●刘　兰[1,2]，吴振新[1]
(1. 中国科学院　国家科学图书馆，北京　100190；2. 中国科学院　研究生院，北京　100049)

Web Archive信息采集流程及关键问题研究*

摘　要：通过对国际网络存档项目和系统的调研，把网络信息采集的基本流程归纳为选择、征求所有者许可、实施采集、抽取元数据、质量审核和网络存档等6个部分，并对采集流程中存在的关键问题进行识别和分析。

关键词：互联网；网络存档；信息采集；采集流程

Abstract: By investigating and studying some typical Web archive projects and systems in the world, this paper summarizes the basic process of Web information acquisition into 6 parts: selecting, soliciting for permission, information acquisition, metadata extracting, quality auditing and Web archiving. The key problems in the information acquisition process are also identified and analyzed.

Keywords: Internet; Web archive; information acquisition; acquisition process

网络信息资源保存通过持续采集网络资源达到对不断变化的网络进行保存的目的，网络信息采集既是网络存档(Web Archive)的起点，也是网络存档的基础，只有对网络信息进行了成功的采集，才能实现有意义的长期保存。因此，作为长期保存关键性的一步，网络信息采集决定着网络信息长期保存的最终效果。但网络信息自身的特点使得网络信息采集成为一项复杂的系统任务，需要一系列的流程来保障采集工作的有序进行，以实现高质量的网络采集。

本文在调研国际网络归档项目和系统的基础上，对网络信息采集流程进行了梳理，并对各个流程需要解决的关键问题进行了识别和分析。

1　网络信息的采集流程

1.1　网络信息保存典型案例系统的采集流程

网络信息长期保存最具代表性的项目是澳大利亚国家图书馆的PANDORA，其采集系统PANDAS的工作流程包括：

1) 识别、选择和登记备选主题。
2) 征求并记录存档许可。
3) 设置采集制度。
4) 采集。
5) 实施质量控制检查。
6) 开始归档过程。
7) 为归档的资源组织用于发现、访问、呈现的相关元数据[1]。

在IIPC的支持下，由新西兰国家图书馆和大英图书馆共同开发的选择性网络采集过程管理工具Web Curator Tool (WCT)实施采集的基本流程是：①获取采集授权，并在许可记录中进行记录；②创建一个对象以定义要采集的资料、采集技术参数和采集进度表；③审批对象；④根据进度表创建对象实例，运行采集，通知采集状态，为审核做准备；⑤审核对象实例，审批采集结果；⑥向外部数字存储库提交采集结果[2]。

由美国国会图书馆资助加利福尼亚数字图书馆进行的Web at Risk项目中，开发了一套用于采集、保存政府信息的基于Web的网络保存服务工具Web Archiving Service (WAS)，其工作流程主要包括：选择、采集、元数据、呈现及访问、维护、长期保存等6个环节[3]。

1.2　网络信息保存系统的基本采集流程

通过以上几个主要采集项目的系统和对其他国家网络信息长期保存项目的调研，笔者把网络采集的一般流程归纳为：①选择；②征求出版者许可；③采集；④元数据；⑤质量审核；⑥网络归档六大基本步骤。当然，具体采集项目对这些流程的顺序可能会有所调整，如对于元数据，可能会在采集之前进行也可能在采集之后，而有些项目也会省略个别采集流程，如那些受网络信息法定呈交法保护的图书馆在网络采集时就不需要征求出版者的许可等，这些都要视具体情况而定。在实际的采集过程中需要参照基

* 本文为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”资助的研究成果之一，项目编号：06BTQ025。

本流程来考虑和处理，以便能够确保网络信息采集的有序进行和成功实施。

2 网络信息采集基本流程中的关键问题研究

在网络信息采集的基本流程中每一步都会涉及一些关键性的问题，只有对各个步骤的关键问题进行识别并妥善处理才能保障网络信息的有效采集。

2.1 选择

由于网络资源数量大、增长快、动态性强、采用技术复杂多样，且不断发展变化，涉及知识产权、隐私等法律问题，因此，即使是完整性采集也不可能覆盖全部的网络资源，也面临着采集对象的选择问题，如完整性采集的最典型案例 Internet Archive选择了可公开获取的网络资源进行采集。

