日本国立国会图书馆互联网资源存档研究与启示

杨云鹏
（国家图书馆，北京 100081）

摘要：日本国立国会图书馆从2002年开始进行互联网资源存档项目WARP，目前已经建立完善的体制。本文从网站筛选、采集技术、网站加工和保存技术4个方面对日本国立国会图书馆互联网资源存档项目进行详细介绍，并从采集方法、数据加工、保存方式、法规建设、国际交流与合作5个方面提出中国开展互联网资源存档的建议，以期互联网资源存档能得到更好的发展。

关键词：互联网资源存档；网站；日本国立国会图书馆；采集
中图分类号：G279　　DOI：10.3772/j.issn.1673-2286.2021.01.004
引文格式：杨云鹏. 日本国立国会图书馆互联网资源存档研究与启示[J]. 数字图书馆论坛，2021 (1)：24-31.

过去人们只能从书籍和文档中了解历史事件。但是，随着互联网和数字技术的发展，纸上留下的信息正迅速被网站等电子信息所取代。而当后代试图回顾历史时，如果网站上没有任何信息，那么大部分历史信息将丢失。为防止这种情况的发生，需要将网站上的信息保存下来。互联网的飞速发展，促使人们的生活、学习和工作逐渐离不开网络，而2020年新冠肺炎疫情的爆发，也加剧了通过网络获取信息这种形式的发展，然而网络资源的寿命一般在90~100天，因此互联网信息存档尤为迫切。互联网资源存档不仅可以保存人类在短期到中期内访问的互联网信息，而且对将来保留历史资料具有长远意义。

互联网资源存档主要由世界各地的国家图书馆和公共机构（世界各地的网络档案馆）负责，IA（Internet Archive）是已知最大的互联网存档内容保存机构，截至目前已拥有PB级别的压缩数据，并保存了3 300亿个网页和网页快照。它成立于1996年，是一个非营利性组织，其成立标志着网络信息资源保存研究的开始[1]。继IA之后很多国家陆续建立了互联网资源存档项目，其中采集规模较大的包括英国、法国和日本。法国和英国的互联网资源存档成分别成立于2002年和2004年，均保存本国域名的网站。日本国立国会图书馆自2002年以来一直在进行本国互联网资源存档项目（WARP）的研究，通过长期互联网资源存档开发了一套包括筛选、采集、组织、保存和发布在内的软件，让互联网资源存档变得更加容易和高效。中国互联网资源存档事业目前还处于初级阶段，亟需改进以跟上互联网的发展脚步，通过剖析其他国家或组织的互联网资源存档的技术和经验，对我国进行互联网资源存档，跟上世界步伐，实现互联网资源的长期保存具有重要意义，而现有研究主要围绕扩大采集范围和增加采集数量展开，缺乏采集技术、数据加工和长期保存方法等方面的研究。

因IA开放度不高且技术相对封闭，所以本文从互联网资源存档项目技术先进、开放程度高且与中国互联网资源存档发展路线一致的日本国立国会图书馆的互联网资源存档项目（WARP）出发，详细分析日本互联网资源存档的机制和支持互联网资源存档的技术，总结可以借鉴的技术和经验，以便更好地推动我国互联网资源存档的发展。

1　日本互联网资源存档项目概述

1.1　互联网资源存档的意义

互联网上的信息很容易更新、修订和删除，并且网站本身也会消失。近年来，政府机构发布报告之类的重要材料已经从纸质媒体转变为网络版本。而部分经过重大更新的网站时，总理办公室的网站将进行重大更新，更新后只保留过去的信息，并不会保留页面显示样式。此外，重大事件的网站也会随着事件的结束而消失。例如，2002年在日本和韩国举行的FIFA世界杯日本组委会的网站在比赛结束后就从互联网上消失了。日本互联网资源存档项目保存了日本政府机构和国家重大事件网站发布的所有内容，其涵盖文化、历史、政治和宗教等多个方面，未来能让更多的国民通过互联网资源存档项目了解国家的发展和变化，对整个日本历史文化的传承甚至人类文明的传承起到非常重要的作用。

