广播电台、电视和网站视听资源馆藏: 连续性与析取

克劳德 姆索
国家视听资源协会（INA）
布尔素马恩，法国

中文译稿由广东省立中山图书馆何燕提供

Session:

148 — *Copyright law and legal deposit for audiovisual materials* — Audiovisual and multimedia with Law Libraries

文摘

自 1974 年创建以来，法国国家视听资源协会（INA）负责搜集和保存法国视听资料，并向用户提供利用。INA 组建的初意是为了满足境内某一公共广播档案机构的职业需求，后来很快发展为法国视听资源遗产的国家级存储仓库。事实上，INA 现已是世界最大的数字音像档案收藏地，馆藏的电视广播资料超过 400 万小时，可追溯到最早的广播节目，作为法定存储机构，每年新增的节目长达 80 万小时。在 21 世纪之初，网络为信息出版发布提供新的平台，广播电台和新媒体公司抓住数字化革命的机会，在网上发布音像信息，法定存储对象随之扩展到法国网站上。有趣的是，法国立法者认为法国国立图书馆（BnF）和 INA 需要分工合作。INA 被法定为音像媒体网站的国家存储仓库,并按需提供音像媒体服务。

2009 年 2 月，INA 开始搜集 8000 多个相关广播网站的信息。这部分内容有效补充了 INA 原有馆藏电台和电视台的广播节目，保证馆藏节目的连续性。同时，为处理和解决新媒体内容的规格标准问题，相继开发和应用了多种信息搜集、储存、访问的新技术体系。

INA：法国音像和网站视听资源遗产协会

自 1974 年依法创立以来，法国国家视听资源协会（INA）负责搜集和保存法国视听资料，并向用户提供利用。INA 组建的初意是为了满足境内某一公共广播存档案机构的职业需求，但随着法国公共广播垄断的衰落，它很快发展为法国视听资源遗产的国家级存储仓库。

事实上，在 20 世纪 80 年代后期，新兴私营机构可申请广播牌照，却没有相应的法例确保它们的节目和产品得以收集和保存，成为国家遗产。相关学术团体多方游说，极力说服当局将电台和电视台的广播内容列为时代证词和证物,为将来的学者所利用。鉴于 INA 搜集和保存视听资料的经验和合法性，1992 年 6 月 20 日法国经投票通过一项法令，指定 INA 为（法国）广播电台和电视内容的法定储存机构。大约 20 年后，协会成为全球最大的广播信息收藏地，馆藏的电视广播资料超过 400 万小时，可追溯到最早的广播节目，每年新增的节目长达 80 万小时,INA 每周 7 天 24 小时不间断地选取 100 个电视频道和 20 个广播电台的数字化记录。

很明显，数字化时代为 INA 日渐老化的馆藏开拓更广阔的发展前景。早在 1999 年，INA 启动一个庞大的数字化计划，对 83 万小时频危的模拟音像资源加以保护，将其转化为数字格式。到 2015 年，这些资源将被多次转换，完全数字化，保证其长期可用性和可获取性。部分馆藏（3 万小时）会就 IPR 问题进行协商，使其可在互联网上向普通公众开放；对于 INA 拥有产权的归档广播资源（超 1 百万小时）将对专业读者提供受限网络访问。所有资源（超 4 百万小时）均可实时查找、评论和分析，方便学习研究。

在音像传播和保存从模拟化向数字化转变的同时，互联网成为发布和获取多种类型视听资源极为重要的工具。来自电信行业的广播电台和新媒体公司显然抓住数字化技术（压缩数字文件，宽度接入）提供的机会，发布网络音像信息、发动跨平台战略，推动声像利用和视频消费的演变（如果说不上是革命的话）。

随着出版技术的快速发展，法国“法定储存条例”的适用面也随之延伸到网络上。有趣的是，法国立法者认为法国国立图书馆（BnF）和 INA 应分工合作，确保各自馆藏的一致性和连续性。INA 因此被指定为网络视听资源和网络平台上音像媒体按需服务的国家存储仓。该法规最近公开颁布，明确规定两机构共同及各自的任务。

馆藏的相关性和连续性是法律赋予 INA 新任务背后的重要概念，在技术上和可行性上,其实施的方法、工具或实践远有别于以往广播资料的搜集归档。本文尝试对网络信息---与传统广播媒体相联、互参和互补---的选择、采购、组织、访问、储存和保存等关键问题作一个总体概述。

