网络资源存档数据质量问题初探

王文玲[1] 曲云鹏[2]
(1. 国家图书馆, 北京 100081; 2. 中国科学院文献情报中心, 北京 100190)

摘要: 网络资源存档的数据质量是影响网络资源存档工作成败的主要因素之一, 本文探讨如何开展网络资源存档数据质量评价, 在分析网络资源存档数据质量问题的表现及其成因的基础上, 提出解决网络资源存档质量问题的方法体系。该体系以存档数据为中心, 严格按照既定的业务标准及工作规范, 利用现有软件工具对采集过程进行全面的数据质量检查, 确保获取高质量的存档数据。

关键词: 网络资源存档; 数据质量; 质量评价; 质量保证
中图分类号: G250.7
DOI: 10.3772/j.issn.1673-2286.2018.04.002

网络资源传递信息手段丰富，除文字信息外，网站外观、浏览体验、交互式操作等都可以向用户传递重要信息。理想的网络资源存档（Web Archiving，WA）不但可以保存文字内容，还可以记录网页外观、浏览体验等信息，重现网站页面时能无限接近其原始面貌，有利于对现代网络资源情况进行研究与利用。但在WA的工作实践中，会遇到多种类型的数据质量问题，包括网站内容文件的缺失、多媒体内容无法展现、版式错乱等。如果对这些数据问题不采取严格的质量控制手段则可能丢失很多重要的信息，导致数据质量偏低，以致保存任务失败，因此，数据质量是影响WA工作成败的主要因素之一。

WA的数据质量保障工作指WA机构为保证所采集的网络资源达到预设的质量标准而采取的一系列措施，包括机器自动执行及人工干预等方式，范围覆盖采集前、采集中及采集后整个流程，是WA的一项重要工作。对数据质量保障工作进行研究，可以让保存机构明确数据质量问题发生的环节所在，明确解决数据质量问题的方法与手段，为保存机构提高采集效率、降低采集成本提供参考。

本文的主要目的是探讨如何开展WA数据质量评价，分析WA数据质量问题的表现及其成因，并提出解决WA数据质量问题的方法建议。

1 WA数据质量评价

WA的目标是将实时的网络资源按照原样复制保存，最理想的结果是能够获得与源网页完全一致的副本，包括网站的外观及所有功能。但在具体的WA实践中，由于各种各样的原因，通过网络爬虫很难获取与源网页完全一致的存档数据，那么，如何评价WA工作、如何评估WA的数据质量成为网络采集专家面临的首要问题。

1.1 WA数据质量评价相关研究

Masanès[1]较早开始对WA数据质量进行系统研究。他认为，WA的数据质量可以从两方面进行评估：①指定范围内资源保存的完整性；②资源是否可以展现网站的原貌，尤其是导航，以及与用户间的交互行为。

通常情况下，以网站主页为入口，链接可以指引用户到一个新的网站或同一个网站的其他元素，此时完整性可以用两个维度来判断：①水平方向，在指定范围内找到的相关入口数量；②垂直方向，从该入口发现的相关链接点的数量。展现网站的原貌可以通过回放软件来判断，如果所采集的数据中保存有相应的脚本和目标网页，并且回放软件支持交互性回放，就可以再现采集时网站的原貌及其交互特性。

Hockx-Yu[2]认为WA的质量应该从以下方面来评

估：①采集完整性，预期内容是否被完整地采集；②知识内容，知识内容（相对页面设计及布局）是否可以通过访问工具回放；③网站行为，采集的网站副本包括实时网站行为是否可以回放，如链接之间的交互能力；④外观，是否可以准确展示网站的界面外观。Hockx-Yu[2]认为，在应用以上质量标准时，首先应强调知识内容，即注重可以体现信息内容的文字、图片和设计；其次，强调网站的外观或网站行为。即使外观不是完全准确，只要网站的大部分内容被采集，且能够合理地回放，采集的副本就可以被提交到档案中长期保存。