1.2　日本互联网资源存档项目采集情况

自2002年以来，日本国立国会图书馆的WARP项目一直在保存即将消失的有价值的网站，如政府网站发布的信息、发布国家重大事件的网站及发布出版物的网站等[2]。互联网资源存档的作用是采集、存储并提供服务，以便让用户可以随时查找消失的网站。

1.2.1　日本互联网资源存档的数据量

日本互联网资源存档项目截至2020年3月已采集12 556个网站，177 154个网页，85亿个文件，数据量达1 678TB。2010年日本修订了《国立国会图书馆法》，允许全面采集网站，因此从2010年开始日本采集数据量快速增长，2010—2013年每年增加100TB左右，2014—2019年每年增加200TB左右。

1.2.2　日本互联网资源存档的数据类型

互联网资源存档项目包括多种文件类型，主要有jpg、png、tiff、pdf、html、php、css、js、xls、xlsx、doc、docx等，其中图片格式、html格式和PDF格式类占71.33%[3]。存档文件类型中图片格式占比最高，这是由于日本政府和大学的网站都是以图文并茂的形式呈现，是为了让更多的人能够快速理解文章的意思。日本政府类网站上公报、公文和政策类文件大多以PDF形式呈现，因此PDF类型的占比也相对较高。

1.2.3　日本互联网资源存档的方法

互联网资源存档从采集技术上可分为两种方法：一种是通过软件采集网站，保存网页内容的原始格式（jpg、pdf、html、php、css、js等）通过数据库进行管理服务；另一种是通过软件对网站进行采集，将采集的内容保存成WARC格式的压缩包，然后通过回访软件进行服务。第一种方法是网站原始格式，文件数量多、数据容易被修改，未经过压缩，占据存储空间大，不便于管理，因此国际上很少用这种方法进行互联网资源存档长期保存。日本国立国会图书馆是通过第二种方法进行互联网资源存档，这种方法是以WARC压缩包的形式保存，数据不能被修改，同时一个压缩包能保存多个文件，不但减少了文件数量而且减少了文件所占存储空间。

互联网资源存档从获取方式上也有两种方法：一种是通过软件采集进行保存，另一种是通过征集赠与或缴存的形式保存。征集赠与或缴存的网站是数据库形式的内容，需要转换成WARC格式。目前转换成WARC格式的技术还不成熟，转换后的网站回放的效果并不理想，会有一部分内容无法显示或出错，因此国际上主要以软件采集的方法进行保存，日本国立国会图书馆同样就是用软件采集方法进行保存。

1.3　日本互联网资源存档项目的特色服务

为更方便快捷地对存档内容进行检索及使用，日本国立国会图书馆对其存档的互联网资源进行了可视化操作和互联网出版物数据加工。①存储站点类别可视化：运用大数据可视化工具对存储的站点进行分类，用不同颜色的圆圈表示，资源容量越大，对应颜色的圆圈所占面积也越大。②公共团体网站可视化：以地图的形式分析采集公共团体网站，分析网站的变化和消失情况。③国家机构文件的可视化：从采集的国家机构文件中选取出1 000万个文件，以图表的形式展示其近5年出现和消失的情况。④互联网出版物的数据加工：从存储网站上提取出版物和受版权保护的作品，如白皮书、会议资料、报告和专著等，并添加标题和作者等数据，以

便可以对其进行有效搜索。

互联网资源存档项目主要保存国家重要文化财产，通过深度挖掘并利用大数据技术对其进行可视化操作，可为不同专业的科研人员提供丰富数据和图表供其研究使用，同时也能让更多的人明白互联网存档的价值和意义。

2 日本互联网资源存档全流程

日本互联网资源存档全流程如图1所示，由5个部分组成，即筛选、采集、组织、保存和发布。网站上发布的信息将随着时间而改变，互联网资源存档项目通过定期重复此流程来跟踪网站中的更改。