选择

和其它大部分国家一样，法国政府机关 CSA 管理国内广播通信事务，主要是电台广播和电视台广播两种。CSA 主要职责是分配和监管广播通信频道，因受制于当局，频道数量有限。当一个新的广播频道出现前---通常不可能像网站那样在一天之内发生---会提前对外公布相关信息。如果符合 INA 的收藏标准，其开播过程往往比较顺利。传统广播媒体的运作与网站运作迥然不同，网站会突然出现或消失。法国域名注册办公室（AFNIC）定期提供以.fr 为后缀的域名服务器列表，但这仅涵盖 30%的法国网站，而且没有根据网站的业务活动或主题进行分类。

音像资料馆藏工作的第一步就是根据网站的大小和更新情况，选择和评估可入藏对象。由于网络的无界性和短暂性，确定馆藏范围和域名抓取范畴显得尤为重要。这是个漫长的过程，涉及到多种人为的判定和裁决。根据法律条例规定的客观标准，INA 的文献资料工作者轮流跟踪相关网站，指定抓取对象。

依据相关法令条例，采选不仅要以广播内容为基础，还要收集发布方的活动情况：与电台和电视台密切相关的主要话题（这类信息包括博客和爱好者网站等），按需视频点播，广播节目的非线性网络访问（允许回放，截拦和仅供网络播放的原始视频内容）等。

INA 的网络音像信息搜集项目始于 2009 年 2 月，已持续 两年多。最初的网站只有 3600 个，如今已扩展到 10000 个。值得注意的是，鉴于当前网络内容的发展，INA 重新考虑“法定储存”的综合性收藏方向。由于无法保证所有相关网站得到正确的辨析和选择，只能用“全力（best effort)”的方法,由特定的职业人士尽力监听互联网。目前 INA 已通过半自动语义抓取方法的测试，当无法用人工方式辨别网页是否符合 INA 馆藏范围时，可用该方法支持人工决策。我们在后面还会说明“全力（best effort)”方式在采购或网页信息爬抓时的应用。

收集

在模拟信息年代，INA 采用图书馆传统的采访技巧和方法,搜集物理介质的广播资源。随着数字技术的广泛应用，早在 2001 年，INA 就组织了大规模的收集行动，对 100 多个频道实行全天候不间断的收录，并直接将收录的广播内容存入其存储系统。

在某种意义上，网络内容搜集的技巧和体系沿用上述信息搜集方式和“深入式”采集手段。网络有别于电台和电视台，是一个非线性媒体，不能“流入”而需要特殊和专用的采集技巧。

直接从网络服务器采集内容是一种惯用手段，目标是尽量在网页上仿效各种人机互动，生成多种响应和内容下载请求，把远程网站的内容尽数收集下来。所有搜索引擎一般都采用这些被称为“收获(harvesting)”或“爬行（crawling)”的技巧，搜集和处理网络数据。应用这些技巧的工具被称为“爬虫”或“蜘蛛”。“蜘蛛”听起来古怪而老土，但可能是最精确描述该方法的词语。它描绘出自动化软件工具为搜集部分互联网内容而发现、跟踪和忽略过的各种路径和节点。这不仅仅是一种收集方式，还是一种严格、系统的技巧。从一个种子列表出发，跟随链接，按特定的爬抓参数，找到并下载网页内容。如果网页不经常变化，链接没有意外出现或消失，节点上没发现新的或更新的内容，这将会很简单。事实上，当变化发生时, 可用某种方式预测或获取变化通知。RSS 源可用作变化的预报，正如蛛网震动会给蜘蛛警报，表明某物正落入蛛网某处。但这种情况较少见，蜘蛛通常不得不在网上来回搜寻以发现新猎物。

“蜘蛛”的比喻现在体现出其局限性，蜘蛛编织和掌控丝网，然而爬虫对网络入侵的深入，会像寄生虫那样扩散到由千百万个蜘蛛共同编织的巨大丝网中。而且，自动化软件工具可落入最不经意的“网络陷阱”(主动设立与否)，正如蚂蚁陷进沙坑圈套里一样。因此我们最后直接使用“爬虫”这一昆虫隐喻来描述当前主流的网络信息发掘工具。