除完整性外，Saad等[3]认为WA数据质量对时效性要求更高，提出需要从一致性角度对数据质量进行定义。他认为网站时刻处于变化中，开展WA时应在一定时间切片内尽快保存网站内容，以确保当前时间切片内资源不会发生变化，一旦网站发生变化，WA的一致性将无法得到保障；此外，采集到的网站版本与采集开始时网站的版本相比较，差别越小，WA质量越高。

通过分析上述观点可知，外观完整性、交互完备性和数据一致性被视为存档数据质量的三大评价指标。高质量的WA数据指在尽量短的时间内，完整采集目标网站中的知识内容，并且完整保存网站的视觉内容和浏览体验。

1.2 建立项目适用的WA数据质量评价标准

对每个存档项目来说，无止境地追求高质量存档数据是不现实的，也是没有必要的，各项目应根据既定目标，结合项目采集需求，综合考虑项目预算，尽量平衡三大质量评估指标，制定本项目适用的质量评价标准。

1.2.1 采集需求

不同采集类型对质量的要求不尽相同，其中广度采集相比深度采集对内容完备性的要求更低。如瑞士国家图书馆的Kulturarw项目采用广度采集策略，致力于对国家域名、互联网全域名进行快照，全程采用自动化方式，除对采集过程进行检查外无其他质量控制措施[4]；法国国家图书馆的WA采用深度采集策略，旨在采集网站的所有内容，甚至包括深度网络（深度网络指存储在数据库中的内容，可通过用户名密码、IP认证等方式来采集）。深度网络的采集无疑会提高存档数据质量，但也会相应地提高采集成本。

1.2.2 成本决策

目前，WA数据质量保证工作是人工或半自动化的过程，追求高质量的数据必然会增加时间成本和人力成本，在实际的WA实践中，每个项目对其采集范围、项目耗时及成本都有既定目标，若盲目地追求高质量可能导致项目超期或预算超标。如澳大利亚的PANDORA项目通过专业的质量控制专家对所有存档网站进行人工核验，但并非所有的项目都有充足的经费来支持精细的质量控制过程，对高质量的追求必须充分考虑成本预算[5]。

1.2.3 平衡三大质量指标

由于互联网资源的复杂性，以及各个存档项目的差异性，存档数据质量的三大指标很难标准化，也很难做到同时满足三个指标的要求。若对采集任务的时效性要求较高，需尽量缩短采集时间，有可能降低外观完整性和交互完备性。因此，在实际工作过程中需要在三个指标中寻找平衡，根据不同采集目的设定爬虫的采集规则，尽可能提高数据质量。

2 WA数据质量问题的表现

WA数据质量问题有多种表现，发现数据质量问题的主要手段是使用回放软件对采集数据进行回放，再通过人工点击和浏览发现存在的问题。通过回放一般能发现以下六类数据质量问题。

（1）数据无法进行回放。WARC（Web Archiving File Format）是目前唯一专门面向网络资源长期保存的文件格式，于2009年5月成为国际标准（ISO 28500：Information and document-WARC file format），2017年7月正式成为中国国家标准（GB/T 33994—2017：信息和文献WARC文件格式），是网络资源采集机构及常用的网络采集系统或工具广泛采用的标准[6]。如果采用非标准的采集软件或软件运行过程中出现问题，则可能导致生成的数据不符合WARC格式而无法通过回放软件进行回放。

（2）内容信息缺失。网页资源是一种复合型资源，其内容信息包括文字、图片、视频及流媒体等，这些内

容信息可能分散在不同的网络位置，由HTML语言整合在同一个页面。通常大型网站会将不同类型的资源放置在不同域名下，如新浪的主域名“sina.com.cn”，图片文件存储在“n.sinaimg.cn”域名下。若回放时发现图片、视频信息无法显示，则有可能是这些资源分布在不同的服务器上，没有被爬虫捕获，或资源已被采集但回放软件不支持这类文件格式。