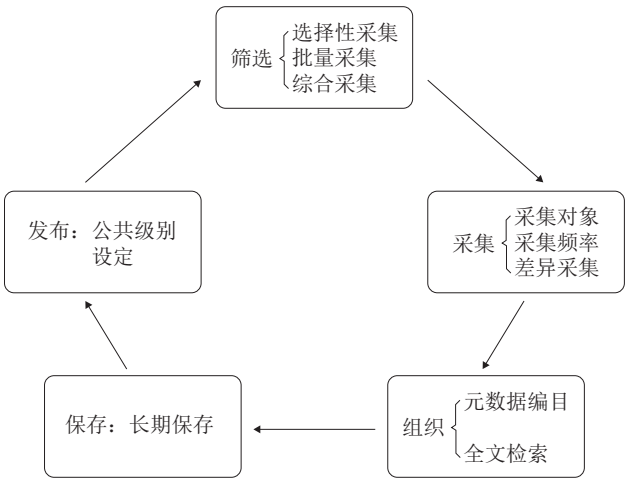

图1 日本互联网资源存档全流程

2.1 互联网资源存档网站筛选策略

根据制定的需求（包括目标类型和规模）筛选要采集的网站，以确定采用何种方式进行信息采集。其中根据目标类型需求按照专题采集特定类型的网站，一般采用选择性采集的方式。根据采集的规模，小规模采集仅采集国内的综合性网站，多采用选择性采集或批量采集的方式进行；大规模采集针对全世界范围采集网站，一般选择批量采集或综合采集的方式进行。

2.1.1 选择性采集

特定主题网站的集合称为选择性采集，需要指定一个采集单位，如站点单位或网页单位。此方法用于中小型互联网资源存档，如奥运会等类专题采集需要采用选择性采集，因为相关网站只有个别栏目是介绍这类专题的，没有必要完全整站采集。针对没有法律许可的网站，如版权声明中明确规定不允许复制保存的新闻类和受版权保护的文学类网站资源，采集部分内容前必须获得创建者的许可，其没有关于“批量采集”法律许可，故此种类型的网站采集也需选用选择性采集的方式。

2.1.2 批量采集

批量采集是跨国家/地区域（如“.com”和“.de”）的大规模网站集合。一些机构，如IA，会聚合世界各地的网站，因此日本国立国会图书馆在采集此类网站信息时，需采用批量采集的方式。

在法律制度下，大部分互联网资源是由国家图书馆等公共机构进行存档。批量采集法律许可的网站，无须事先获得创建者的同意。根据2010年4月生效的《国立国会图书馆法》（修订版），日本国立国会图书馆有权批量采集公共机构网站的资源。

2.1.3 综合采集

综合采集是将选择性采集和批量采集相结合的方式进行采集。日本国立国会图书馆通过立法可以对一部分网站进行批量采集，但是对于社交网站、视频网站和私人网站等并没有批量采集的权限，因此当采集需求涉及这类没有权限的网站时，只能采取选择性采集的方式进行采集。综合采集是采集特殊需求的内容，如发生的全国性热点事件既涉及官方网站内容又涉及社交网站内容，就需要运用综合采集，对法律允许采集的网站进行批量采集，不在法律规定范围内的网站须征得同意后才可进行选择性采集。

2.2 互联网资源存档采集技术

在实际采集目标网站时，日本国立国会图书馆使用自动采集程序——采集机器人（抓取工具）进行采集，在采集之前制定采集频率和采集深度。

2.2.1 采集对象

根据《国立国会图书馆法》第24条规定可以对以

下机构进行采集，如国家机关（立法、行政、司法，包括当地分支机构）、独立行政机关、国立大学法人（包括大学联合机构法人）、特殊法人等。第24-2条规定的机构包括地方公共组织（包括法定的委员会）和地方公社（港务局、房屋供应公司、道路公司、土地开发公司、地方独立行政机构、全国地方赛马协会、地方公共组织金融组织、日本下水道公司）等[4]。除法律规定外，WARP项目还会与网站创建者沟通，采集创建者允许的私人网站。

