历史学堂

HISTORICAL STUDIES WITH BIG DATA

舒健 主编

大数据时代的历史研究

上海译文出版社

图书在版编目(CIP)数据

大数据时代的历史研究/舒健主编. —上海：上海译文出版社，2018.1
(历史学堂)
ISBN 978-7-5327-7623-8

Ⅰ.①大… Ⅱ.①舒… Ⅲ.①史学—文集 Ⅳ.①K0-53

中国版本图书馆CIP数据核字(2017)第206631号

本书由上海文化发展基金会图书出版专项基金资助出版

图字：09-2014-119号

大数据时代的历史研究

主编/舒　健
责任编辑/钟　瑾　装帧设计/半和创意

上海世纪出版股份有限公司
译文出版社出版
网址：www.yiwen.com.cn
上海世纪出版股份有限公司发行中心发行
200001 上海福建中路193号 www.ewen.co
上海信老印刷厂印刷

开本 890×1240　1/32　印张 14　插页 2　字数 270,000
2018年1月第1版　2018年1月第1次印刷

ISBN 978-7-5327-7623-8/K·258
定价：68.00元

目录

第一部分　大数据与史学研究

3 / 变革亦变异?
——大数据时代的史料收集与历史书写
…… 陈文俊

21 / 大数据时代历史研究的可能性
…… 马建强

40 / 行为的可数据化：大数据时代的人文关怀
…… 朱锋刚　李　莹

第二部分　数据库介绍和建设

55 / 日本国内近代史研究相关基础史料的数据库建设现状
…… [日] 田中智子

66 / 地方历史文献的数字化、数据化与文本挖掘
——以中国地方历史文献数据库为例
…… 赵思渊

80 / 从高第书目到 Bibliotheca Sinica 2.0
——兼论数字化与汉学史研究
…… 王国强

92 / 汉语基督教文献书目数据库编目实践
…… 黄　薇　徐锦华

102 / 古籍数字化在历史研究领域的应用
——以韩国事例为中心
…… [韩] 李惠源

115 / 喜马拉雅研究与“喜马拉雅多媒体数据库”的建设
…… 姚　勇

126 / 跨越三地的史料数据整合
——闽渝档案馆藏台湾“光复”前后之档案的现状与利用
…… 吴巍巍

138 / 论近代中国报纸广告的蜉蝣性与数字档案化
…… 陈　静
151 / 浅析 “中国地方志书目数据库” 的建设
…… 胡艳杰
167 / 历史文献的视觉档案
——以上海图书馆“上海年华”数字资源平台为例
…… 黄　薇

第三部分　大数据与历史研究结合案例

183 / 数字人文时代的关系型数据库：中国历代人物传记资料库（CBDB）的应用
…… 徐力恒
199 / 社会关系网络与范成大《吴郡志·人物》之编撰：以 CBDB 及 Pajek 作为分析工具
…… 李宗翰　郑　莉
209 / 大数据在蒙元史研究中可能遇到的困难与挑战
——读钱大昕《廿二史考异》元史部分札记
…… 翁沈君
236 / 大数据视域下我国历史地理学研究现状与趋势
——基于《中国历史地理学论丛》（2005—2014）的统计
…… 郑　星
250 / 古籍数字化在明代科举研究领域中的应用与展望
——以《天一阁藏明代科举录》为例
…… 卞　梁
258 / 大数据视域下的夏商文化起源研究
…… 丁　新
271 / 田野调查方法和 GIS 技术支持下的山区聚落时空演变研究
…… 霍仁龙