由于法定网络收藏范围相对集中，且不可能达到大型域爬工具所覆盖的网站数量，INA 开发了一个可伸缩的内部爬行系统,以更好适应网站的多样性(例如更新频率、深度和交互特征)。

该系统建立在一个双层架构的基础上，由一个主调度程序向大量爬虫发送命令，每个爬虫负责某个时段的某个网站，每台单机上通常有 500 到 1000 个这样的爬虫在运行。

显然，主调度程序（双层架构的最上层）处理调度事务和每个网站的配置。采用多抽样的策略，监听更新频率和节奏：根据频率特征（经常、每小时、每天、每周更新等等）以半自动

方式对网站进行分类，进而相应地调整网页内容爬抓的节奏和速率。由于外网页（根据有限的交互量或点击率来定义）的更新程度往往比内网页多，因此每个网站外网页的爬行量总会比内网页多。

一些网站采用联合提要，而非链接新内容的方式发布消息。这些提要源可用来启动特定的爬行，当某个网站发布文章或信息时，会抓取其中某一单页中的文章或内容，并根据更新和评论情况自动进行重访。

这种方式分别牵涉到调度策略（频率和适合各个网站特定要求的深度爬行参数）和爬行问题（网络陷阱回避、爬行规律、礼貌执法和储存）两个方面。它允许同时使用不同的爬虫。由于互联网是一个由大量不同技术系统和格式组成的虚拟丛林，爬虫在触动交互进程的时候，容易引起错误和遗漏。

这种专用的多爬虫方法目标是采用前面提到的“全力（best effort）”收集方式搜寻各类型网络内容，显著提高入藏（档）质量。同一个调度程序可连接多达 3 个不同的内部爬虫，每个爬虫专注于某一类特定任务：

PhagoSite（网站吞啮）是一个通用的爬虫，可应对大量网站而不需要过多的计算机资源。

Fantomas（方托马斯，音译）是一个比较特殊的爬虫，基于 Phantom-JS 网络工具包，采用与谷歌（Google）公司 Chrome 浏览器和苹果（Apple）公司 Safari 浏览器相同的核心功能。该爬虫可爬行大部分用时下潮流的 JavaScript 交互软件编制的２． 0 网站，没有严格的计算机资源限制。

Crocket（克罗克特，音译）基于火狐（Firefox）浏览器，可爬行结构复杂，内容丰富的网站（富媒体和富交互），但需要密集的计算机支持。

此外，由于视频内容（无论数量和大小）是音像档案的重要组成部分，需开发专用的爬行工具，下载 YouTube 或 Dailymotion 上 UGC 格式的视频或收集（网络）直播流内容。

这种定制的专用爬行系统自２００９年以来一直运行着，当前每年请求达６０亿。显然，工具需不断更新和升级以适应变幻莫测和逐渐成熟的互联网世界。

描述，元数据和访问

电影和广播档案与数字和网页档案一样，需要技术支持方可读取。这个要求对于以往的出版物是不存在的。书本上的内容可马上看到，只要你有阅读能力；但影片、磁带、磁盘或数字文档等信息内容储存的物理载体需要重新装配、计算、解析才能为人们利用。