选择采集对象是网络采集的第一步，也是制约采集能否达到归档目的的最关键一步。由于网络资源不像传统出版物有统一目录，所以无法了解网络资源的整体情况，因此，选择采集对象就变得比较复杂，主要涉及怎么选和选什么两个关键问题，即选择方式和选择对象。

2.1.1 选择方式 选择方式主要由采集策略决定，不同的采集策略决定了不同的选择方式，如完整性采集策略，在选择采集对象时主要依赖于自动化的工具，如瑞典的 Kulturarw3项目使用 Whois软件鉴定本国的站点，自动化地选择 .se国家域的网站，以及服务器在瑞典的 .com的网站[4]。而对于选择性的采集，主要根据制定的采集指南或由来自相关领域的专家组成网络采集选择委员会来确定要采集的对象，人工方式占有较大比例，通常还会借助传统信息源来选择采集对象，如澳大利亚的 PANDORA项目根据网络采集指南，由图书馆员来选择要采集的对象。实际上，通常采用自动化和人工方式相结合，只是不同的采集策略对自动化和人工方式所依赖的程度不同，即完整性采集多是以自动化方式为主，人工选择为辅，而选择性采集则是以人工选择为主，自动化方式为辅，目前综合利用两种方式选择采集对象的混合型采集较为流行。

2.1.2 选择对象 选择对象主要从对象的位置、内容主题、格式、版本、网站边界和采集深度等几个方面来界定。涉及的主要问题有种子 (Seedlist) 站点的选择、网站边界的选择、内嵌数字对象的选择、采集深度的选择、数字对象格式的选择和对 Robots. txt文件的处理等。

1) 种子站点的选择。种子站点即采集的起始网站。由于网络资源不同于传统印刷型资料，没有统一目录，种子站点的选择相对比较困难，不同采集策略种子站点的选择是不同的。完整性采集，通常是首先对采集域加以界定，然后通过各种方法对采集域加以描述和表示。以国家域采集为例，目前可以利用国家顶级域名、IP所在物理地址、网站语言、Whois数据库提供的网站注册者信息等来确定，法国国家域采集就是综合运用域名是 .fr的、语言是法语的和服务器物理位置在法国的 3种方法[5]。选择性采集，通常是由图书馆员或网络存档员根据选择标准，利用多个搜索引擎进行搜索和过滤，从而选择一定数量的种子站点 (通常是几百个)，并在采集过程中不断补充，如关于“9·11”的相关网站、与总统竞选有关的网站的选择等。

2) 网站边界的选择。由于网络资源自身的链接特性使得网站并不是一个个的孤岛，而是相互间彼此链接的，站点之间存在纵横交错、错综复杂的联系，因此网站边界难以确定，在选择阶段要确定网站的外部链接如何抓取。完整性采集在于尽可能地穷尽采集范围内的网络资源，外部链接作为重要的信息源是重要的采集对象，因此，只要是采集域内的外部链接都进行采集，如 Internet Archive认为外部链接是发现新的信息源的主要途径，因此只要外部链接是可以公开获取的网站就实施采集。选择性采集出于采集成本和版权方面的考虑，通常只采集那些符合选择标准的外部链接，如澳大利亚 PANDORA项目，在采集前会征求所有采集对象所有者的许可，而对外部链接，如果不在采集范围内，或没有征得出版者的许可，将不予采集。

3) 内嵌数字对象的选择。内嵌数字对象通常是图片或多媒体文件，是网站内容的重要组成部分，因此为了保存网站的完整性和真实性 (Look and Feel)，有些项目把这些内嵌的对象也包含在采集范围内。但也有些项目考虑到知识产权问题，对未经许可的内嵌文档和图片等资源不予采集。

4) 采集深度的选择。对于网站抓取深度的确定，目前还存在很大的困难。对于一个新的种子网站我们无法确定其层深，需要采集几层才能保持其相对的完整性，而且各网站层深差别很大。有研究人员采用探测的方法，在一定时间范围内进行采集，或向内采集一定数量的链接或文件，或一定规模的资源等，然后查看采集效果，随后调整参数进行不断的测试，但效果并不理想。目前还没有理想的方法对网站抓取深度进行准确的定义，只能在实践中根据具体情况进行不断的尝试。

