模型，来实现长久保存和利用。

OA1S 参考模型是对与数字信息存储系统相关的环境、功能模块以及信息对象的概念化，它为数字信息长期保存的设计提供了一个很好的参考模型。我们遵循此参考模型，进一步结合系统的体系结构、存储或处理过程、数据库设计、处理平台等技术方面的细节，就能构建符合档案业务管理和服务的数字信息长期保存实际应用系统。

通过该参考模型阐明了网络信息资源收、管、用全过程，即网络信息的产生、加工、存储、数据管理、访问和发布。它同时阐述数字信息向新媒体及格式迁移，表述信息的数据模型、信息保存时软件的角色，以及档案间数字信息的交换。

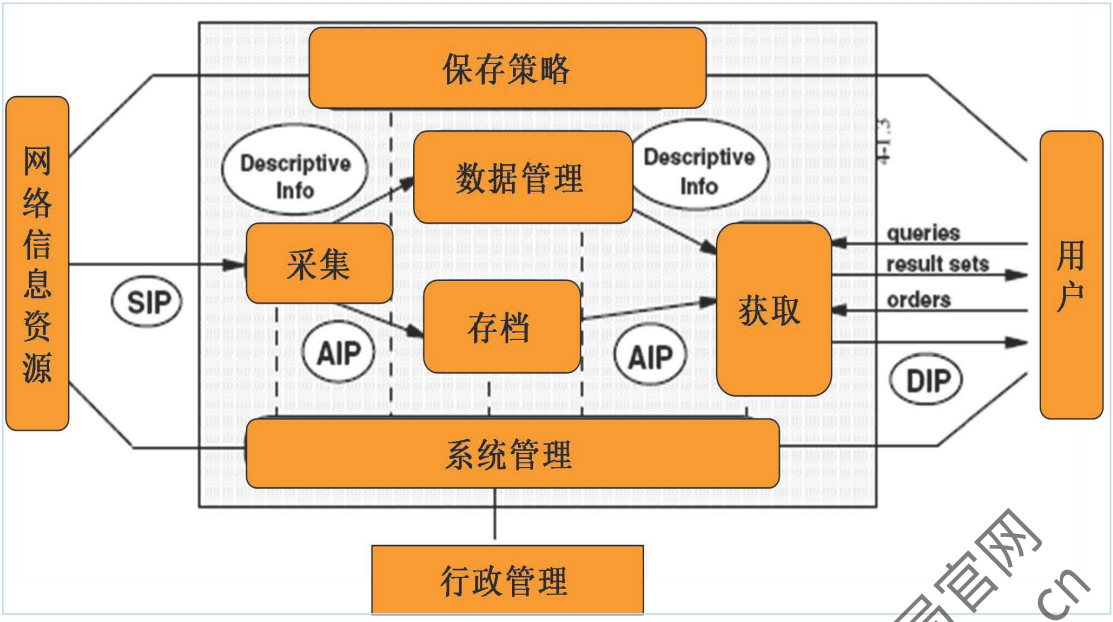


图 8 OA1S 参考模型

（2）采集技术策略

网络信息资源归档程序，采用基于HTTP协议的网络爬虫技术，从目标网站上采集网络资源文件，力图保存目标网站上各个不同时期的历史文件，而不是仅仅保存目标网站上最新的文件。

网络信息资源归档平台建设初期采用全面式归档方式，将站点信息基本无选择性的全部进行归档。力求保持站点原貌及网络资源的完整性、全面性、原始性。

在进一步的建设中，通过全面对站点评估，鉴别有长久保存价值的信息，进行有选择的归档。并重于对信息含义的理解，准确的分析信息资源类型、具体内容等，例如信息文件载体格式及文件标题、文件号、日期等。通过知识挖掘和知识发现,为用户提供新的知识。最终在现网络信息的原貌，达到长期保存的目的以及满足利用者要求。

3.项目建设总体内容

该项目主要内容包括网络信息资源归档与利用的机制、体制的建立，网络信息资源采集平台研发，信息采集技术研究。网络信息资源归档管理平台研发，WEB 存档文件格式转换技术研究。网络信息资源垂直搜索引擎平台研发，垂直搜索引擎技术研究，网络信息资源回放系统研发，网页指纹技术研究，相关基础设施及硬件设施的配备等。

4.平台建设相关技术要求

（1）界面要求

系统界面设计的主要原则是“简单大方，统一实用，快捷灵活，提示清晰”。考虑到系统的使用者计算机操作水平参差不齐，界面设计应做到简洁、友好、符合用户使用习惯。

系统界面应具有智能化特点，应提供充分的即时、在线帮助，能在需要时提供下一步操作提示，方便用户使用。

（2）安全性要求

系统须结合安全体系的建设方案，提供足够的安全措施保障系统的安全性。

数据访问的安全：登录本系统的用户必须经过权限核查，只能访问本人权限范围内的数据。对系统数据的权限设定应达到字段级。此外，系统还应充分保证数据库不被非法访问，开发方须针对这一要求提出完整的安全性保证方案。

