请注明引用地址:Web Archive 存档策略分析*[J],现代图书情报技术,2009(1):16-21.

Web Archive 存档策略分析*

林　颖[1]　吴振新[2]　张智雄[2]
[1](北京师范大学数字图书馆技术研究中心　北京 100875)
[2](中国科学院国家科学图书馆　北京 100190)

【摘要】选择 Web Archive 有效存档研究中几种典型的存档策略:基于外部索引的压缩存档、基于多文件服务的存档、基于格式迁移的存档、基于特征抽取的存档,对它们的保存背景、策略应用以及实现特点进行分析,希望为我国在 Web Archive 存档研究的发展提供一些参考。
【关键词】Web Archive　网络信息保存　存档策略
【分类号】G250.76

An Analysis of Web Information Archiving Strategies

Lin Ying[1]　Wu Zhenxin[2]　Zhang Zhixiong[2]
[1](BNU Research Center for Digital Library Technology, Beijing 100875, China)
[2](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

【Abstract】Archiving strategy as the important research area of Web Archive, has been concerned by many projects, this paper selects several typical strategies: compressed archiving with external index, archiving with multi-master, format migration archiving, and characteristic described archiving , and analyzes their preservation context, characteristics , as well as the application of the strategy to achieve, and provides valuable reference for Web Archive research in China.
【Keywords】Web Archive　Digital preservation　Archiving strategy

随着网络应用的蓬勃发展,越来越多的信息资源开始直接依赖网络发布,适时保存这些 Web 资源并持续提供对这些资源的服务变得相当重要,于是 Web Archive(网络信息资源保存服务)的相关研究应运而生。随着研究的深入,越来越多的研究人员已经认识到 Web 资源不仅是目前更是将来长期保存的重点。从技术角度,可以将 Web Archive 分为三大部分:采集、存档以及访问,它们之间关联缺一不可,但是存档作为采集和访问的基础,是 Web Archive 首先必须解决的问题[1]。

Web 资源的存档不同于一般的数字资源存档,一方面 Web 资源动态性强、更新快,数量呈指数增长,所以有限的物理空间不仅需要存档更多的资源,更需要在将来能够快速安全地扩展存储空间。另一方面,Web Archive 的目标决定了其必须要能够支撑现在和将来的访问服务,所以必须考虑存档 Web 资源的稳定性、可访问性和长期使用性。另外,Web 资源类型丰富,简单的文本资源保存无法满足 Web Archive 的需求,存档对象包括软件、视频、音频、书籍等多媒体资料,可以看到日益丰富的资源类型也给存档带来了不小的挑战。

数据存储体系结构和存档策略是 Web Archive 有效存档研究的两个主要方面,数据存储体系结构的研究是通过选择合适的数据存储技术、存储媒介和部署策略,组成多种存储技术融汇的多级存储体系,较好的解决存储体系的大容量、高性能以及动态可扩展性的需求,PANDORA的SAN架构存储系统DOSS[2]、Internet Archive的集群

收稿日期:2008-09-19
收修改稿日期:2008-10-09
* 本文系国家社会科学基金项目"网络信息资源保存的理论与方法研究"(项目编号:06BTQ025)的研究成果之一。

存储系统 Petabox[3]、LC – SDSC 基于网格的协作共享存储框架 Chronopolis[4]，都进行了这方面的探索。而存储策略有着更广的研究领域，也是一个更有价值的研究内容，一个好的存储策略不但保障数据存档的可扩展和高效可用性，还可以保障数据内容的安全和可恢复性，即数据格式的永久有效性，因而成为 Web Archive 项目的重点研究内容。如非盈利项目 Internet Archive 是以压缩存档为重点、澳大利亚国家图书馆的 PANDORA 着重强调了多文件机制、北欧各国国家图书馆联合的 NWA 采取格式迁移的存档方式、美国国家档案馆的 ERA(Electronic Records Archive)则认为只有特征抽取才能彻底保证存档资源的永久有效性。各个项目在存档策略上侧重点各有不同，但不论它们采取何种策略都是为了实现对 Web 资源的长期保存和永久使用。

本文选择了几个典型的 Web Archive 项目，希望通过对这些项目的分析，了解这些项目 Web 资源保存的特色和规模、存档策略和实现特点，为我国在 Web Archive 存档方面的研究发展提供一些参考。