5) 数字对象文件类型和格式的选择。由于电子资源格式的多样性和不断发展变化性，以及出于长期保存的考虑，有必要对用于长期保存的网络信息资源的格式进行限制和选择。通过对采集数字对象文件类型和格式的统计分析，丹麦研究人员发现采集对象中，15种主要的文件格式占所有收藏的 95%，HTML，JPEG，PDF，AV等格式排在前面；澳大利亚的 PANDORA对采集的存档库进行统

计，也得出了相似的结果。因此选择主要的文件格式进行采集，基本能满足存档的需要，从而降低采集成本。当然，也有很多国家出于保存网络资源原貌的考虑，采集所有的文件类型和格式，然后再考虑对那些稀有格式文件的处理和保存问题。

6) 对 Robot. txt文件的处理选择。对于受 Robot. txt文件保护的站点如何采集，不同的国家策略不同。瑞典 Kulturarw3项目、奥地利 Austrian On-line Archive (AOLA) 项目[6]都选择了尊重 Robot. txt规则，不采集受其保护的网站 (页)，但也有国家如挪威认为尽管遵循 Robot. txt原则是基本的网络礼仪，但对于国家图书馆，其责任是保存网络文档，因此应忽略 Robot. txt原则。

2.2 征求出版者许可

尽管很多人认为网络属于公共领域，但由于网络资源本身也受知识产权保护，所以如果没有明确的关于网络信息资源的存缴法，即使是国家图书馆，在网络信息采集时也需要征求出版者的许可或采取相应的处理措施。而即使存在明确的存缴法，法定的存缴机构也需依法进行网络信息采集和保存，如新西兰国家图书馆按照网络资源存缴法采集新西兰国内的所有公开网站。

目前关于征求出版者许可有 3种处理方案："Opt-in"，"Opt-out"和混合型方法[7]。

"Opt-in"是指在采集前征求出版者的许可，典型案例是澳大利亚的 PANDORA项目，由于没有存缴法的保护，PANDORA项目规定在采集前都要征得所有采集对象所有者的许可。

"Opt-out"是指在采集前不征求出版者的许可，但在网站的显著位置提出声明，那些不希望被采集的网站所有者可以提出申请，要求从存档中删除已经采集的网站，同时以后也不再对其进行采集，或者是网站所有者利用"Robots. txt"文件保护那些不愿被采集的网站 (页)，如 Internet Archive, 这种方案主要是考虑到征求出版者的许可需要太多的人力和太长的周期，不适合自动化的完整性采集。

混合型方案是综合利用这两种版权处理策略，如对在线出版物采用"Opt-in"方式、对网页采用"Opt-in"方式 (很多网站或网页的寿命短，如果必须在采集前征求出版者的许可，有可能在获取许可后其内容已经不复存在)，既能较好地处理版权问题，又能缩短采集周期。因此，在实际采集中要根据数字对象的不同特征选择不同的征求许可方式。

2.3 网络采集

在实际网络采集的过程中，最关键的两个问题是采集工具的确定和配置。

2.3.1 采集工具的确定 网络采集过程中，首先需要确定采集工具。由于 Web Archive需要持续地采集网络资源，需要较强的版本管理、变化检测等有利于积累性资源长期保存的诸多功能，尽管目前存在很多网络信息爬行工具，但无法满足网络存档的要求，不能直接用于网络存档的信息采集，需要进行修改或重新开发。