数据传输的安全：在数据转换或数据更新等环节，必须保证数据传输过程中的安全性。

数据库的安全：系统应提供应用级的数据库安全保障。

（3）系统开发要求

系统的开发必须按软件工程的实施规范进行，保证系统选型先进、操作方便、维护简单、实施规范，基本要求如下：

系统开发应采用模块化的开发方式，各模块之间相互独立，模块接口开放、明确，任何一个应用模块的损坏和更换

不能影响其他软件模块的应用。允许用户通过参数定义有选择地使用系统提供的应用模块。

系统模块设置要合理，保证不会由于功能模块的增加或改变而影响数据库的独立性和完整性。

系统应具备易用性和易维护性，使用图形化界面对各种信息进行访问；在修改、维护该系统时，简单方便、易于操作。

系统应具备可移植性，不受操作系统及其他参数的限制。

系统的程序开发必须规范化，要有统一的命名规范，包括模块名、变量名、函数名等的命名。程序要有良好的编码风格，代码布局要有统一的格式规范，程序中必须给出详尽的注解。

（4）运行环境描述

为了保障现有大量低端台式 PC 机的使用，系统须满足客户机端应能在最低办公用机（配置为 PIII，512M 内存）上运行。

操作系统： Windows XP、Windows 7 及更高版本。

（二）网络信息资源归档与利用平台整体构架

系统建设采用 B/S 构架，沿着档案归档管理工作收集、管理、利用的思路建立三层结构模式，整体划分为数据资源层、系统控制层、用户 UI 层。如图 9 所示：

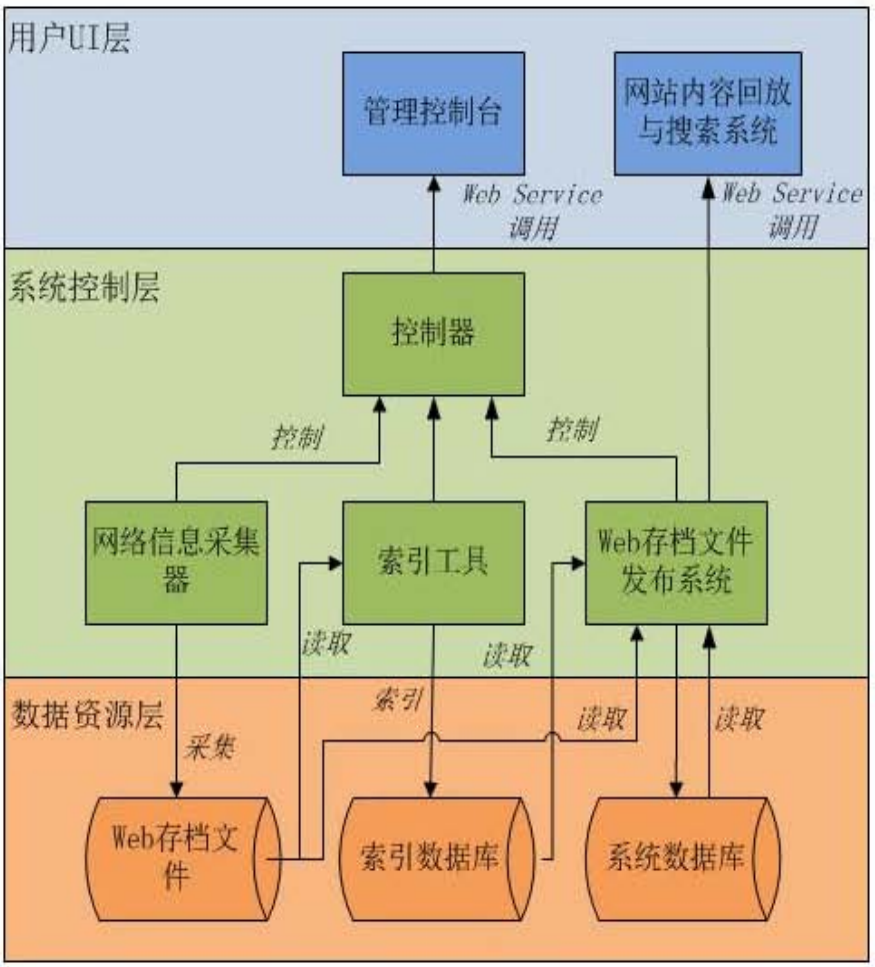


图 9　归档与利用平台平台整体构架

1. 数据资源层

该平台建设的核心部分，由 WEB 归档文件、索引数据库和系统数据库构成了网络信息资源归档管理平台，用于保存从网络采集来的各种 Web 信息资源文件及归档信息。

Web 归档文件用于保存从网络采集来的各种 Web 信息资源文件，如 HTML 网页、图片、视频等文件。Web 归档文件将采用 ISO 28500:2009 标准所规定的 WARC 格式。

索引数据库用于保存索引工具对 Web 归档文件中的 Web 文档所建立的索引数据。索引数据为快速检索到 Web 归档文

件中的文档提供数据基础。索引数据库由开源的搜索引擎Lucene所产生的.cfs索引文件构成。

系统数据库用于存储控制协调整个系统工作的系统数据，如所采集网站的基本信息、用户权限等。

2.系统控制层

系统控制层由维持整个系统运转的所有后台核心组件构成。它提供如网络信息资源抓取、建立Web文档索引，发布Web存档文件等核心功能。系统控制层不直接提供用户界面，但它提供Web Service接口，使其他系统可以通过调用Web Service接口中提供的函数，使用系统控制层的功能。

网络信息采集器又称网络爬虫，网络信息采集器负责将指定网站上的Web文档，采集到典藏系统，并保存于Web归档文件中（WARC文件）。网络信息采集器将以开源软件Heritrix作为基础，添加或修改特定功能，以适应网络信息资源典藏系统的特殊要求，例如处理URL地址中包含未编码的中文字符的问题、增量式Web文档采集的问题。

索引工具用于为存储在Web归档文件中Web文档，建立索引数据。索引数据储存于.cfs索引文件中。索引工具将利用开源软件Lucene提供的技术建立两种索引：URL索引和全文索引。利用URL索引，系统可以通过URL地址快速找到相关的Web文档存档。利用全文索引，系统可以通过关键词快速找到相关的Web文档存档。

Web 归档文件发布系统用于将存储在 Web 归档文件中 Web 文档发布。发布系统通过 Web Service 接口和 Http Service 接口，使其他系统可以获取发布后的 Web 文档。

控制器对网络信息采集器、索引工具和 Web 归档文件发布系统进行统一控制，协同它们工作。并提供 Web Service 接口，可以通过调用 Web Service 接口，控制、监督整个系统的运行状态。

3. 用户 UI 层

用户 UI 层即网络信息资源利用平台为管理员和普通用户提供了便捷的交互界面。

管理控制台为管理员提供了与系统交互的用户界面。管理员通过管理控制台，可以控制、监控整个系统的运行状态。例如：管理员通过管理控制台，管理要归档的网站、启动网络信息采集任务、建立 Web 文档索引，发布 Web 归档文件，控制用户访问权限等。

网站内容回放与搜索系统为普通用户提供了浏览查询 Web 文档存档的用户界面。用户可以用两种方式来浏览查询 Web 文档存档。一是以网站回放方式，即用户可以像浏览在线网站一样，输入网址，指定要浏览的存档时间点后，打开该网站的首页的 Web 存档，然后通过点击首页的链接，来浏览整个网站以前的 Web 文档。二是以搜索方式，即用户可以像使用百度这样的搜索引擎那样，通过关键词搜索，查找 Web 文