1　Internet Archive 基于外部索引的压缩存档

Internet Archive 成立于 1996 年，其初衷是为了帮助研究人员、历史学家、科学工作者实现对 Internet 历史资源的永久访问[5]。最初，IA 存档的资源以 Web 文本为主，从 1999 年开始存档对象开始涉及多媒体资源类型。目前 IA 的存档对象已经包括文本、音频、视频、软件等 Web 可见格式，并对公众开放访问。截止 2006 年，IA 已经保存了 3PB 的资源，并且还在以每月 20T 的速度增长。

IA 的 Web Archive 范围相当广泛，只要是 Internet 上的数字资源都是 IA 的存档目标或潜在存档目标，而且 IA 也积极引导公众将自己的资源纳入 IA 的存档，因此 IA 的一个工作重点就是实现海量数据的有效存储。在存储系统上，IA 自行设计了基于 Linux 的集群存储系统 Petabox，通过增加集群节点的方式扩充存储，IA 的每个节点都能够支持 TB 级的存储，由这些节点再构成 PB 级存储体系[6]。但仅仅实现海量存储是远远不够的，IA 更希望能够实现有限物理空间的最大化应用，于是 IA 提出了基于外部索引的压缩文件存档，其实现方式为：

(1)在外部数据库中保留对象的名称、存储对象的地址、存储对象在系统中的位移量、对象的大小；

(2)系统查找位移地址，读取与对象大小一致的字节获得相应的资源。IA 将这种存档格式定义为 ARC(或 WARC)，每 100M 的数据封装为一个数据包，在文件系统中进行保存。作为一种便于海量资源保存的压缩格式，ARC 具有以下特点[7]：

①封装独立，每个包都无需索引文件进行标识和解压缩；

②格式支持网络协议的传递，包括 http、ftp、news、gopher、mail；

③数据包支持多文档的数据流方式，即 1 个数据流可以组合多个存档文件；

④数据一旦写入即能保证有效，即数据完整性不依赖数据写入包后的内容索引。

在 IA 的每个 ARC 包中不仅封装了对象数据，还包括对象的元数据。ARC 提取的元数据主要是包含 URL 来源、创建时间等的描述性元数据。在 ARC 的数据封装中，多个 Web 站点的 Web 资源可以封装在一起，相同 Web 站点的资源也可以封装在不同的 ARC 数据包中。IA 系统结构如图 1 所示，所有 Web 资源均以 ARC 格式保存，并在数据库和 ARC 文件中对相关的资源建立索引。

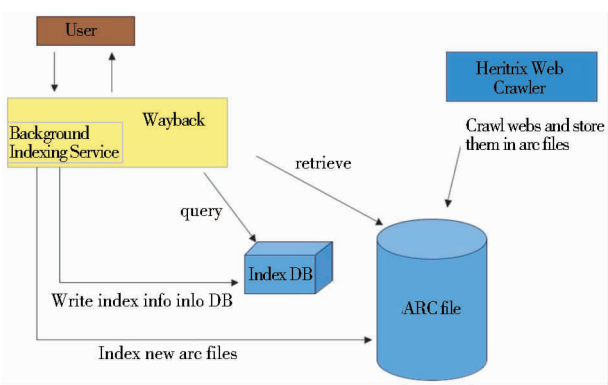

图 1　IA 系统结构[8]

这种基于外部索引的压缩存档策略的最大优势在于，既能对 Web 资源这类海量数据进行压缩存储，又对内容数据内外部同时建立索引，直接访问压缩资源。这种存档方式简单、直接，使得基于 ARC 的压缩存档成为当前应用范围最广的 Web Archive 存档模式，受到了大多数研究机构的亲睐，法国国家图书馆就在其网络资源长期保存的项目中应用了这种模式，截止 2006 年已经保存了超过 700TB 的数据。

但是这种方式对元数据描述不够详细,特别是缺少结构性元数据和技术性元数据的描述。尽管为了防止 Web 资源的格式退化,IA 通过收集相关格式软件的方式保证将来的数据重现,但这并不是一个一劳永逸的方法。

2　PANDORA 基于多文件服务的存档

PANDORA(Preserving and Accessing Networked Documentary Resources of Australia)项目是澳大利亚国家图书馆(National Library of Australia, NLA)对澳大利亚的在线出版物(包括电子期刊、机构站点、政府出版物)、有重要文化价值网站的长期保存计划[9]。作为一项由政府主导的旨在保存网络历史文化遗产的研究,该项目中尤其关注存档数据的可用性和长期性。