1) 相关采集工具和系统。随着网络信息资源长期保存项目的深入开展，目前已开发出一些专门用于 Web Archive的网络爬虫，同时有些项目对通用的网络爬虫进行了修改以满足网络存档的要求，如美国的 Internet Archive项目开发了专门的爬虫 Heritrix; 北欧 NEDLIB项目研发了专门网络存档采集器 NEDLIB; 瑞典的 Kulturarw3项目使用了修改后的 COMBINE Robot等。还有些国家和机构利用一些基本的采集工具开发出了专门的网络资源存档采集系统，如澳大利亚国家图书馆的 PANDORA项目开发出的用于选择性采集网络出版物的数字化存档系统 PANDAS; 新西兰国家图书馆和大英图书馆在 IIPC的支持下共同开发的选择性网络采集的过程管理工具 WCT; 丹麦的 netarchive. dk项目开发了网络存档软件包 NetarchiveSuite; 加利福尼亚数字图书馆 Web At Risk项目开发了一套基于 Web的网络保存服务工具 WAS等。目前主要网络存档项目使用的采集工具情况如表 1所示。

表 1 主要项目使用的采集工具情况

项 目	采集工具
PANDORA	HTTrack
Minerva Prototype	HTTrack
Internet Archive	Heritrix
中国国家图书馆 WICP	Wget
Kulturarw3	COMBINE Robot
Nordic Webarchive	Combine/Nedlib
Domain. UK/UKWAC	Bluesquirrel Web Wacker/HTTrack
Web at Risk	Heritrix
AOLA	Nedlib/Combine

2) 采集工具的选择与确定。无论是选择已有的爬虫工具还是自行开发新工具，都有必要对现有的采集工具进行详细的调研和分析，找出已有的采集工具，尤其是开源工具，比较分析各自的优缺点，选择满足采集需求的工具，或者是借鉴各种已有采集工具优点，开发出新的采集工具。开发新的采集工具，要考虑开发的必要性、自身的技术实力、经济成本、开发周期、可行性等问题。因此，通常尽可能选择已有的 (开源) 软件，这样可以集中财力物力用于实施网络采集和存档。即使开发新的采集工具，也尽可能在开源软件的基础上进行优化和扩展，同时尽量寻求国际合作，减少投入的人力、财力，分摊技术开发风险，加强国际合作，优势互补，缩短开发周期，回避

开发的技术调整等，如目前 IIPC成员们合作致力于 Smart Crawler的开发。

2.3.2 采集工具的设置 采集工具的设置是采集策略与选择的具体实现，决定着采集的效果。其中主要包括网络边界设定、采集频率设定、采集调度、网络礼仪设置等。

1) 网络边界设置。网络区别于传统信息媒体的一个最大特点就是其超链接的特性，但对于网络采集，超链接的特性却使得网络边界的确定非常困难。目前的采集工具和采集系统提供了多种网络边界确定的方案。Heritrix中通过设置采集范围 CrawlScope来限制网络边界，提供了抓取范围不受限制的 BroadScope; 在采集域内抓取的 DomainScope; 在当前 Host范围内抓取的 HostScope; 抓取限制在主机上某些路径的 PathScope多种选择。WAS通过设置针对某个站点的最长爬行时间; 最大采集文档规模或最大采集链接数来确定采集边界，超过限度就自动停止采集。

2) 采集频率的设置。比较理想的方式是能够随着网站页面的更新立即实施采集。但由于网站（页面）更新的识别和监察比较困难，网络采集需要一定的时间，而网站变化随时都在发生，页面的更新没有规律，因此，无法设定完全按需进行的采集频率。目前完整性采集，由于资源海量，采集耗时长，其采集频率比较低，一般是每年几次，如丹麦每年 4次对国家域进行完整性采集。而选择性采集，其采集频率相对要高一些，会根据采集对象的性质来决定采集频率，如澳大利亚的 PANDORA项目对网络连续出版物，根据其出版周期确定采集频率，进行周期性的采集，而对于专著进行一次性的采集; 日本的 WARP项目把采集对象分为网站和连续出版物两类，对网站每月采集一次，对连续出版物，按出版物的出版频率采集。而基于事件的采集，如总统竞选、“9·11”事件等，由于采集的对象网站更新频繁，因此每天都对其进行采集。采集工具和系统通常都会提供采集频率的设置选项，如每天一次、每周一次、每月一次或每年一次等不同选择，NetarchiveSuite甚至提供了每小时一次的选择。采集者要根据采集策略和采集对象的特点，设置合适的采集频率。