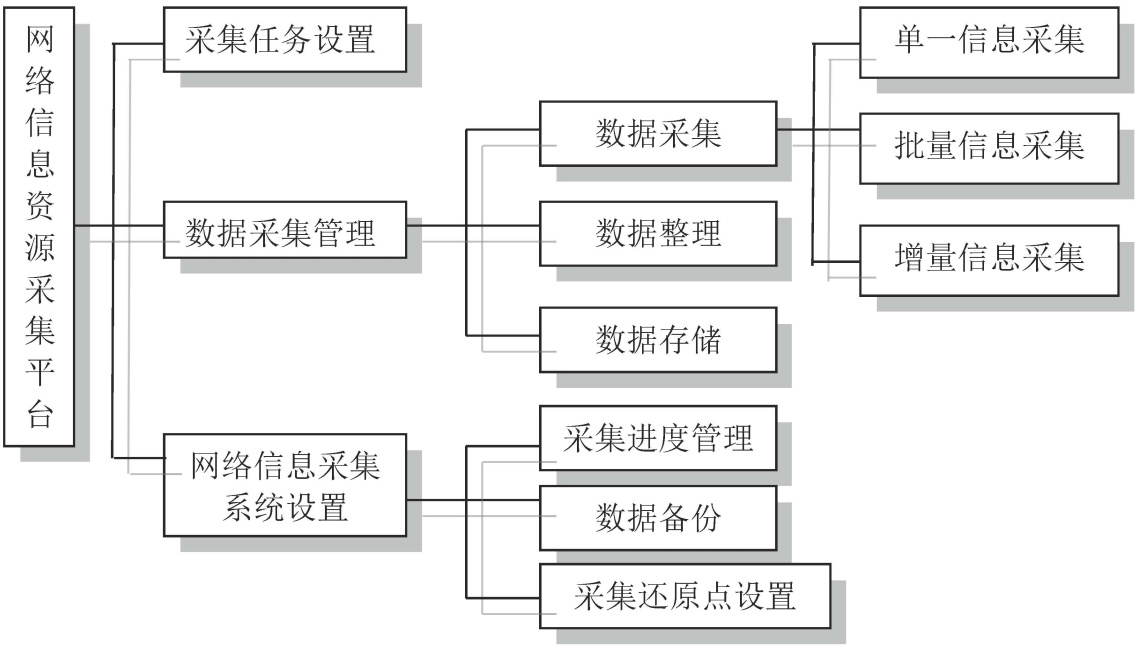
档存档。

（三）各应用系统设计方案及其主要功能

该项目整体框架涵盖三大应用子系统。

1.网络信息资源采集平台

（1）方案设计



（2）主要功能

该系统负责分析指定网络信息资源，指定相应的采集策略，将网络上发布的所有Web信息资源采集下来，利用分词索引技术将信息资源分词处理并保存为Web归档文件。主要功能分解如下：

采集任务设置：设定特定网站的网络信息资源的采集过程。包含任务名称、任务描述、种子列表、采集规则列表等。

数据采集管理：执行采集任务后将采集数据存储于WEB存档文件中。

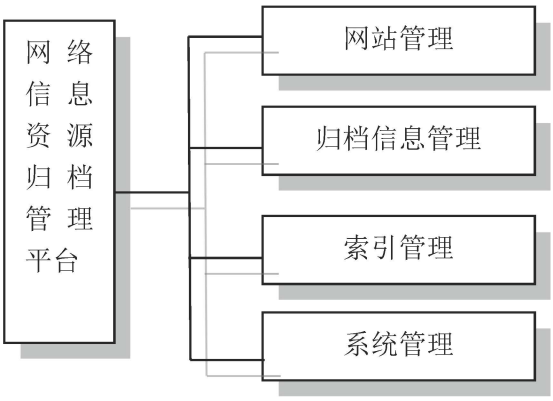
增量数据采集：通过对网络信息资源的新旧比对仅对目标网站上发生变化或新增文档进行采集。

采集进度管理：设置可查看采集完成进度、经历时间、完成时间，对任务进行暂停、恢复操作。

采集任务设置：根据工作需要对采集任务的执行时间段和周期进行设定。

2.网络信息资源归档管理平台

（1）方案设计



（2）主要功能

该系统设置归档管理功能，将采集完毕的网络信息资源著录诸如归档时间、密级等相关背景信息，形成基于档案管理的标准工作模式。

网站管理：对要归档的网站进行增加、删除、修改操作。

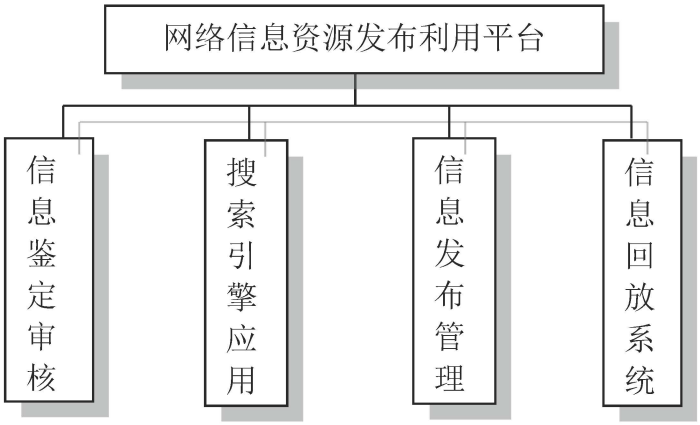
归档信息管理：对采集结果中的信息资源进行元数据信息抽取，按照标准完善归档著录项，最终打包形成 WEB 存档文件等待发布。

索引管理：通过启动索引工作，整理归档文件数据信息，为 WEB 存档文件建立索引。

系统管理：包括数据差错控制、日志管理、用户权限管理。

3. 网络信息资源发布利用平台

（1）方案设计



（2）主要功能

该系统包括网络信息资源搜索引擎平台及网络信息资源内容回放平台，为用户提供便捷的历史网络信息资源的在线全文检索和内容浏览。

信息鉴定、审核：按照相关鉴定标准对所采集信息进行信息内容鉴定、审核时候可以开放。

搜索引擎应用：通过输入任意关键词对所采集全部信息进行垂直搜索。

信息发布管理：经鉴定后的信息且通过审核的历史信息发布至互联网提供利用。

信息回放系统：展现指定网站在任意历史时间的原貌。

4.安全系统设计

从软件、硬件两方面入手搭建网络信息资源归档与利用平台的安全体系，保障该项目成果运行安全稳定。软件开发方面，采用代码加密机制、管理员及用户权限控制、日志管理等安全管理模块；硬件设备配备独立的的服务器，由本单位自行管理，安装部署专业的防护设备与备份设备，以应对当前互联网上病毒、木马、非法攻击等事件频繁发生，保证网上档案信息不受侵犯，保障终端计算机与服务器的安全。同时制定严格有效的安全管理工作机制，确保工作有序进行。