2.2.2 采集原理

日本WARP项目使用自动采集程序（Heritrix）自动采集网站。采集机器人采集网站原理如图2所示，采集机器人首先访问起始网页（起始URL）。然后，在采集页面html文件的同时，分析html文件中的结构并开始采集文件，包括文档、图像、音频、视频、样式表和脚本文件。从起始URL跳转到其他链接页面，然后重复相同的操作直至到达设定的采集深度或者没有链接为止。为了减少对采集网站服务器上的网络负载，每次采集之间将保留1秒或更长的下载间隔[5-6]。

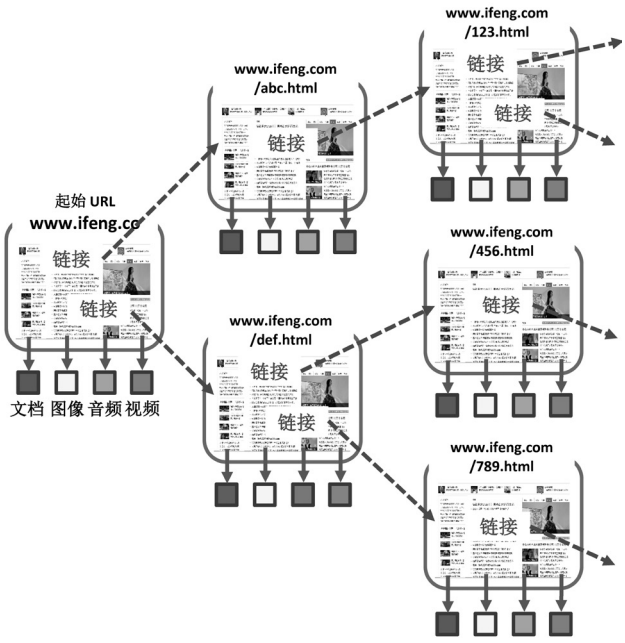

图2 采集机器人采集网站的原理

根据《国家国立图书馆法》第25-3条第2项的规定，对于设置了爬虫协议（robots.txt）的网站，要求网站必须将日本国立国会图书馆添加到爬虫协议中。

2.2.3 采集频率

最理想的采集频率是网站每次更新便采集，但这需要一种能实时监测网站更新信息的爬虫。一些大学研究机构正在开发配合大数据分析的高性能爬虫工具，但目前还没有互联网资源存档操作机构使用这种爬虫，因为这种爬虫对服务器的要求特别高，并不适合大批量采集，且还处于实验阶段。

日本国立国会图书馆根据不同网站设定不同的采集频率，如表1所示，基于法律规定尽可能保存“国家机关”的信息，因此采取国家机关每月采集一次，都道府县、政府条例制定部门城市等每季度采集一次，电子杂志根据发刊频率进行采集，重大事件网站根据需要采用选择性采集或综合采集的方式随时采集。

表1 网站保存采集频率和方式

采集机构	采集频率	采集方式
国家机关	每月	批量采集
都道府县	每季度	批量采集
政府条例制定部门城市	每季度	批量采集
直辖市	每季度	批量采集
独立行政机构等	每季度	批量采集
大学	每季度	批量采集
电子杂志	发刊频率	批量采集
重大事件	随时	选择性采集 或综合采集

2.2.4 差异采集

互联网资源存档会定期采集相同的网站。因此，部分新采集的文件相比之前采集的文件发生了变化，有一部分文件则与之前的完全相同，造成时间和存储空间的浪费。为了解决数据重复采集的问题，日本国立国会图书馆提出差异采集法。每次采集时保存所有文件的方法称为完全采集，而仅保存更改过的文件的方法称为差异采集。