由于广播流和互联网的数据量庞大，ＩＮＡ的文献资料员逐步调整工作任务，元数据提取和管理逐渐成为惯例。

为了与馆藏组织的传统方式保持一致，需依据特定的分类体系对网站进行著录和分类，缩短广播馆藏与网页馆藏间的差距。

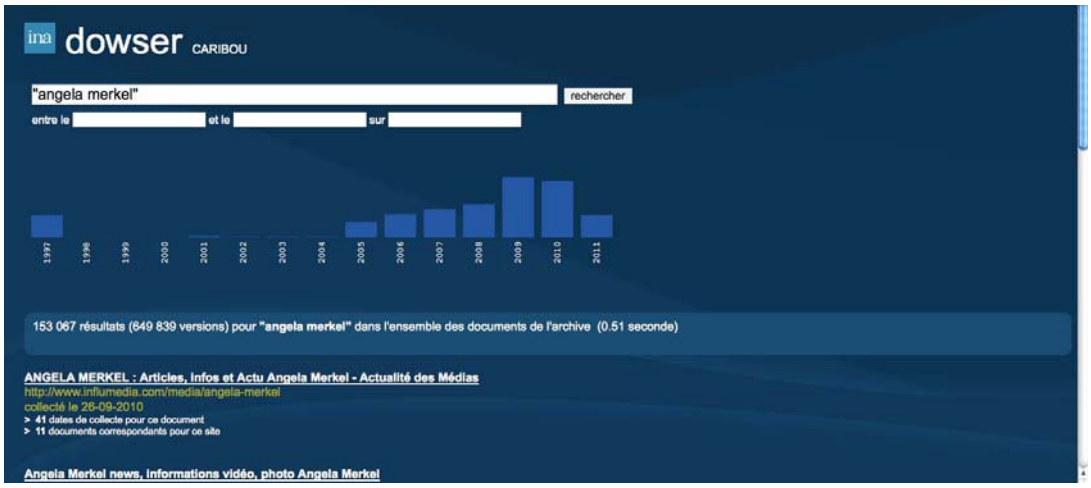
自动生成已归档（入藏）网站内容的索引，让用户可通过 URL（全球资源定位器）和日期组合，随意访问。存储在磁盘的文档和元数据形成档案资源。因为数字文档储存的离散性和网络作为发布工具的本质，访问已归档网页（最近估计，平均由 5 0 多个独立文档组成一个单网页）意味着重组发布过程，即访问某一特定时段爬行抓取的文档，并将这些文档在一个重构页面上显示出来。

档案浏览工具是一个定制的火狐（Firefox）浏览器。经过某次特定的抓取，通过及时的页面前进和后退，最终显示出所有可用网页。大部分人机交互功能仍存在（链接导航和基本交互活动仍体现着 95%的用户体验），但由于技术原因，某些交互功用已失效（交互的 Flash 内容，复杂的 JavaScript 交互，以及一些视频播放）。显然，档案网页的目标是抓取和恢复原始网页和内容的“外观和感觉”，但在很多情况下，由于技术障碍，这些尝试并没能完全成功。

例如，一些失效的交互功能需要以某种方式重建，确保完整的浏览体验。譬如：原网页内嵌播放器无法播放视频时，可用外置播放器代替；增加附加功能，如视频内部查找等。

由于无法将 Google(谷歌)或 Bing（必应）等搜索引擎归档（所有用户交互体验的仿真实际上是不可能的），INA 开发出一个特殊的搜索引擎，用以检索其网络档案，这是为处理档案时间限度和副本群集或副本近似而定制的。

Dowser 是我们神奇的搜索引擎，允许用户通过类似 Google 式的文本查询，访问任意存档内容。柱状图用来帮助研究者浏览给定查询条件内的日期和事件。



真实性是档案合法性的支柱，但这观念在数字化环境下受到极大的质疑。原始文献消失得快，数字数据注定是要被复制，转移或加工。网络内容档案可连续呈现网站的过往面貌，但无法让用户访问已不存在的网站。INA 的做法是，将原网站内容或环境的变化（网页间链接的暂时中断，外部视频播放代替原有播放，由于爬行限制出现的内容缺失等等）通知用户，强调网络档案文件并非是仿制品，而是不完全地重构原始载体。同样，“全力”方式依然是我们认为最好的重建及还原方式。

储存与保管

网络发布的信息内容绝大部分是数字化的，没有存储实体。倘如它收纳大量的时代历史纪录，是我们社会风貌的证物，那如何长期保存网络信息是至关重要的问题。同样重要的还有大型存储库的解决方案，保证数据在不停发展的磁带和硬盘存储技术之间，顺利转移，安全保存。

网络信息（档案）界意识到网络信息长久保存系统设计的重要性，网络信息长远保存战略仍在制订中，许多机构宁愿把注意力放在尽可能提供更广更多的可访问馆藏资源上，令馆藏建设保持活力。我们在后面会指出网络信息（档案）工作者讨论的问题，事实上这些均与数字信息相关。

假设可保持数据的完整性和数字信息的位阶，人们可否在未来正确地重新解析这些已存数据？正如必须维护电影胶片和视频带的可读性一样，文档转换格式以及浏览器或插件的维护等也同等重要。

长期字节保存（使用短期转移方式）

归档文件长期位保存是典型的数字信息工作流程，采用的是短期到中期的转移（约 2 0 年）策略。在 INA，两个档案副本分别存储在不同代的存储硬盘里，另有两个脱机备份副本存储在磁带上。这是基于当前文件归档的操作习惯：每 3 — 5 年转移到新磁盘，每 4 — 5 年转移到新磁带。