（四）网络信息资源归档平台系统部署模式

根据用户使用情况的不同，系统可以采用两种部署模式，即单一服务器部署和分布式部署。

1.单一服务器部署

单一服务器部署，即将整个系统的所有组件都部署到一台服务器上，如下图所示，这台服务器负责完成从网络信息采集、Web 存档文件索引和发布，到为普通用户提供网站内

容的回放与搜索服务的所有工作。这种部署模式适合需要采集的网站信息比较少的情况使用。比如只采集本单位的一个或几个网站，且网站的规模不是很大。

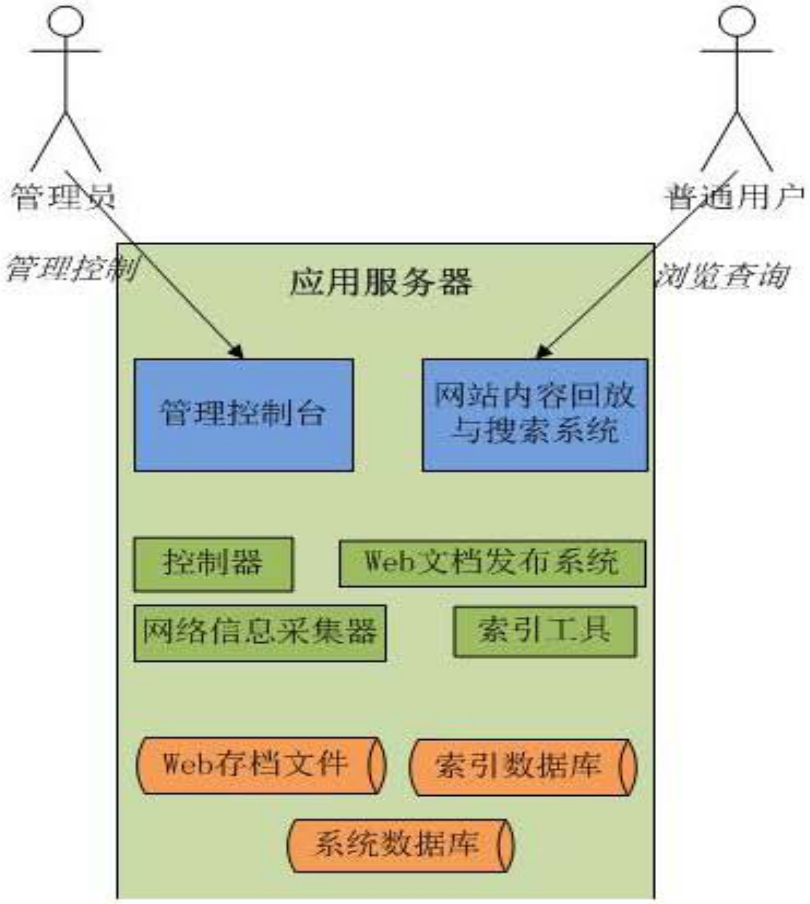


图 10　单一服务器部署示意图

2. 分布式部署

分布式部署即将数据资源层组件和系统控制层组件部署在若干个子服务器上。然后将管理控制台和网站内容回放与搜索系统部署到一个中心服务器上。中心服务器起到一个中转站的作用，它将管理员和普通用户发出的各种请求分别转发给相应的子服务器，子服务器处理请求，然后将处理结果返回给中心服务器，中心服务器再将处理结果返回给用

户。

比如，可以部署若干子服务器，每一台子服务器负责某一地方市县的档案馆网站信息的采集、索引和发布。子服务器可以在地理位置上位于地方市县的境内，以降低信息采集过程的网络流量。管理员可以登录部署在中心服务器上的管理控制台，控制监督每一台子服务器的运行情况。普通用户可以使用部署在中心服务器上网站内容回放和搜索系统，查看浏览被采集并存储到每一台子服务器上的 Web 文档存档。