在差异采集中，通过比较哈希值来判断文件是否相同。哈希值是通过使用某种计算方法（哈希函数）来处理电子数据而获得的值。由于不同电子数据的哈希值很少相同，因此可以将其比作电子数据中的指纹。电子数据的任何细微变化都会改变哈希值。

日本国立国会图书馆的差异采集是在开源软件DeDuplicator的基础上进行的二次开发。差异采集中，

首先分析网站结构，筛选出不易变化的文档类型，避免由于网页微小噪音导致哈希值变化进行错误采集。然后选择文本文档、非文本文档或者两者都进行过滤。最后对比以前的采集日志，如果文件名称均不相同，则进行保存；如果存在相同名称的文件，需比较通过SHA-1算法自动计算出的网页文档的哈希值，若相同则不保存，反之保存。

回放差异采集保存网站时，如果存在保存的文件，则显示该文件，如果当时没有文件，则显示最近保存的同名文件。被保存的文件因为拥有相同的哈希值，所以即使采集时间不同，也可以在保持原始状态的同时对其进行再现。

通过差异采集，不但可以减少要保存的文件数量，而且可以减少保存文件所需的存储空间。如前所述，WARP项目每月都会对国家机关的资源进行采集，每季度对其他机构进行采集。经测试，与完全采集相比，差异采集方式约能减少70%的采集量。换言之，差异采集所需的存储容量约为全部馆藏的30%。通过差异采集方式进行采集，有效地节省了互联网资源存档的存储空间。

2.3 互联网资源存档的内容组织

为了给用户提供更好服务，日本国立国会图书馆对采集到的网站进行了深度加工，分别是网址深加工、元数据编目、全文内容挖掘的处理。

2.3.1 网址深加工

采集网站回放的URL虽与原始URL不同，但保留了与原始URL的关系。图书馆回放地址通过两种形式呈现，即日期和网址组合、标识符和网址组合。

日期和网址组合的回放网址由三部分组成，即互联网资源存档域（http://web.archive.org/web/）、日期（20040618115539）和原始URL（/http://www.meti.go.jp/），其表示该网址是在2004年6月18日11:55:39开始采集的。

与日期和网址组合不同的是，标识符和网址组合将日期替换为标识符信息（info:ndljp/pid/285403/），而其他则保持不变。

2.3.2 元数据编目

日本国立国会图书馆会根据文档大小、用户需求和目标内容3个方面来控制元数据的粒度。

（1）在批量采集的情况下，由于文档数量巨大，难以提供细粒度的元数据；而在选择性采集的情况下，由于文档数量很小，因此会提供相对详细的元数据。

（2）互联网资源存档内容最终的目的是服务用户，因此元数据应满足一般用户的需求。当用户需要详细的元数据时，图书馆会尽可能提供。

（3）元数据的粒度还取决于目标内容。按特定目标采集互联网资源时，在采集之前，会将标题、发布者和时间等元数据添加到待采集的目标互联网资源中。发布网站时，一部分会直接使用原始网站的元数据，如标题、发布者和原始URL；一部分会在原有基础上增加一些必要的元数据字段，如摘要、主题事件、主题人物和关键词等，因此元数据并不统一。此外，图书馆会从保存的网站中提取出特定的出版文档，如白皮书、会议资料、报告、年鉴和论文，并为其添加详细的元数据。这样，用户就可以集中、有效地搜索和浏览散布在整个互联网网站上的出版物。

2.3.3 全文内容挖掘

互联网资源存档的搜索服务与元数据编目是互补的关系，但是只基于元数据的搜索服务并不完善，因为透过元数据搜索不会搜到存档内容的详细信息，因此在元数据搜索的基础上还需开发全文搜索服务。目前，全球60%的互联网资源存档机构都具备全文搜索功能。

日本国立国会图书馆WARP项目利用开源软件Solr进行二次开发，在Solr服务器上使用warc-indexer插件对存档文件进行索引，实现对所有采集资源（html页面、pdf、不同媒体类型的元数据、URL等）的全文和元数据检索。全文检索功能除了需要对存档内容进行索引加工外，还需要硬件设备的支持，由于全文索引和搜索需要具备高速计算和快速响应的搜索服务器，同时由于存档网站的数量巨大，因此还需要具备快速读写的存储设备。