为了解决大规模数据保存和访问的效率冲突,PANDORA 将存档分为三个层次:

(1)持续工作所需的存档:主要是预存档数据;

(2)确保长期保存的存档:这是最主要的存档内容,主要包括长期保存资源、元数据等;

(3)提供访问的存档:主要用于访问的派生物。

同时,PANDORA 出于存档安全的考虑,同一份资源不能用于多种服务,因此在 PANDORA 的数字存档系统 PANDAS(PANDORA Digital Archiving System)中对数字对象进行了分类复合,数据模型如图 2 所示,其中:

(1)Web 对象经过数据检测后以压缩格式保存到 DOSS 存储系统中;

(2)同时再保存一份与源文件一致的资源,并且是一个未经压缩的备份,主要用于 Web 访问服务,其中 DOSS 是 PANDAS 基于 SAN 结构的底层存储系统[10]。

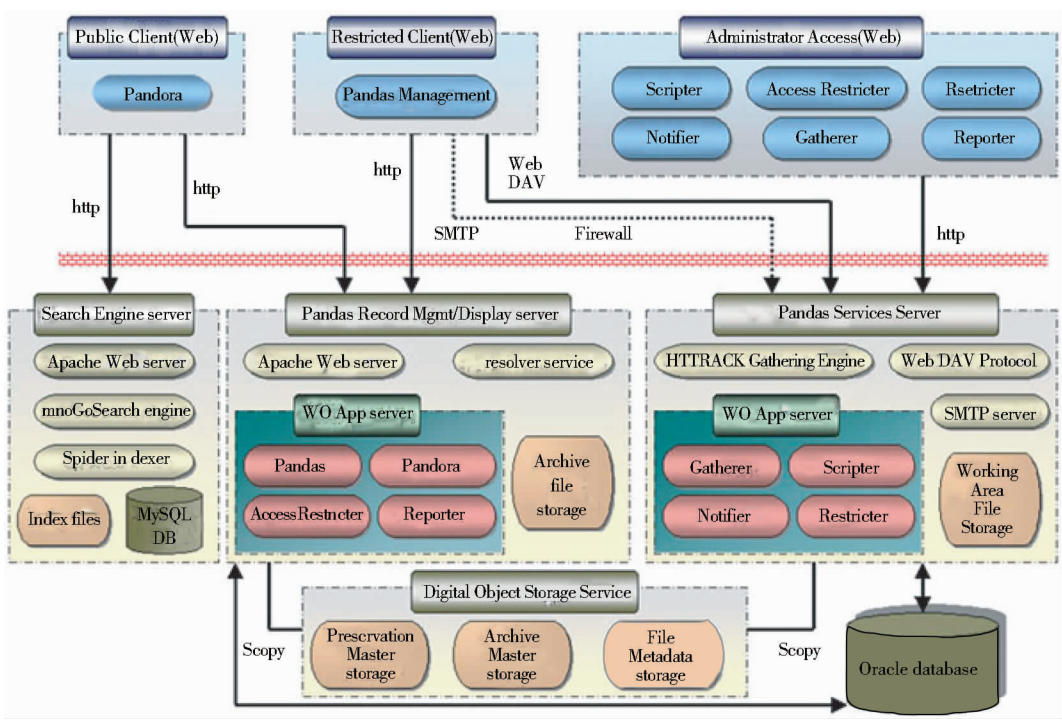

图 2　PANDORA 数据模型[10]

即 PANDORA 分别创建了三个不同用途的 AIP 包进行存档处理,此外还将采集资源建立了单独的备份,专门用于用户服务:

(1)保存主文件(Preservation Master),采集获得的一个未经任何改动的备份,以 tar 格式保存在 DOSS(Digital Object Stroage Service)存储系统中,见图 2 中 Preservation Master Storage;

(2)显示主文件(Display Master),经过数据检测处理的备份,同样以 tar 格式保存在 DOSS 存储系统中,见图 2 中 Archive Master Storage;

(3)元数据主文件(Metadata Master),依旧保留目录结构和源 Web 服务文件名,包括从每个文件中的

HTTP 应答的元数据,同样也以 tar 格式保存在 DOSS 存储系统中以描述性元数据为主,见图 2 中 File Metadata Storage;