（3）功能模块缺失。网页上的一些功能模块，本身并不提供信息内容，而是为网站内容（尤其是动态内容）提供容器或入口，如菜单栏、广告栏和浮动窗口能极大提高网站的可用性和交互性。这些功能模块通常采用样式单、JavaScript、嵌入式播放器等技术来实现，这些技术目前还处于WA尚未完全解决的技术清单中，因此功能模块缺失是WA中常见的问题。如国家图书馆的公开课资源采用JavaScript调用视频播放软件对课程资源进行播放，若未采集到JavaScript代码或嵌入的视频播放软件代码，则整个播放功能模块都会缺失。

（4）交互性缺失。网站的交互性可能通过多种方式展现，如功能性链接及按钮、下拉式列表、Flash等富媒体，这些功能的实现需要实时与后台数据库进行通信，或调用其他服务器上的脚本，或使用代码进行封装，这些对网络爬虫都存在一定技术难度，很可能无法获取，从而造成网站交互功能的失效。如新浪微博的交互性在于可以对微博转发、评论和点赞，但是爬虫程序抓取网页后，缺乏后台数据库的支撑，即使回放软件支持回放，这种交互性也很难重现。

（5）网页外观无法重现。网站内容虽然可以展示，但外观简陋、排版错乱的情形时有发生。出现类似问题的原因是负责网站外观的CSS文件和其他与网页样式有关的内容未被采集。

（6）网站内容变更或错误。在WA过程中，目标网站内容发生改变，导致采集到的数据与目标网站存在差异。典型的案例为采集频率设置不恰当，如果网站更新频率为1个月，而采集频率低于1个月，那么就会丢失很多更新的内容信息。而更为严重的问题是由于采集种子IP地址错误，或原网站内容被恶意篡改，导致采集到的内容与既定采集目标完全不符。

3 WA数据质量问题的影响因素分析

WA是一项复杂的工作，任何环节的疏忽都会导致存档数据质量问题，甚至导致采集失败，对影响WA数据质量的因素进行归纳有助于分析造成质量问题的深层次原因，以便有针对性地提出解决质量问题的方法和手段。本文将影响WA数据质量的因素归纳为策略因素、技术因素、权利因素及环境因素。

3.1 策略因素

由于网站开发语言的灵活性，以及每个网站独特的业务逻辑和技术框架，为达到最佳的采集效果，最理想的方法是对每个网站制定个性化的采集策略，明确采集的深度、广度和采集频率，多个域名的作用及其关系，是否遵守蜘蛛协议，是否存在爬虫陷阱，制定采集规则及种子列表等。然而在实际的采集实践中，尤其是大规模采集，针对每个网站制定不同的采集策略并不现实，只能采用一种或多种采集策略，这样势必导致诸多数据质量问题。如某网站的主页地址是“主域名/pages/default.aspx”，访问主域名不会显示任何内容信息，而是直接跳转到网站主页，若此时只将主域名列入种子列表，爬虫访问时即使会爬到网页主页进行采集，但由于已经发生一次跳转，相当于减少了采集深度，会导致采集深度达不到既定目标。

3.2 技术因素

技术因素主要指目前WA技术发展滞后，无法跟上网站开发技术的发展，从而导致无法采集到相关内容。比较典型的网页开发技术有动态脚本或应用、流媒体和嵌入式播放器、表单或数据库驱动的内容等，一些网站融合数据库类型资源、复杂脚本类型资源的技术特点，如社交媒体网站、网页游戏、在线地图等都是网络爬虫很难攻克的技术难点。

除上述客观的技术障碍外，还有一种网站主动设置的技术壁垒——爬虫陷阱。爬虫陷阱是一种根据爬虫工作方式而设置的技术陷阱，爬虫经过网页时，会陷入无限循环的爬行中，无法跳出。这种陷阱与在线日历类似，爬虫一页接一页地进行爬行，采集不到更多有意义的网站内容，直至达到预设的最长采集时间才能停止，否则无法跳出。

3.3 权利因素

虽然网络资源及网站信息均可公开访问，但并不意

味着WA机构有权利免费获取这些资源进行长期保存。博客、微博、微信等社交媒体网站，以及当下流行的网络表演平台，绝大多数资源由用户提供并上传，用户拥有这些资源的知识产权，WA机构并不能完全确定是否有权利抓取并存取这些资源。