2.4 互联网资源存档保存技术

无论是书籍还是数字内容，图书馆都必须保证其

保存的内容可以长期使用（100年或更长）。这种措施和尝试被称为长期保存。

2.4.1 存档资源内容的保存技术

互联网资源存档的长期保存主要通过数据冗余和不同介质备份两种方式完成。

数据冗余主要用于防止由于硬盘故障而导致的数据丢失，目前通过使用RAID（磁盘冗余阵列）等技术来实现。

不同介质备份主要是定期将硬盘上的数据备份到光盘等其他介质，以保留多代数据。划分存储位置以进行风险分配（灾难恢复）也是防止数据丢失的一种重要手段。数据的存储介质需要存储在稳定的物理环境中，并且需要定时进行介质转换以防止存储介质的劣化。

2.4.2 存档资源质量的保存技术

互联网资源存档不仅要保存数据内容，而且要保证数据能被正常使用。日本国立国会图书馆采用数据迁移和虚拟软件的方法来保证数据的可用性。

数据迁移是文件由于硬件或软件环境的变化，在技术上变得不可读之前，需要转换格式或迁移到另一种存储介质的方法。例如，使用老式处理软件创建的文件转换为最新的处理软件的数据格式，或者在硬件设备更改时将介质从软盘更改为光盘来保存数据。

虚拟软件是在新的硬件和软件环境下，模拟原来的文件和软件的使用环境。例如，可以通过使用虚拟软件在最新的Windows环境中重现Windows 3.1环境来使用仅在Windows 3.1上运行的软件。

为了有效地管理和实施数据迁移和虚拟仿真，有必要创建与保存相关的元数据，以记录数据存档时的播放设备、播放环境、创建应用程序、文件格式版本等。通过将存储在元数据中的信息与最新的技术趋势进行比较，可以及时掌握文件的过时情况并进行数据迁移和准备合适的虚拟仿真环境。

2.5 互联网资源存档的发布

2.5.1 互联网资源存档发布范围

在世界各地的互联网资源存档机构中，很少有将其存储的所有内容无条件地发布在互联网上，资源的发布经常受到一些限制，如访问的位置、资格、范围等。

存档机构采集并保存资源必须要使用它，否则毫无意义。日本国立国会图书馆综合考虑版权、个人信息和许可条件等采用了不同的发布形式。对于法律允许的采集内容在互联网上公开发布。对于一些版权要求严格或者包含许多个人信息的资源，出于研究目的，只在图书馆内部发布。

2.5.2 互联网资源存档的发布形式

日本互联网资源存档项目为了给用户提供更好的服务，通过多种形式对采集资源进行发布。①网站搜索服务：将采集的资源进行整合、编目、索引发布到官方网站上，用户通过搜索找到自己所需资源，这是世界上通用的发布形式。②专题服务：每月确定一个专题，按照专题的需求整合存档内容，发布到专题页面。③特色服务：将采集的内容进行整合和深度挖掘，通过可视化和数据再加工的形式展示给用户，让用户能更加直接地了解存档项目的使用价值。④历史网站服务：日本用户通过浏览器浏览网站如果出现错误或者打不开时，将提供跳转到WARP历史网站界面选项，进入后可以选择不同采集日期的页面，让用户能够浏览被修改和删掉的网站内容。

3 日本互联网资源存档对我国的启示

3.1 开发互联网资源存档的采集软件

目前国内存档机构还在采用完全采集的方法对网站进行采集，这导致许多数据被重复采集，造成人力资源和存储资源的浪费。日本国立国会图书馆利用差异采集方法实现了只采集修改的网站数据，节省了时间和存储空间。