第四部分　存在的问题和新的领域

293 / 少数民族古籍数字化在大数据时代下的发展前景及存在问题

…… 刘　琳

304 / 大数据视域下古籍文本可视化分析及挖掘在中国史定量研究中的应用

…… 欧阳剑

326 / 古籍数字化在中国医学史研究领域的应用

…… 孙灵芝

348 / 网络工具与古汉语语言文字研究

——以出土文献、古文书为中心

…… 魏郭辉

364 / 清代女性别集规模化整理的现状和方向

…… 肖亚男

373 / 上海图书馆“全国报刊索引”：数字出版中的史学情怀

…… 徐华博　彭　梅

384 / 对当前中国大陆高校图书馆所购史学类数据库的统计与分析

…… 张晓宇

399 / 从视觉进入医学史研究的新视野

——《中华图像文化史·医药卷》绪论

…… 张树剑

410 / 大数据在历史气候学研究中的应用与展望

…… 韩健夫

论近代中国报纸广告的蜉蝣性与数字档案化[①]

陈　静[②]

作为一种营销方式的广告在中国古已有之：街头叫卖的吆喝声、梆子声，店铺外悬挂的实物商品、招牌匾额和幌子旗帜，文人吟咏或街头传唱的诗文歌赋以及商家自制的包装仿单，处处可见，类型多样，皆以非广告之名在行广告之事。然而，真正现代意义上的“广告”，时至近代才在中国流行起来：“广告”作为一个专有名词被普遍接受并得以自成学科从而跻身现代学堂，[③] 各种舶来及本土的“广告学”研究的文章见诸报端和著作，[④] 广告生意更发展出了当时炙手可热的新兴行业。[⑤] 广告的类型也更为丰富，在传统的各种广告形式之外更多了基于纸质媒介的报纸广告、期刊广告、海报广告和月份牌广告，附着于街头的灯箱广告、墙面广告和电

---

① 本文为莱斯大学白露（Tani Barlow）教授领衔的“The Ephemera Project”项目的阶段性成果。部分内容曾在2015年数位典藏与数位人文国际研讨会（台湾大学）、2015年“大数据时代下的历史研究”国际学术研讨会（上海大学）上宣讲发表。在此感谢白露教授、清华大学邓柯教授对本文的帮助。

② 作者单位为南京大学艺术研究院。

③ 参见如来生：《中国广告事业史》，新文化社1948年版。

④ 参见郭瑾：《民国时期的广告研究及其当代意义》，《广告大观理论版》2006年第6期。其中对民国时期出版的有关广告学和与广告学相关的部分著作进行了整理，可为参考，但其所整理的论文部分不够全面。在民国时期期刊全文数据库（1911—1949）和晚清期刊全文数据库（1833—1911）以“广告”为主题搜索，结果共有4 329条。

⑤ 参见如来生：《中国广告事业史》，新文化社1948年版。

车广告，以及嵌于新的大众媒体的广播广告等。

广告成了近代中国商业经济的重要组成部分。以报纸广告为例，白瑞华（Roswell Sessoms Britton）指出，在华文报纸发展的早期，“大多数报纸的广告都是大型外国海运公司、保险公司和贸易公司发布的告白。报纸广告收入比订阅费的赢利要少。中国商业尚未开始转型，还不需要报纸上刊登广告”①。20 世纪 20 年代以降，在华中文报纸，尤其是商业报纸刊登广告已逐渐成为自然之惯例。据学者统计，《申报》《新闻报》《沪报》《时报》《大公报》《益世报》《中国报》《忠言报》《时闻报》等报纸广告版面逐渐超过了 50%以上。② 考虑到近代报纸可观的种类和发行量，可见当时报纸已经成为最重要的广告传播载体，而广告亦成为报纸，尤其是商业报纸最为依仗的经济来源。③

诚如戈公振先生在其《中国报学史》（1927）中所言：“我国广告事业，年有进展，自为可喜之现象。如《申报》、《新闻报》、《益世报》之经济充裕，不可谓非广告之赐。然就上列各表观察，则外货居十之六七，国货仅十之二三。”④ 这些“洋货”广告中涉及门类众多，包括药品、日化用品、乳制品、谷物、糖果、食物、化肥及杀虫剂、机动车、烟草、化工产品、纺织品、饮品、橡胶制品、机械、摄影器材、电子设备、钟表、乐