(4)用于访问的显示备份,显示主文件的未压缩版本,用于公众访问的 Web 服务,见图 2 中 File Metadata Storage。

可以看出,PANDORA 采取的是一种基于多文件服务的存档策略,通过几个阶段不同备份的方式实现存档,并且采取了独立访问的资源方式将保存和使用分离,这不仅在一定程度上缓解了保存和访问的冲突也更好地保障了数据的可还原性。但是这种方式容易在存档过程中引发 AIP 间的混乱。此外这种方式也存在相当的数据冗余,会给存储带来不小的压力。

3　NWA 基于格式迁移的存档

NWA(Nordic Web Archive)是北欧 4 国(瑞典、芬兰、冰岛、挪威)合作的 Web 资源存档项目,主要是开展彼此间的资源共享和 Web Archive 技术共享,目前 NWA 的存储规模:挪威 22TB、瑞典 400GB、芬兰约 200GB、冰岛约 10 - 20GB 左右。

NWA 通过对 Web 资源格式的统计分析认为,超过 97% 的 Web 资源都是 HTML/JPEG/GIF 格式,所以该项目从保存角度认为,格式越集中,保存的难度就相对越小,因此 NWA 主张的存档策略是对不同的格式进行统一的处理,即数字对象的格式迁移[11]。

NWA 的存档迁移过程如图 3 所示,分别处理元数据和内容数据:

(1)首先对采集对象进行数据抽取,获得源 Web 对象的元数据和内容数据;

(2)判断内容数据是否为 HTML 格式,是否属于可转换格式,将符合格式转换的内容数据转换为 HTML 格式;

(3)从已经被转换的 HTML 格式文件中抽取相关资源再以 NWA 的文件格式进行保存。NWA 认为,保存 Web 资源首先需要实现内容的数据保存,至于原有表现形式则并不是 Web Archive 的重点[12]。

在元数据格式上,NWA 采用了 DC(Dublin Core)格式并支持 OAI 元数据收割。由于 NWA 采取的是基于格式迁移的存档策略,因此元数据基本不涉及系统性、结构性元数据等方面的内容,以描述性元数据为

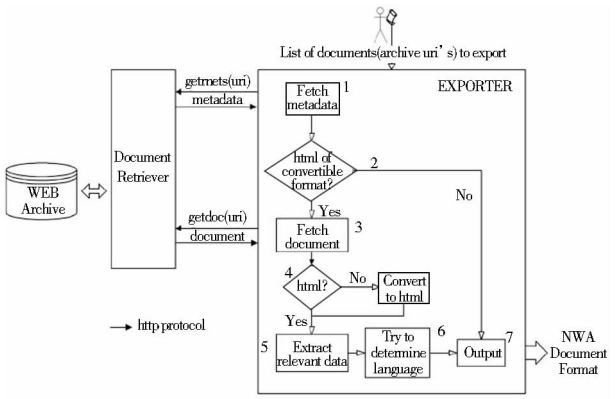

图 3　Web 资源的 NWA 文档格式迁移[12]

主,主要包括源链接、采集时间戳等元数据信息。

尽管这种基于格式迁移的存档策略易于管理和数据还原,但并不是所有资源都能够实现完整的格式迁移,在迁移过程中不可避免地存在着数据遗失的情况。另一方面,格式迁移也会在一定程度上影响数据的历史还原程度,特别是在表现形式上,如何协调资源内容在格式迁移前后的表现,将是格式迁移所面临的难题。

4　ERA 基于特征抽取的存档

ERA(Electronic Records Archive)是美国国家档案馆(National Archives and Records Administration, NARA)开展的一项对(政府)电子记录(以电子公文等为主)提供保存和访问的研究项目[13]。NARA 认为数字记录的存档,不仅是简单的对象保存,更要保留记录的基本特征,例如记录之间的内在关系,通过这种基于特征抽取的存档方式,存档资源就可以在将来的任何时间无需技术支持就能够有效地使用。即存档资源能够脱离对特定软硬件格式的依赖。因此,NARA 希望通过 ERA 系统为各种类型的数字对象建立一个全面的、系统的、动态的且独立于任何软件或硬件的存档方式。

因此在 ERA 项目中,NARA 重点应用了 PAM(Persistent Archives Method)模式,这种模式基于两种假定:

(1)可以详细描述每个对象的基本特征;

(2)每个对象都可以以各种不同的形式表现,并且总存在一种可以将其物理化的外在形式。作为一种面向对象的方式,PAM 的特点和重点在于,它强调记录对象的特征,而不是通过其他方式解决对象的技术退化。例如,文档在计算机上可以以 pdf 等格式传递,但是也可以打印在纸上。每个记录的不同表现形式在