一些网站会主动发布蜘蛛协议，明确网站是否欢迎网络爬虫，以及欢迎哪种类型的爬虫；同时，规定网站哪些目录可以采集，哪些目录不允许采集。目前，蜘蛛协议尚未成为国际标准，也不是强制执行的规范，只是互联网行业约定俗成的协议。虽然现有的网络资源采集软件支持是否遵守蜘蛛协议进行配置，但从保存的角度出发，采集机构通常会在法律允许的范围内忽略蜘蛛协议，尽可能完整地保存互联网资源。

3.4 环境因素

环境因素指网络采集软件的运行环境及网络环境等。运行环境主要指服务器的硬件及软件状态，若服务器发生硬件故障，或内存溢出导致宕机，或突然断电，则会导致采集生成的WARC文件发生损坏，无法进行内容的读取，从而导致采集失败。若服务器网络连接发生故障，或由于防火墙设置无法访问采集目标，在重试连接达到一定次数后，采集软件会将采集对象备注为“404错误”；若发生域名劫持，那么DNS服务器可能将采集对象解析为错误的IP地址，导致采集软件无法采集到正确的内容。

4 WA数据质量保证方法

通过以上对WA质量问题表现及影响因素的分析，本文提出WA数据质量保证的方法体系（见图1），该体系以存档数据为中心，通过制定一系列严格的业务标准及工作规范，利用现有软件工具对采集过程开展全流程的数据质量检查工作，同时以团队建设、环境维护及授权获取网站备份为补充手段，确保获取高质量的存档数据。

4.1 制定严格的采集业务标准和工作规范

由于质量控制专家的背景知识、技术水平及工作熟练程度各不相同，为避免人为因素导致的数据质量问题，应当为WA工作制定统一的业务标准和严格的工作规范。推荐设定的标准应包括7种：①数据质量评价标准。根据上述结论，WA机构应当综合考虑项目既定目标、采集类型、项目预算，平衡三大质量评估指标，制定项目适用的质量评价标准。②数据标准及元数据规范。包括资源保存格式及网站对象元数据规范。③软件使用标准。明确采集软件、回放软件、杀毒软件的选择标准，确保生成数据的一致性。④种子筛选标准。确定种子筛选的原则，保证所采集到的数据既符合网站原貌，又符合WA的要求。⑤种子排序标准。根据采集目的确定种子优先级排序的标准，可以更有针对性、更有效地进行存档。⑥爬虫配置标准。设置深度采集和广度采集两种采集方式的爬虫默认配置，每次采集根据具体的采集需求，爬虫配置做尽可能小的修改，以避免人为因素导致的爬虫设置错误。⑦操作规范。应当为每一个程序化的质量保证步骤制订便于执行的工作规范，如病毒检查工作规范、软件回放质量检查工作规范等。

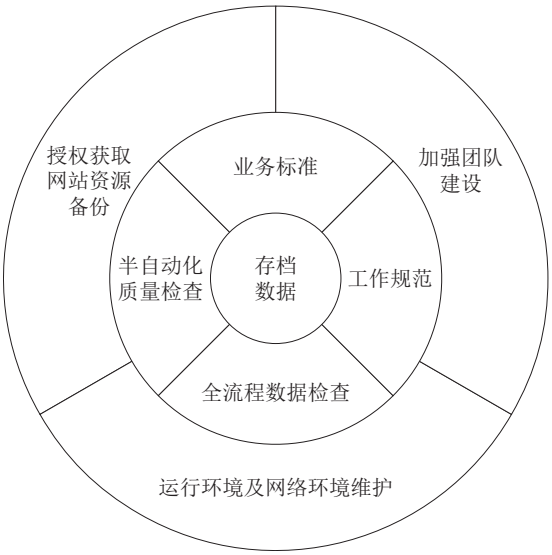

图1 WA数据质量保证的方法体系

4.2 开展全流程的数据检查

开展直接的数据质量检查工作是保证数据质量最有效的方式，根据国际互联网保存组织的调研[7]，绝大多数成员机构都会进行质量控制，只有不到5%的机构从来不进行质量检查，对采集过程开展全流程质量控制的机构约为11%。质量检查工作是质量保证工作的核心，应贯穿WA的整个工作流程，包括采集前、采集中和采集后三个时间段。采集前的质量控制工作主要针对网站，检查采集该网站信息所需要注意的特征等，进行预