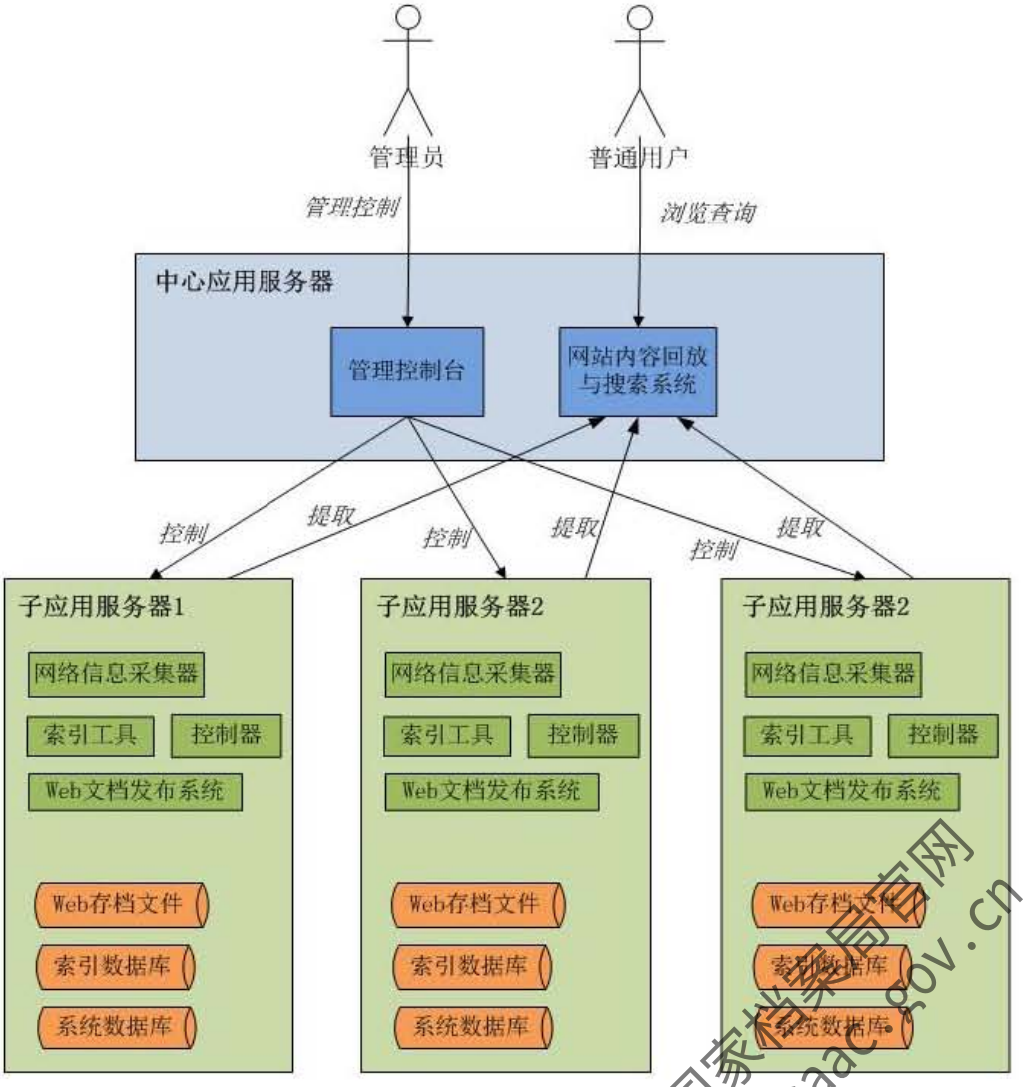


图 11 分布式部署示意图

（五）网络信息资源归档技术实现及其原理

1.开发语言：Java Flex

Java作为一个高层次的，面向对象语言，是一套丰富的，高质量的开源库。为系统平台建设提供了模块化设计和部件设置的支持，为逐步扩展和单独更换提供便利。

Flex 是一个高效的开源框架，可用于构建具有表现力的 Web 应用程序，它们利用 Adobe Flash Player 和 Adobe AIR 技术，运行时跨浏览器、桌面和操作系统，实现一致的部署。虽然只能使用 Flex 框架构建 Flex 应用程序，但 Adobe Flash Builder 软件可以通过智能编码、交互式遍历调试，以及可视设计用户界面布局等功能加快开发。

2.应用程序形式：Browser －Server 方式的Web 应用程序

B/S 结构（浏览器和服务器结构）是随着 Internet 技术的兴起，对C/S 结构的一种变化或者改进的结构。在这种结构下，用户工作界面是通过WWW 浏览器来实现，极少部分事务逻辑在前端(Browser)实现，主要事务逻辑在服务器端(Server)实现，形成所三层结构。是WEB 兴起后的一种网络结构模式，WEB 浏览器是客户端最主要的应用软件。这种模

式统一了客户端，将系统功能实现的核心部分集中到服务器上，简化了系统的开发、维护和使用。客户机上只要安装一个浏览器（Browser），如 Netscape Navigator 或 Internet Explorer 等，服务器安装 Oracle、Sybase、Informix 或 SQL Server 等数据库。浏览器通过 Web Server 同数据库进行数据交互，这样就大大简化了客户端电脑载荷，减轻了系统维护与升级的成本和工作量，降低了用户的总体成本(TCO)。

B/S 结构最大的优点就是可以在任何地方进行操作而不用安装任何专门的软件。只要有一台能上网的电脑就能使用，客户端零维护。系统的扩展性非常容易，只要能上网，再由系统管理员分配一个用户名和密码，就可以使用了。

3.应用程序框架与采用技术：Struts 框架 Ext-JS AJAX 技术

Struts 的优点主要集中体现在两个方面：Taglib 和页面导航。Taglib 是 Struts 的标记库，灵活运用，能大大提高开发效率。另外，就目前国内的 JSP 开发者而言，除了使用 JSP 自带的常用标记外,很少开发自己的标记,使用 Struts 是一个很好的起点。

ExtJS 是一种主要用于创建前端用户界面，是一个基本与后台技术无关的前端 ajax 框架。功能丰富，特别在界面设计及 ext 的表格控件上。因此，可以把 ExtJS 用在.Net、

Java、Php 等各种开发语言开发的应用中。Ext 是基于 Web 的富客户端框架，其完全是基于标准 W3C 技术构建设的，使用到的都是 HTML、CSS、DIV 等相关技术。Ext 最实用之处，是有一系列非常简单易用的控件及组件，只需要使用这些组件就能实现各种丰富多彩的 UI 的开发。

Ajax(即“Asynchronous JavaScript and XML”)是指一种创建交互式网页应用的网页开发技术。AJAX 不是一种新的编程语言，而是一种用于创建更好更快以及交互性更强的 Web 应用程序的技术。通过 AJAX， JavaScript 可使用 JavaScript 的 XMLHttpRequest 对象来直接与服务器进行通信。通过这个对象，JavaScript 可在不重载页面的情况与 Web 服务器交换数据。AJAX 在浏览器与 Web 服务器之间使用异步数据传输（HTTP 请求），这样就可使网页从服务器请求少量的信息，而不是整个页面。AJAX 可使因特网应用程序更小、更快、更友好。AJAX 是一种独立于 Web 服务器软件的浏览器技术。AJAX 基于下列 Web 标准：JavaScript XML HTML CSS 在 AJAX 中使用的 Web 标准已被良好定义，并被所有的主流浏览器都支持。AJAX 应用程序独立于浏览器和平台。Web 应用程序较桌面应用程序有诸多优势，它们能够涉及广大的用户，它们更易安装及维护，也更易开发。不过，因特网应用程序并不像传统的桌面应用程序那样完善且友好。通过 AJAX，因特网应用程序可以变

得更完善，更友好。

4.运行平台：J2EE 服务器，如 Tomcat ，WebLogic 服务器。

采用 J2EE 的三(N)层结构的特点主要有六点，一是能有效降低建设和维护成本，简化管理。二是适应大规模和复杂的应用需求。三是可适应不断的变化和新的业务需求。四是访问异构数据库。即应用服务器能够提供广泛的异构数据库访问和复制能力。五是能有效提高系统并发处理能力。六是能有效提高系统安全性。