记录创建时各有不同的用处。PAM 就是建立在这样一个概念上,通过增强对记录的基本属性的外在规范的要求,来保证该记录的可信赖性的[14]。

ERA 将内容(Content)、结构(Structure)、上下文(Context)以及外在形式(Representation)定义为一个对象必须具有的本质特征。它认为在记录这些本质特征的时候,不仅需要那些不可改变的特征,更需要详细描述每个特征的可变性,而不是随着时间丧失这些可变性特征。表 1 是 ERA 提出的关于 E - mail 邮件的特征示例,内容类型是通信,数据类型是 pst 格式文件,彼此间通过相应的基本属性(文本、字体等)关联[15]。

表 1　ERA 的 E - mail 邮件的特征示例

Record Type	Data Type	Essential Characteristics
Correspondence	MS Outlook .pst file	Appearance/ Layout - Text/Font - Structure ("To:" & "From:") - Association w/Attachments Behavioral - None

事实上,在 ERA 的存档策略下,存档对象的元数据就不再局限于描述性元数据,而更多地包含了相关结构性元数据等,它实现了对存档对象的完整保存元数据描述。这对当前的 Web Archive 影响深远,因为随着研究的深入,如何避免存档对象格式退化的问题日益突显,毕竟只有在格式可辨识的情况下才有可能能还原内容数据,而格式的解析则要受制于当前存档对象的软硬件环境,包括硬件设备、软件工具、数据模型、规范标准等。

特别是随着 Web 资源在不断发展,类型在不断丰富,存档对象的格式更新速度加剧,这不仅扩大了格式退化的范围,而且也加剧了格式退化的速度。格式退化对 Web Archive 而言是一个非常严峻的问题,一旦存档对象无法识别,那么 Web Archive 也将变得毫无意义。尽管不少研究机构采用格式保存等方式来保证将来对内容数据的还原性,但这毕竟只是延缓了格式退化的速度,并没有从根本上解决格式退化的问题,

所以 ERA 针对这种现状提出了基于特征抽取的存档策略,在保存内容数据的同时解析并保存数据的生存特征。一旦将来出现格式退化,即便已经完全失去内容数据的生存环境,那么还是可以根据已经保存的特征信息重新传递内容数据所依赖的软硬件环境等,然后才能基于这些技术信息模拟或重现存档对象的生存环境,包括适时的格式迁移。

尽管这种基于特征抽取的存档策略具有明显优势,它通过特征抽取实现了内容与格式分离,但是需要对每种格式进行详细的特征说明,因而这种存档模式复杂程度高、需要昂贵的人力物力的支持。

5　结　语

在上述分析基础上,笔者对这 4 种存档策略的主要特点进行了进一步总结,如表 2 所示:

表 2　4 种存档策略的分析

	基于外部索引的压缩存档(Internet Archive)	基于多文件服务的存档(PANDORA)	基于格式迁移的存档(NWA)	基于特征抽取的存档(ERA)
元数据	描述性元数据	描述性元数据	描述性元数据	描述性元数据 结构性元数据 技术性元数据
AIP	与元数据一起封装 ARC 或 WARC 格式	与元数据一起封装多个 AIP	与元数据一起封装 XML、HTML	与元数据一起封装结构化(4 个层次)
特点	压缩存储应用范围广	多备份区别应用	标准格式易于还原	独立性强 无关格式退化

(1)所有 Web Archive 项目都有一个共同的目标,即在海量存档的基础上确保采集的 Web 资源能够在将来得到长期有效的使用,因此 Web Archive 需要平衡保存和访问之间的冲突。PANDORA 认为不能将同一份资源用作多种服务,所以在该项目中对于访问服务重新创建了一个资源备份来使用。除此之外考虑到长期保存的需要,PANDORA 还分别对未经处理的采集资源和经过处理的采集资源分别建立备份,以确保能够在任何危机情况下真实还原数据。

(2)对于 Web 资源的海量特性,必然要合理安排存档的密度,因此 IA 提出了压缩存档。ARC 格式的最大特点在于既能够对数据进行压缩存储又能对数据建立内外索引从而直接访问压缩文件,这极大地提高了存储空间的利用率。NWA 项目则直接摈弃了无关数据,它认为保存的重点是网络上(以文本为主)的内容信息。