采集，从而制定或修改相应的采集策略；采集中的质量控制工作主要监测爬虫的运行状态和运行日志，及时排查并解决错误；采集后的质量控制工作主要对所采集的数据及软件的日志进行检查，保证数据格式和内容正确，并针对错误进行排查解决。

4.3 开展半自动化质量保证工作

为提高WA质量保证工作的自动化程度，减少人工参与度，越来越多的WA软件工具开始集成质量保证的功能模块，也有不少专门的质量保证辅助工具出现[8]。如果能够充分利用这些工具，则可以省时省力地完成质量保证工作。如Heritrix可以提供详细的爬行任务运行日志信息和报告，以及强大的采集规则配置工具，确保能完成各种高精度的采集任务[9]；Web Curator Tool提供专门的质量检查工具，可将采集到的种子地址以直观的树形结构进行展示，用户可以根据需要直接对这些资源进行修剪[10]；NetArchiveSuite提供专门的质量检查工具Viewerproxy，可以对采集的资源进行回放[11]；Monitrix是一个专门针对Heritrix 3设计的前端监控分析软件，目前还单独开发了一个日志信息可视化平台Kibana[12]。

4.4 其他策略

（1）加强WA团队建设。WA数据质量保证工作目前主要通过人工完成，质量控制专家的专业能力直接影响质量保证工作的效果。作为合格的质量控制专家应熟练掌握互联网相关知识，包括互联网数据传输技术、网站开发技术、网络硬件相关知识；此外，还需具备较强的数学能力、逻辑推理能力和编程能力等。面对互联网技术突飞猛进的发展速度，应当加强对WA团队的培训，提升其专业能力。

（2）做好运行环境及网络环境维护。维护良好的软硬件运行环境及网络环境是保证高质量WA工作的前提，网络采集团队应制定严密的服务器及硬件管理制度，运用各种监控网络硬件的设备及软件，定期对服务器软硬件运行环境进行检查，为WA工作提供良好的软件、硬件运行环境及网络环境。

（3）直接获取网站资源备份。网络爬虫是一种有瑕疵的WA技术，可模拟人类浏览网页时的情形，然而又不能完整地进行模拟，因此这种方法永远不能完美呈现原始网站面貌。保存机构可以采取与网络资源所有者进行合作，在解决知识产权等相关问题的前提下，直接从提供商处获取网站及资源的数据备份，包括后台数据库、嵌入式资源及动态脚本等。

5 未来展望

目前，WA机构所采用的及本文所提出的数据质量控制手段主要针对策略因素造成的质量问题，针对技术因素和权利因素造成质量问题的研究及实践相对较少。在未来的工作中，WA机构可以在两个方面进行努力，推动WA工作的发展与进步。

（1）加强WA技术的研究，增强网络爬虫的采集能力。在权利问题解决前，网络爬虫仍然是WA工作的主要手段之一。目前网络爬虫面临的主要技术问题是富应用封装的网络资源无法采集。提高网络爬虫解释和处理复杂逻辑的能力，从富客户端和沉浸式环境中获取原本无法获取的资源，从而提高WA数据质量的完整性。

（2）推动相关立法，从出版机构获取网站原始数据备份。有超过半数的WA工作都是由图书馆开展的非营利性工作，由于没有相关法律的支持，只能采用效率比较低的网络爬虫技术。WA机构应该推动本国相应网络资源呈缴立法工作，从而合法获取网站原始数据备份。

参考文献

[1] MASANÈS J. Web archiving methods and approaches：a comparative study [J]. Library Trends，2005，54（1）：72-90.