随着互联网的快速发展，越来越多的资源需要采集，差异采集方法是必然趋势。我国存档机构目前正面临采集数据量快速增长导致存储空间不够的问题，而差异采集能够减少存储空间的占用从而提高采集效果，因此国内存档机构可以借鉴现有的差异采集软件，如DeDuplicator、OutbackCDX和warcrefs的技术经验，开发出适合中文数据资源的差异采集软件[7-9]，解决存储空间不够的问题。差异采集方法实现之后，不但能够解

决国内存储空间紧张的问题，还能解决后期发布人工删除重复页面的工作，大大节约了人力和时间成本。

3.2 建立互联网资源存档的元数据库

中国国家图书馆互联网资源存档项目的编目仅采用一种统一的编目格式，并没有针对文档大小、用户需求和目标内容控制元数据的粒度。中国互联网资源存档数据量巨大，国家图书馆由于受到人力和财力的限制，像日本一样将元数据添加到所有互联网资源存档内容中是不现实的。当前国内图书馆互联网资源存档项目采集的资源随着互联网的发展越来越多，单独通过网址查找资源已经不能满足用户的需求（并不是所有用户都知道准确的网址），因此亟需建立自己的元数据库，让用户能够通过元数据准确查找资源。虽然因存档数据量巨大，无法通过人力实现对所有的存档数据建立详细元数据库，但是可以借鉴日本国立国会图书馆的经验，将采集的出版物提取出来，单独制作详细的元数据，为用户提供服务。

中国互联网资源存档解决用户通过元数据查找资源的需求，需要开发一套资源采集系统，理解所采集网站的内容，并利用语义网等技术自动添加元数据。存档编目还可以引入社交标签的机制，让用户自行将主题的元数据添加到正在观看的内容当中。元数据库建立后不但能让用户通过元数据准确查找资源，而且还能通过元数据建立资源之间的关系，提供关联服务。

3.3 强化互联网资源存档的长期保存

互联网资源存档不仅是把网络资源做一个备份存储下来，而且还要保证采集到的资源能够通过浏览器回放。国内存档机构目前还处在扩大采集规模和数量的阶段，对于保存只是做了硬盘备份和服务器RAID设置，并没有考虑到资源的长期使用和长期保存。

日本国立国会图书馆从存档数据长期保存和长期使用的角度出发，在硬件上利用服务器RAID设置和定时转换存储介质的方法来保证数据的长期完整性，在软件上利用数据迁移技术和虚拟软件的方式来保证数据的实用性。中国存档机构可以借鉴日本的经验，根据国内存档情况制定定时存储介质转换计划和积极开发虚拟软件模拟资源的原始运行环境，保证存档的数据能够长期保存和使用。互联网资源长期保存的目的就是让消失和被修改的资源能够以原始的样式重新展示，让更多的人通过互联网存档计划了解真实的历史和文化，因此保证长期保存数据的可用性是十分重要的。

3.4 完善互联网资源存档的法规建设

合法性通常是网络资源存档面临最大的非技术性问题[10-11]。在所有者没有明确许可的情况下，是否拥有复制内容和提供独立于原始网站访问的合法权利？是否侵犯了所有者的版权？一些网站明确标出了版权许可或版权授权信息，如知识共享或官方版权，可以部分解决网络存档合法性问题。但是，很大程度上取决于有关国家规定和存档机构的职权范围。

日本国立国会图书馆采用法律授权和创建者授权的方式，解决了互联网资源采集和服务的合法性问题。目前，中国国家图书馆正在积极准备互联网资源存档相关法律的提案，如果提案被通过，国家图书馆将能够对互联网信息进行复制、编辑、长期保存和公共服务。在此之前，国内存档机构需要积极与网站创建者沟通获取采集和发布权限，尽最大可能保存即将消失的互联网资源。

3.5 加强互联网资源存档的国际合作

中国的互联网资源采集机构主要有国家图书馆、北京大学、国家档案馆、台湾图书馆和台湾大学图书馆。不同机构虽然采集策略不同但还是有重合的地方，会形成对一个站点重复存档的问题。国内存档机构的交流与合作有助于避免网站的重复采集和技术升级，实现更大规模的互联网资源存档。