数据库：My SQL Server 5.0。

客户端需求：无需安装任何软件，用户有浏览器即可登录使用系统。

（六）网络信息资源归档平台的软硬件环境

1.客户端配置

硬件配置：

内存：1G 及以上

CPU：1.6GHz 及以上

软件系统：

Windows 2000、Windows XP、Windows Vista、Windows 7 操作系统

Internet Explorer 7.0 及以上版本、Firefox 3.5 及以上版本 Web 浏览器

2.Web 服务器配置

硬件配置：

内存：4G 及以上

CPU： 双核 2GHz 及以上

硬盘：容量 400G 以上的磁盘阵列

软件配置：Windows Server 2003、Windows Server 2008、UNIX、LINUX 操作系统

J2EE 服务器，如 Tomcat 或 WebLogic

My SQL Server 5.0 服务器

（七）网上信息资源归档利用管理体系建设

网络信息资源归档规范化是实现网络信息资源科学管理的一项基础工作。为建立健全统一的网络信息资源管理体系，提高工作效率，实现工作目标，深入开发、利用档案信息资源，充分发挥网络信息资源在社会中的作用，结合我省实际情况，制定网络信息资源归档技术规范、网络信息资源归档与利用管理规范。

网络信息资源归档技术规范

一、适用范围

本规范规定网络信息资源归档所采用的术语、归档范围、采用元数据、归档方法等。

归档机构范围包括辽宁省各级各类档案局（馆）。其他机关、团体、企业事业单位，其他社会组织申请归档可参照本标准执行。

二、采用规范及标准

《中华人民共和国档案法》

《中华人民共和国著作权法》

《信息网络传播权保护条例》

《辽宁省档案条例》

《辽宁省电子文件归档与管理办法》

三、术语和定义

1.网络信息资源（Network information resources）

通过计算机网络可以利用的各种信息资源的总和。

2.网络信息资源采集（Network information resources collection）

利用网络采集技术，将某一网站的上发布 Web 文档下载，并保存到 Web 存档文件的过程。

3.Web 文档（WEB document）

网站上的各种信息资源文件，如静态 HTML 网页、动态 JSP、ASP 网页、MS Word 文档、XML 文件、图片文件、音视频文件等。

4.网站地图（XML sitemap ）

网站地图，又称站点地图，指网站页面上放置了本网站上所有页面的链接。

5.网络机器人（Network robot）

网络机器人（又称网页蜘蛛，网络爬虫等），是一种按照一定规则，自动抓取万维网信息的程序或脚本的技术。

6.网络信息资源元数据（Metadata）

用于描述网络信息资源背景、内容、结构及其整个管理进程的数据。

四、归档网络信息资源范围

1.有效满足人们目前或者未来对信息资源需求。

2. 面临网站管理机构变更、网站改版、更新、关闭。

3. 记录重大事件、受欢迎程度高。

4. 具有信息稀缺性、学术性、新颖性，以及信息可获得性。

5. 被大众认可的，与我国社会、政治、文化、宗教、科学或经济相关的网络信息资源。

五、归档办法

省档案局（馆）实施的网络信息资源归档项目，任何组织或个人都可以向省档案局（馆）提出网络信息资源归档申请（具体方法参见“申请办法”），省档案局（馆）将对网站内容进行审核，审核通过后进行归档。

省档案局（馆）根据申请单位或个人填写的申请表，确定网络信息资源的保管期限和是否公开提供利用。申请单位或个人在填写申请表时应注明信息资源是否允许公开、公开时间、信息密级等。对于不允许公开的信息，省档案局将不对外公开，资源所有单位或个人可以到省档案局查询利用或提取。允许公开的信息，省档案局将在互联网上提供查询利用。

对于由于技术原因无法采集的网页，要求归档单位或个人提供网页具体内容，或指导提供整改方案。除特殊情况外，不接收网上或用移动介质传递的归档内容。

六、归档时间

网站归档每年进行 3 次。如果面临预归档网站管理机构变更、网站改版、更新、关闭等特殊情况，归档操作也可应要求随时进行。

七、归档网络资源元数据定义

归档网络资源元数据，即网络信息资源版权、资源本身描述、资源硬件及软件环境等用于满足长期保存需要的元数据。网络信息资源归档所采用的元数据，申请单位或个人在提出申请通过审核后，需将信息资源的元数据信息及网站地图提交。

归档网络资源元数据表

类型	元数据	说明
说明内容描述类	题名(Title)	用于说明由创建者或出版者赋予资源的名称，例如网站名称、网页名称等。
	主题（Subject）	用于说明有关资源主题内容和学科内容的关键词、词组、短语或分类号
	描述(Deicription)	用于以文本形式说明资源的内容，例如文档、目录、版本说明、注释或视觉作品的内容等。
	来源 （ Source ）	资源的出处表示资源是从何处得来的。
	语种(Language)	用于说明资源内容所用的语种。
	关联(Relation)	该资源与其他相关资源的关系，该项目允许在相关资源描述间建立关联。
	覆盖范围(Coverage)	用于说明资源知识内容的时空特征。
	速度(Speed)	多媒体的播放速度或文本的动态显示速度。
	位置(Location)	所描述对象在网述中的具体位置。
	访问量(Visits)	用户访问该信息的流量。
	链接类型(Linktype)	该信息所属的链接类型。
知识产权描述类	创建者(Creator)	用于说明创建资源内容的主要责任者。
	出版者(Publisher)	用于说明负责使资源或为可取得和利用状态的责任者。
	其他责任者(Comrfbuor)	是指在创建者中未列出的对资源赋予的知识内容的创建作出主要贡献的个人、机构或团体。
	权限(Rights)	用于说明资源本身所具有的或该赋予的权限信息，一般包括知识产权等信息。
外部属性描述类	日期(Date)	用于说明当前资源的创作日期。
	有效期(Validity)	信息可以展示在网站上的时间。
	类型(Type)	资源本身的类型。
	格式(Format)	用于说明资源的格式，注明需要什么软件或硬件来显示和执行这一资源。
	标识符 (Demifiet)	唯一识别资源的字符串或字，如URL。

八、申请办法

预申请网络信息资源归档，请确保网站内容在归档范围内、并拥有网站版权。然后提供网络信息资源的详细信息，使用下面的表格。省档案局将根据归档标准，评估网站是否可以归档。