磁盘保存

当存贮对象从原来的格式网页文本（HTML，千字节大小），转变到数量不断增长，容量达百万字节的音像文档时，储存要求是磁盘保存最基本的问题。与磁带储存不同的是，磁盘储存通常附有较高的电源和冷却费用。考虑到磁盘技术的安全性和故障率，系统的正常运行期限一般在 3-5 年内。为加快转移和节约成本，每次转移都需关注存储空间的增大程度（例如，从 1.5 增长到 3TB 的磁盘空间）

LTO（Linear Tape-Open,开放线性磁带）

LTO 是 20 世纪 90 年代晚期研制，与当时盛行的专有磁磁带相反的，一种基于开放式标准的磁带数据保存技术。自 2000 年起，LTO 标准规定了胶卷盒标准尺寸，在大约 20 年间（2000 到 2017），磁带经历 8 代变化，容量不断增大。每代容量均以近两倍的比率扩大，推动转移战略变化和物理储存仓库的标准化。

尽管磁带理论上的储存期长达 25 年，但无法保证 LTO 厂商在未来还会生产可读取旧磁带的设备。事实上，LTO 规格只要求每一代的 LTO 必须可兼容读取前两代的产品（例如 LTO-4 代，可读取 LTO-2 代的磁带）。因此，基于安全转移的考虑，当一种新格式被广泛采用，经济性价比高时，需将数据转到新格式的磁带上。例如，虽然 LTO-5 驱动器在 2010 年第 2 季度已有供应，但只有到 2012 年第 2 季度，LTO-5 磁带的价格才优于 LTO-4 磁带。

类似地，做出转移决定后，需充分考虑经济价格因素，在两代 LTO 产品上最大限度地提高存贮效率和储存空间，譬如，为让容量提高四倍，可将 4 盒 LTO-4 转到 1 盒 LTO-6 中。

Checksums（校验）

自档案文件创立以后，可用不同工具校验其内容。如果发现文件内容有效，将对文件进行一次校验（Checksums），并存储在独立的数据库中。储存时，校验（Checksums）数据会连同其他相关的文档一起写进磁带。这样，可定期检验文档内容和有效性，如果需要，可取缔已损坏的后备副本。

长期档案保存：仿真或转移？

必须从两个不同的角度思考网络档案：

－网页的＂外观与感觉＂－－体现创建者设计风格、想象力和天赋的审美形式

－数据部件，反映作者思想的文字语言、图像、声音和视频

第一种形式似乎属于仿真保存策略范畴。目前许多组织机构相继成立各种项目，分类和模拟旧浏览器和插件；创建网页内容的录像或快照，检验仿真网络的正确运作。

第二种形式则归结于转移保存策略，其关键问题在于衡量获取（或几年内）的标准文件格式能否为中期（例如２０年）的数据转移所识别。

在上述提到的例子中，PDF 或 ASCII 是文本选用的格式，图像用 JPEG2000，声音用 AIFF，视频格式则用 MPEG4 H264。这些都是当前 INA 长期音像信息转移所采用的标准格式。这种方式有利于元数据提取，因为可利用多种软件工具识别文件和文件格式，并将结果用现行的标准格式保存下来。最后，网络信息保存就是尽量抓取网页内容。即使采用最好的转移和(或)仿真策略，仍无法完整恢复某一网页的原来状态。页面的视图可保存下来，但网页内的交互功能却不能用同样的方法还原。

上述策略的混合应用，会使未来用户用到多种软件工具。例如，他或她可能看到一个仿真器所创建的页面，但无法读取重建页面内的关联文件。然而，仿真工具通过对照网页快照的图像，映射出其视觉形式，然后转向元数据提取，从而还原原文档内容。

结语

我们无法真正地评估出数字储存对知识的长远影响，但档案馆、国家机构或学术机构都在努力，让知识可长期保存。考虑到网络信息生命短暂，格式繁多，变化多样，网页互动或用户体验维护复杂等因素，网络归档确实是知识保存的一大进步。目前，收藏网络信息已成为业界的共识，尽管不同机构有各自不同的解决途径和选择。ＩＩＰＣ内各单位的合作－－牵涉到多个国家的学术或职业用户－－促进彼此的对话和实践交流。万维网的历史和内容将被后人续写出新篇章。

2012 年 7 月 31 日