3) 采集调度。网络采集的调度包括设定采集的具体起始时间、批处理采集的优先顺序等。采集的起始时间一般考虑网站本身的特点及其稳定性，通常选择晚上或周末开始采集，这样能减轻采集对网站的影响，同时这个时段网站自身的变化更新比较少，可以减少网站采集的误差率。WAS提供“每天一次”、“每周抓取一次”和“每月抓取一次”3种频率选择，对于每周一次的抓取安排在周五的午夜开始排队，整个周末进行抓取，对于每月一次的抓取安排每个月的 15号开始排队采集。WCT为每一个目标添加一个 Scheduler以确定这个目标的采集频率和采集时间，每个目标可以添加多个 Scheduler。正常情况下采集是可以在设定的时间里准时开始的，但由于带宽限制可能会出现网络拥堵现象，当有太多的网站在排队等待时，抓取可能不会按照预定的精确时间进行，这就需要确定采集的优先顺序，设置不同的优先级，如 WCT中当等候开始的 Target Instance数超过收割代理数时，Scheduler就按照优先级来决定先收割哪个（通常是先进先出），WCT提供了高、中、低 3种优先级别，通过对不同级别的任务分配不同的带宽来调度采集的优先顺序，如可以为重要任务设置 75%的带宽。为了更快更好地完成网络采集，WAS还提供了批处理的功能，可以一次同时开始多个抓取任务，最多可以添加 100个网站创建批量抓取，选择的这些网站将同时排队等待抓取，但仍是单独抓取它们。

4) 网络礼仪。在网络采集过程中，既要保证快速高效的采集网络资源，还要确保采集过程尽可能地对被采集网站造成较小的影响，因此在采集过程中还要注意网络礼仪，设置合适的带宽、下载速度和爬行间隔，如 Heritrix提供每秒钟使用的总带宽和每个主机每秒钟使用最大带宽两个设置选项，并可限定两次访问同一主机的不同 URL之间的时间间隔; WAS在实验测试的早期，为爬虫设置最高的礼仪，即爬虫在两次访问同一个服务器之间保持足够的间隔。遵从 robots.txt文件也是基本的爬虫礼仪，很多爬虫都有关于 robots.txt文件的设置选项，Heritrix提供了遵守、完全忽略等 5种处理选项，用户可以根据自己的决策，进行相应的选择。

2.4 元数据

为便于对网络存档的管理和利用，需要对采集到的资源进行元数据编目，包括元数据创建方式、元数据标准、应用的内容规则、元数据应用层次和元数据存储位置等问题。

2.4.1 元数据抽取方式 大多数的项目是在网络资源存档后对采集到的资源进行元数据编目，但采集阶段也需要及时抽取一些与采集相关的元数据，如采集时间、从采集对象网页头标（Header）字段中抽取的网页标题、上次修改时间等相关元数据。这些简单的元数据可以利用自动化的工具抽取，但如果要进行更加详细的描述，就需要专门的元数据抽取工具甚至图书馆编目人员的参与了，如澳大利亚的选择性项目 PANDORA是由图书馆员对采集到的网络出版物进行编目，把编目数据加入到国家图书馆书目中，供读者检索使用。目前由于完整性采集的资源体量巨大，基本不进行人工编目; 专题采集中，投入的人工相对多一些。人工编目的元数据质量好但效率低、成本高，因此亟待开发高效的元数据抽取工具。

2.4.2 元数据标准及要素 目前比较常用的网络资源描述元数据标准是都柏林核心元数据集 DC，但由于网络资源长期保存项目一般由图书馆承担，很多图书馆沿用了图书馆使用的元数据标准，如 MARC，UNMARC，AACR2 等。元数据要素则需要根据具体项目的要求进行确定，限于篇幅原因，这里不再详述。

2.4.3 元数据描述层级和存储 IIPC发布的元数据描述框架中定义了 3种层次的元数据描述方案：对象层、站点层和收藏层。元数据的存储目前有两种方案：元数据与数字对象分开单独存储、元数据和数字对象封装在一起存储。