1.访问省档案局网站（www.lndangan.gov.cn），成为网站注册用户，填写申请表。

序号	填写项目	说明	填写内容
1	网站域名	对应于互联网中某台计算机数字地址（IP 地址）的字符标识，即计算机在网络空间中的地址和名称。	
2	单位名称	网站所有者单位名称	
3	申请人姓名	提出网络信息资源归档申请人姓名	
4	版权持有情况	信息资源的原创作者或享有信息资源印刷出版和销售的权利。	
5	网站建立目的	建立网站用于发布工作信息、公益服务、宣传推广、资源共享、信息交流等。	
6	申请归档原因	（参考归档资源范围）	
7	资源所占空间	资源大小，以比特率数表示。	
8	网站类型	政府网站、企业网站、商业网站、教育科研机构网站、个人网站、其它非盈利机构网站以及其它类型等。	
9	开发语言	Java、ASP、JSP、、PHP 等	
10	采用数据库	ACCESS、SQL SERVER、MYSQL、ORACLE、db2 等	
11	网站硬件环境	网站所在计算机的硬件配置，包括 cpu 频率、内存容量、硬盘容量等。	
12	网站初始硬件环境	网站运行所需的软件环境，包括操作系统（windows、linux 等）、网站发布平台（iis,tomcat 等）等	
13	网站变更历史	网站改版、变换域名、更改服务器等情况	
14	信息资源类型或格式	信息资源的表现形式，如视频、图片、文本等	
15	信息资源是否允许开放	信息资源访问限制	
16	申请日期		
17	联系方式	电话号码	
		通讯地址	
		E-mail	
		传真	

网络信息资源归档申请表

2.也可以访问省档案局网站（www.lndangan.gov.cn）下载打印“网络信息资源申请表”，填写完整，邮寄到省档案局相关部门。

地址：

传真：

电话：

E-mail:

如果归档周期、归档方法、归档内容等有特殊请求，可以致电省档案局进行协商。

3.辽宁省档案局（馆）网络信息资源归档项目免责声明

（1）本项目是公益性科研项目，旨在保存与保护网络中有价值的信息资源，为当今或后世提供历史网络信息服务。申请网络信息归档需仔细阅读该声明并签字确认。

（2）本项目所保存和公开的网络信息均系公共网络信息，其中的内容与观点不代表辽宁省档案局（馆）的观点与立场。

（3）本项目所保存和公开的信息是在申请人授权的情况下通过网络机器人自动获取的，如原作者、信息发布网站等著作权人或其他相关责任者不同意其内容或作品被本项目保存和公开，请直接和本项目组取得联系，本项目将遵照著作权人和其他相关责任者的意愿，对其拥有版权并被本项目所保存的信息进行删除。对期间发生的任何侵权、盗用等事件，省档案局不承担任何责任。

（4）本项目所保存和公开的内容和作品的版权归原作者所有，若作者有版权声明的，其版权归属以附带声明为准。若无专门声明的内容或作品以如下原则为准：

在用于非商业、非营利、非广告性目的时需注明作者及文章出处。

在用于商业、营利、广告性目的时需征得文章原作者同意，并注明作者姓名、授权范围及文章出处。

任何修改与部分删除均需保持作者文字原意并征求原作者同意，并注明授权范围。

本项目所公开的所有信息仅供参考使用。

（5）项目网站如因系统维护或升级而暂停服务时，将事先公告。若因非本馆

控制范围外的硬件故障或其他不可抗力而导致暂停服务，于暂停服务期间造成的一切不便与损失，辽宁省档案局（馆）不负任何责任。

（6）本项目使用者因为违反本声明的规定而触犯中华人民共和国法律的，一切后果自负，辽宁省档案局（馆）不承担任何责任。

（7）凡以任何方式登录本项目网站，直接或间接使用本项目网站资料者，视为自愿接受本声明约束。

（8）本声明未涉及的问题参见国家有关法律法规，当本声明与国家法律法规冲突时，以国家法律法规为准。

（9）本声明及其修改权、更新权及最终解释权属于辽宁省档案局（馆）。

申请人签字确认：

九、归档网站的技术要求和指导

由于技术原因不能归档网站，在网站建设和网站管理方面提出以下建议和指导。

1.确保所有信息资源信息网站根URL下，如网站中的图片、视频、Flash等资源存储在第三方网站上，将不能被正常被捕获和归档。

2.允许归档系统捕获到网站信息。归档平台采用“采集技术”整合网站信息，如果采用反采集（anti-robot）技术阻碍或减缓了归档传送，将不能被正常被捕获和归档。

3.网站应尽量避免使用动态生成当前日期的功能。日期应该使用服务器的日期，而不是客户端日期。

4.动态网站（即采用后台数据库支撑网站内容），在通过查询字符串能检索到的情况下，归档网站只能捕获到网页快照，而不能捕获到数据库。如果网页形成过程通过HTTP GET请求来完成。

例如www.lndangan.gov.cn/p.asp?id=5&d=true，当采集系统提出查询请求的时候，这个网页能够动态形成，同普通用户查询请求一样。但是采集系统不能通过HTTP POST请求检索到数据库驱动网页，因为没有查询字符串产生。使用POST参数在一些特定情况有用，例如搜索查询。但是网站管理者和供应商应该确信通过查询字符串URL，内容也是可以被存取的。需要网站管理者对通过直接链接不能正常显示的网页提供相应的xml sitemaps。

5.确保归档网站上需要归档的信息都是公开的，并能够被普通用户所访问和

利用。对需要注册才能访问的网站或网站上的信息栏目则可能不适用于网站归档平台。

6.使用有意义的网址。使用有意义、易于理解的网址利于增强归档网络信息资源归档的可用性、安全性和搜索引擎优先选择性等。

7.网站的管理员和开发人员应该使用简单而标准的网络技术。对于网站的建设者而言， world wide web consortium(WC3)-国际浏览器规范标准和目前的XHTML 1.0、CSS2对网站的开发有许多既定的规定。而使用过于复杂和不标准的网站设计则可能无法归档。

8.随着许多不能被“无障碍网站1.0版”兼容的格式的产生，客户端脚本应该基于最终目的，被恰当的使用在最适合的地方。对于网站的管理员和开发者而言，考虑到各种使用者的需求，制作开放式的脚本并且提供访问信息的多种可选方案。建立针对残障人士设计的无障碍网站，即任何环境，任何人及任何设备都能顺利浏览的网站。