日本国立国会图书馆积极参与国际交流，利用开源软件进行二次开发，实现了互联网资源存档的快速发展。因此，国内存档机构应积极参与国际交流并吸收国外经验，让国内互联网资源存档尽快达到国际标准。互联网资源存档是一个全球化的工作，国际交流和合作是必不可少的，通过交流不但能够获取先进的技术，而且保证了所保存的内容符合国际标准。

4 结语

随着互联网的快速发展，越来越多的行业从线下

转到了线上，在网络上产生了大量有价值的资源，同时由于互联网资源与实体资源相比具有寿命较短的不足，互联网资源存档势在必行。但目前中国互联网资源存档还处于初级阶段，没有完善的法律保障、先进的技术支持和充足的资金保证，因此面对海量的网络资源，如何进行批量采集、加工、编目、保存和发布，突破知识产权和采集技术两大难题，成为亟待解决的问题。日本国立国会图书馆互联网资源存档项目的成功，为我们做了很好的示范和启示，我们应该吸收和借鉴日本的成功经验，包括差异采集方式、元数据规范、长期保存技术等，建立完善的法律法规、加强国内外交流学习先进的采集技术，建设适合中国的互联网资源存档项目，实现中国互联网资源存档的快速发展。

参考文献

[1] Internet Archive [EB/OL]. [2020-09-07]. https://archive.org/about/.

[2] 国立国会図書館インターネット資料収集保存事業（WARP）[EB/OL]. [2021-01-02]. https://warp.da.ndl.go.jp/.

[3] 国立国会図書館インターネット資料収集保存事業統計 [EB/OL]. [2020-05-11]. https://warp.da.ndl.go.jp/info/WARP_statistic.html.

[4] 国立国会図書館法によるインターネット資料の収集について [EB/OL]. [2021-01-02]. https://warp.da.ndl.go.jp/bulk_info.pdf.

[5] 陈瑜. 日本国立国会图书馆网络信息资源采集保存项目介绍研究 [J]. 图书馆杂志，2014，33（3）：91-94.

[6] 闫晓创. 日本网络资源存档项目实践研究 [J]. 浙江档案，2017（12）：20-23.

[7] 孟庆浩. 互联网数据增量采集系统的设计与实现 [D]. 北京：北京邮电大学，2015.

[8] 孟庆浩，王晶，沈奇威. 基于Heritrix的增量式爬虫设计与实现 [J]. 电信技术，2014（9）：97-101.

[9] 高婷，白如江. 基于OutbackCDX的增量式Web信息采集研究 [J]. 山东理工大学学报（社会科学版），2020，36（4）：99-105.

[10] 陆媛媛. 《公共图书馆法》应关注网络信息资源长期保存问题 [J]. 安徽电子信息职业技术学院学报，2017，16（1）：104-107.

[11] 张林华，徐维晨. 浅析国外网页档案实践及其对我国的启示 [J]. 档案与建设，2020（6）：9，38-41.

作者简介

杨云鹏，男，1986年生，硕士，工程师，研究方向：数字资源整合与互联网资源存档，E-mail：syzyyp@126.com。

Research and Enlightenment of Internet Resource Archiving in the National Diet Library of Japan

YANG YunPeng
(National Library of China, Beijing 100081, China)

Abstract: The National Diet Library of Japan started the internet resource archiving project WARP in 2002 and has established a complete system. This paper gives a detailed introduction to the internet resource archiving project of the National Diet Library of Japan from four aspects of website screening, collection technology, website processing and preservation technology. Meanwhile, it puts forward a proposal for China to carry out internet resource archiving from five aspects of collection methods, legal construction and international exchanges to get better development.

Keywords: Internet Resource Archive; Website; National Diet Library of Japan; Collection

（收稿日期：2021-01-02）