2.5 质量审核

统计表明，采集来的网络资源如果不进行质量审核和处理，40%的资源无法保证其功能 (PCWA)。因此，在网络存档过程中，必要的质量审核具有重要意义。但由于完整性采集资源的海量性，无法进行逐个资源的质量审核，因此，很多完整性采集过程中没有质量审核，Internet Archive中很多资源无法正常显示也是很正常的。但对于选择性采集，存档质量是非常重要的，质量审核可以在存档之前对采集的资源进行最后的把关，对失败的存档进行审查和重新采集，以确保存档资源的质量。

2.5.1 质量审核的目标和方法 质量审核主要从采集对象的完整性、功能性和冗余性几个方面来衡量，即是否想要采集的对象都采集到了，是否保持了原站点的功能特点，冗余垃圾信息是否被有效的过滤等。为了实现质量审核的目标，采集工具和系统提供了一系列的质量审核方法。WAS通过监督抓取过程、查看抓取结果、检查抓取日志和报告等几种方式对采集质量进行控制。WCT提供浏览工具 (可以把采集的资源与原网站进行比较，从而确定采集的效果，进行必要的修改)、收割历史查看工具 (通过新采集资源与历史采集资源的比较，让用户明确采集的状况、下载错误情况、数据量和其他统计信息) 和修正工具 (提供爬行期间收割来的所有资料的树形列表，可以浏览收割状态、下载的规模、要下载的资源总量、成功下载数、失败下载数等，并可删除已采集的冗余及垃圾信息) 等 3种专门工具。同时，WCT还可以采用其他方式进行质量审核，如通过日志文件视图查看目标的日志文件来辅助质量审核。还可以用第三方工具辅助质量审核，如 Firefox的网络开发工具条。

2.5.2 质量审核存在的问题与挑战 尽管有很多工具和系统辅助，但质量审核仍需要大量的人工参与，对失败的采集要分析原因，组织重新采集或者请求技术专家的帮助。如采集过程中的信息更新、版权保护问题造成的禁止采集以及需要特定技术才能完成而现在不具备或成本太高等，这些原因造成的采集失败、无法保持站点的“Look And Feel”目前还无法解决。另外由于所需人工干预较多，质量审核过程也是网络采集中比较耗费人力、财力的模块，目前只有澳大利亚的 PANDORA项目对每个采集对象都实施了完整的质量审核。

2.6 网络存档

通过质量审核的采集对象就可以进行归档，以供随后的长期保存和检索利用。关于网络存档涉及的问题很多也很复杂，如存档格式、存档介质等。

3 结束语

笔者在对国际网络信息长期保存的主要项目和系统调研的基础上总结出网络信息采集的基本流程，并对该流程中需要解决的关键问题进行了初步的识别和分析。但限于篇幅，各个问题的论述都比较简略，而实际上这些问题都是相当复杂的，很多问题目前都还没有完全解决。希望通过对这些关键问题的梳理，能为网络存档项目提供有益的参考，起到抛砖引玉的作用，从而引起对各个具体问题的深入探讨和分析。□

参考文献

[1] PANDORA digital archiving system [EB/OL]. [2008-04-07]. http://pandora.nla.gov.au/pandas.html

[2] The Web curator tool [EB/OL]. [2008-04-10]. http://webcurator.sourceforge.net/.

[3] Web archiving service: release 1 guide [EB/OL]. [2008-04-10]. https://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=181.

[4] The kulturarw3 project- the royal Swedish Web archiw3e-an example of “complete” collection of Web pages [EB/OL]. [2008-04-10]. http://www.ifla.org/IV/ifla66/papers/154-157e.htm.

[5] ABITEBOUL S, COBENA G, MASANES J, et al. A first experience in archiving the French Web [EB/OL]. [2008-05-06]. ftp://ftp.inria.fr/INRIA/Projects/verso/gemo/GemoReport-229.pdf.

[6] Austrian on-line archive [EB/OL]. [2008-04-10]. http://www.ifs.tuwien.ac.at/~aola/.

[7] Web at risk. Collection planning guidelines [EB/OL]. [2008-05-06]. http://wiki.cdlib.org/WebAtRisk/tiki-download_file.php?fileId=327.

作者简介：刘兰，女，1983年生。研究方向：信息资源管理。

吴振新，女，副研究员。

收稿日期：2009-03-11