9.管理员和开发人员应确保客户端脚本在互联网上是完全公开的、可见的。并且脚本文件尽可能的单独形成文件，例如用.js格式文件，而不直接在网页上加入脚本。

10.对于不支持或者禁用Javascript脚本的浏览器，网站设计者应该提供相应的技术方案，以及获取信息的其他办法。

11.网络采集软件不能存取使用诸如Javascript语言等脚本编写的，动态产生的网址。采集系统在读取那些由客户端脚本动态产生的网络内容时也会出现问题。这些可能会影响到那些以这种方式建设的网站的归档。(类似“上一页”、“下一页”、“第几页”，动态生成网址的网页)，建议代码写成易于存取的脚本文件，或使用<noscript>元素来确保在这种情况下内容可读，链接有效。

12.使用Javascript脚本设计的网站，其代码应尽量简单并且层次清晰。这种方法在网站的“layer”(层)设计中体现得尤为明显：

(1)代码语义，标准规格(XHTML)

(2)使用CSS添加展示层

(3)使用Java脚本增添交互

13.尽量不要使用繁杂的 Javascript 脚本设计网页资源连接，这种形式的脚本不利于对资源的归档、索引、搜索等操作。

例如：

<ahref="

javascript:_dopostback(‘ctl00$contentplaceholder1$gvsectionitems’,’page$1’)”>1

14.尽量使用简单链接网址设计方式，例如：<ahref=”content/page1.htm”onclick="

javascript:_dopostback(‘ctl00$contentplaceholder1$gvsectionitems’,’page$1’)”>1

15.被归档网站提供所有内容的网站地图(xml sitemap)，有助于采集系统在动态产生的网址中获取内容。如果动态的网址不能正确获取，采集系统还可以使用这些链接继续浏览。

16. 网站建设应让网站结构更透明并使网站内容更易获取，并提供给访问者多种途径访问。如果一个网站有大量的内容，把他们全部直接列出来是根本不可能的，应在网站内建立交互式导航，和通俗易懂的 HTML 网站地图。

17.目前的网站归档技术还不能归档交互搜索工具（如搜索引擎），或者网站其他的交互功能。这部分归档需专门建立 Sitemap。

18.如果归档网站有子网，部门管理者应提供相应的子网网站地图。

网络信息资源归档与利用管理规范

1.管理对象

管理的对象主要为归档的网络信息资源和与网络信息资源相关的元数据。

2.管理机构及职责

（1）信息资源管理机构

档案局（馆）设置网络信息资源管理领导机构。主要负责审定网络信息资源管理的规章、制度、办法，负责审核有网络信息资源归档资格等。

（2）指派专人按照网络信息资源管理的有关规定，负责信息资源的组织、

协调、服务等日常管理工作，主持信息平台、信息整合等系统建设、系统管理工作，负责有关技术支持工作。

3.管理内容及要求

（1）信息梳理

确定各申请单位的网络信息和需求，将申请信息进行分类梳理，制定相应的分类目录，明确采集信息内容、采集时间和采集的方式等。并由部门负责人签字确认。

信息归档审核确认后，信息采集部门负责采集，采集由信息管理部门统一归口管理。

（2）信息分类与分级

信息管理部门依据本规范和归档单位的协议及其它相关保密规定，对归档的网络信息资源实施分级，确定普通信息、秘密信息等，制定不同资源的存储介质和应履行的存储职责。网络信息类别暂时按照组织机构类型划分，最终形成各个网络信息归档资源的馆藏模式表，如表所示，形成信息密级目录列表，由部门负责人签字确认。

信息类别	保管期限	存储介质	密级	保密期限	存储份数
国家机关\企业\事业\社会团体\其他	临时保存级	光盘、磁带	公开\限制\秘密\机密\绝密		保存两份
国家机关\企业\事业\社会团体\其他	不定期保存级	磁盘、磁带	公开\限制\秘密\机密\绝密		保存两份
国家机关\企业\事业\社会团体\其他	长期保存	磁盘、磁带	公开\限制\秘密\机密\绝密		保存三份

网络信息资源资源馆藏模式表

4.信息安全管理

为保证网络信息资源的安全，分别施行相应安全控制策略，加强管理员权限、用户权限管理，合理信息资源授权，保障信息资源存储安全、传输安全。

（1）管理员权限管理

根据网络信息资源归档系统的不同功能，建立相应的管理员角色，包括系统

管理员、负责用户管理的管理员组、负责信息资源采集、目录管理的管理员组，以及相应的二级管理员组。用户可以维护自身的用户信息，只有被授权的用户才可以进行相应的操作。

对于网络信息资源密级为非公开的，只能采用针对单个管理员角色直接授权的方式。

（2）传输安全

根据信息资源不同的安全级别和对访问效率的不同要求，对信息资源的传输进行非加密传输、部分加密传输或者完全加密传输。

（3）存储安全

根据信息资源不同的安全级别和对访问效率的不同要求，对信息资源的存储进行非加密存储、部分加密存储或者完全加密存储。备份介质应妥善保管，特别重要的应异地异质存放，并定期检查，发现问题及时迁移。定期监控并登记软、硬件备份系统的变化情况，并形成监控报告。

5.档案网络信息资源利用规则

（1）信息管理部门汇总各部门信息密级目录列表，作为信息资源管理和信息访问控制的基础。

（2）信息利用部门负责建立外部公共信息平台，实施信息集中和信息整合，采用统一标准接入、存储、处理和发布各类信息。

（3）网络信息资源的集中和整合

信息集中和信息整合主要通过网络信息资源利用平台实现。信息平台包括四个层次：构建于网络之上的信息资源接入层、信息资源目录与共享交换层、信息资源应用层、信息资源门户展现层。

（4）提供网络信息利用部门负责已发布信息的维护，保证信息的时效性，适时更新。

（5）提供网络信息利用部门负有平台的维护职责，保证信息处理可靠及时，信息使用灵活方便，并提交月度、季度、年度使用和运维报告。

（6）信息质量反馈

网络信息资源发布部门针对信息利用情况，跟踪改进效果。如：对信息的准确性、时效性、完整性和一致性提出意见，及时汇总送达相关信息管理及提供部门，协助提高共享信息的质量。