(3)Web 资源类型非常丰富多样,那么避免格式退化就是 Web Archive 所必须解决的难题。IA 一直在存档相关 Web 资源的格式软件,NWA 则将数据格式进行统一转换,但这都不是一劳永逸的技术方法,因此 ERA 在不断实践数字对象的特征抽取。ERA 认为,通过对本质特征的详细说明与记载,可以实现内容数据与格式无关。如果将其与 PREMIS 元数据体系进行参

照可以发现,所谓的特征实际就是数字对象的技术元数据和结构元数据,可以说 ERA 项目是 PREMIS 的实例化应用。

(4)不同的存档目标会影响 Web Archive 的存档策略。例如,NWA 不注重资源在表现形式上的还原,因此它选择标准格式对数据进行统一转换并存档。这样不仅有利于数据的组织与管理,同时还可以在最大程度上减少格式退化带来的影响。

不难看出,这些 Web Archive 项目因应用环境和需求不同,在存档模式上各具特色,这同时也表明在网络资源保存中可以采用多种策略合理有效地存档海量 Web 资源。深入分析这些策略对于开展 Web Archive 项目具有重要的参考意义。如何在各个项目研究的基础上取长补短,根据实际情况制定出既有利于存储容量动态扩展又能够保障优良性能的存档策略是 Web Archive 领域未来需要继续探索的内容。另外,还应该更多地从应用(或使用)角度关注海量数据存档策略对于高效访问的影响,同时还应关注存档策略与长期保存研究之间相辅相成的关系,从而最终实现 Web Archive 的目标:保存 Web 历史资源以便于今后的长期永久使用。

参考文献:

[1] Netpreserve. org (International Internet Preservation Consortium) [EB/OL]. [2008 - 07 - 09]. http://www. netpreserve. org.

[2] PANDORA(Australia's Web Archive) [EB/OL]. [2008 - 07 - 09]. http://pandora. nla. gov. au/about. html, http://www. nla. gov. au/nla/staffpaper/2004/koerbin2. html.

[3] Large Scale Data Repository: Petabox[EB/OL]. [2008 - 07 - 09]. http://www. archive. org/web/petabox. php.

[4] SDSC Chronopolis[EB/OL]. [2008 - 07 - 09]. http://chronopolis. sdsc. edu/.

[5] Internet Archive[EB/OL]. [2008 - 07 - 09]. http://www. archive. org/about/about. php.

[6] Large Scale Data Repository: Petabox[EB/OL]. [2008 - 07 - 09]. http://www. archive. org/web/petabox. php.

[7] Burner M, Kahle B. WWW Archive File Format Specification[EB/OL]. [2008 - 07 - 09]. http://pages. alexa. com/company/arcformat. html.

[8] Library of Congress. Data Center for Library of Congress Digital Holdings: A Pilot Project[R/OL]. [2008 - 07 - 09]. http://chronopolis. sdsc. edu/assets/docs/SDSC_LC_data - storage_report_2. pdf.

[9] PANDORA(Australia's Web Archive) [EB/OL]. [2008 - 07 - 09]. http://pandora. nla. gov. au/about. html.

[10] Koerbin P. The PANDORA Digital Archiving System (PANDAS): Managing Web Archiving in Australia: A Case Study[EB/OL]. [2008 - 07 - 09]. http://www. nla. gov. au/nla/staffpaper/2004/koerbin2. html.

[11] Nordic Web Archive Introduction[R/OL]. [2008 - 07 - 09]. http://www. lib. helsinki. fi/tietolinja/0100/nwa. pdf.

[12] Hallgrímsson P, Bang S. Nordic Web Archive. 3rd ECDL Workshop on Web Archives[R/OL]. [2008 - 07 - 09]. http://bibnum. bnf. fr/ecdl/2003/proceedings. php? f = hallgrimsson.

[13] Electronic Records Archives (ERA) [EB/OL]. [2008 - 07 - 09]. http://www. archives. gov/era/.

[14] National Archive and Records Administration. Electronic Records Archives Program Management Office. Electronic Records Archives: Introduction to Preservation and Access Levels Concepts [R/OL]. [2008 - 07 - 09]. http://www. archives. gov/era/pdf/preservation - and - access - levels. pdf.

[15] Lake D. SAA Presentation[R/OL]. [2008 - 07 - 09]. http://www. archives. gov/era/pdf/2006 - saa - lake. pdf.

(作者 E - mail:wuzx@ mail. las. ac. cn)