[2] HOCKX-YU H. How good is good enough? – Quality Assurance of harvested web resources [EB/OL]. [2018-02-05]. http://britishlibrary.typepad.co.uk/webarchive/2012/10/how-good-is-good-enough-quality-assurance-of-harvested-web-resources.html.

[3] SAAD M B，GANÇARSKI S. Using visual pages analysis for optimizing web archiving [C] // Proceedings of the 2010 EDBT/ICDT Workshops，Lausanne，Switzerland，March 22-26，2010.

[4] 龙正义. 网页长期保存的策略与方法研究 [J]. 档案管理，2010（3）：20-23.

[5] 闫晓创. 国外Web Archive项目对我国的借鉴和启示——以澳大利亚的PANDORA项目为例 [J]. 档案学研究，2012（5）：79-83.

[6] 曲云鹏. 网络资源存档文件格式WARC研究 [J]. 图书馆学研究，2014（24）：20-25.

[7] AYALA B R，PHILLIPS M，KO L. Current quality assurance

practices in web archiving [EB/OL]. [2018-02-05]. http://digital.library.unt.edu/ark:/67531/metadc287034/.

[8] 吴振新，曲云鹏，李成文，等. 基于开源软件搭建网络信息资源采集与保存平台[J]. 现代图书情报技术，2009(7/8)：6-10.

[9] MOHR G，STACK M，RANITOVIC I. An introduction to Heritrix：an open source archival quality web crawler [C] // Proceedings of the 4th International Web Archiving Workshop，Bath，UK，2004.

[10] Web Curator Tools [EB/OL]. [2018-02-05]. http://webcurator.sourceforge.net/.

[11] NetArchiveSuite [EB/OL]. [2018-02-05]. https://sbforge.org/display/NAS/NetarchiveSuite.

[12] JACKSON A. Watching the UK domain crawl with Monitrix [EB/OL]. [2018-02-05]. http://blogs.bl.uk/webarchive/2013/09/monitrix.html.

作者简介

王文玲，女，1981年生，硕士，馆员，研究方向：文献资源建设、数字图书馆技术，E-mail：wangwl331@163.com。
曲云鹏，男，1980年生，博士，副研究馆员，研究方向：数字图书馆技术。

An Overview of the Issues of Data Quality in Web Archiving

WANG WenLing[1] QU YunPeng[2]
(1. National Library of China, Beijing 100081, China; 2. National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Data quality issue is one of the key factors affecting the result of web archiving. The aspects of assessing data quality in web archiving are clarified in this article. Then the representations and the cause of data quality issues are analyzed. At last a framework of methodologies is developed to solve those issues. In this data-centered framework, business standards and working specifications are complied strictly, all kinds of softwares and tools are used to perform comprehensive data checking during the harvesting period to ensure high data quality.

Keywords: Web Archiving; Data Quality; Quality Assessment; Quality Assurance

（收稿日期：2018-02-06）

书 讯

《汉语主题词表》（工程技术卷）

《汉语主题词表》自1980年问世以后，经1991年进行自然科学版修订，在我国图书情报界发挥了应有的作用，曾经获得了国家科学技术进步二等奖。为了适应网络环境下知识组织与数据处理的需要，2009年由中国科学技术信息研究所主持，并联合全国图书情报界相关机构，完成《汉语主题词表（工程技术卷）》的重新编制工作。

全书共收录优选词19.6万条，非优选词16.4万条，等同率0.84。在体系结构、词汇术语、词间关系等方面进行改进创新。为了方便工程技术领域不同专业用户使用，《汉语主题词表》（工程技术卷）按专业分13个分册出版，同时建立《汉语主题词表》服务系统，提供在线概念检索和辅助标引服务，通过可视化技术展示各类概念关系，是图书馆、档案馆、出版社、期刊杂志社、文献信息中心等专业工作者及科研、教育及工程技术领域人员必备的参考书。

《汉语主题词表（工程技术卷）》已于2014年由科学技术文献出版社出版，全书2 300余万字，总定价3 880元，可分册购买。