① ［美］白瑞华：《中国报纸（1800—1912）》，王海译，暨南大学出版社 2011 年版，第 294 页。

② 苏士梅：《中国近现代商业广告史》，河南大学出版社 2006 年版，第 155—157 页。

③ 王润泽：《中国新闻媒介史（1949 年前）》，北京大学出版社 2011 年版，第 321 页。

④ 戈公振：《中国报学史》，上海古籍出版社 2014 年版，第 230 页。

器、纸制品和办公设备等 30 多个类别。[①] 其生产商则主要来自英国、美国、加拿大、法国、日本、德国、印度、意大利、荷兰、瑞士、新加坡等国。这些公司或自立广告部，或直向报社广告部，抑或间接委托广告社在各大报纸上投放广告，每年花费不菲。以英美烟草公司为例，据统计，其 1924 年在 47 种中、外文报纸杂志（含中文报纸 15 种）上投放广告，总计费用为 131 009.35 元。其中，上海地区中文报纸的广告费用为 100 486.32 元，占 76.7%。[②] 在“洋货”通过广告营销挺进中国市场的同时，国货也开始重视广告的作用，逐渐大规模地在各种媒体上与“洋货”开展了广告战。国货与洋货之战在 1915—1925 年随着中国民族资本主义“黄金时代”的到来而愈演愈烈。以《申报》为例，在此期间日刊登广告中华商广告刊登比例明显提高，个别时期几乎占到了 3/4。[③]

就内容而言，近代报纸广告的变化也具有一定的代表性。第一是类别的丰富，从早期“告白”已渐细化出商业服务广告、娱乐广告、文教类广告和社会广告等，涵盖了日常生活中的方方面面。第二是内容的多样，从单纯的文字发展为文字与插图、摄影的相辅相成、相得益彰，尤其是图像在广告中起到的作用越来越重要。第三是设计的美感提升，专业化、职业化广告人才与艺术家参与到广告字体与图画设计中，使得广告的审美性大大提高，成为一种近代商业艺术。第四是市场营销技巧的广泛应用，心理学、营销学、社会学等现代理论被有意识地用于广告文本中。第五是消费

---

① 此处依据的是“中国商业广告档案库”分类，而非民国时期的实际商品分类。

② 秦其文：《中国近代企业广告研究》，知识产权出版社 2010 年版，第 221 页。

③ 苏士梅：《中国近现代商业广告史》，河南大学出版社 2006 年版。另参见［美］葛凯：《制造中国：消费文化与民族国家的创建》一书中有关“国货运动和反帝抵货运动 1905—1919”和“国货运动和反帝主义抵制活动 1923—1937”的章节。

市场和对象的细化，针对不同的商品有的放矢地进行广告宣传和市场推广。这些都使得广告的商品和文化内涵得到了极大的丰富，成为一种独特的历史记录。

近代广告研究一直是广告学、报刊研究、文化研究、历史研究、社会史研究等人文学科内的一个重要主题。特别是基于晚清以降的广告图像，对中国近代宗教、文化、商业、政治、国家、性别、消费文化、民族认同等问题进行研究颇为多见。比如，陈平原教授以晚清画报为例，对传统中国的“左图右史”与西学东渐之“图像叙事”结盟，进而汇入到以“启蒙”为标识的现代化进程的研究；梅嘉乐（Barbara Mittler）教授从妇女杂志、广告切入，对以上海为代表的近现代中国城市的视觉现代性的研究；安克强（Christian Henriot）教授对摄影技术和照片塑造近代中国形象的研究；葛凯（Karl Gerth）教授对近现代中国的消费文化和民族国家的研究等等；白露教授所开展的关于近代广告中“摩登女性”形象的研究，将广告、中国现代商品社会的出现和本土社会学（借助中国启蒙社会思想来解释日常生活，阐释精英化社会关系与社会实践等的思想或著作）置于特定的历史语境中进行考察，以 Cutex（蔻丹）、BAT（英美烟草公司）、“双美人”牌化妆品、“安迪生”电灯泡、“卜内门”化肥等一系列广告为例，说明商品在现代中国社会是如何与社会进化思想联系起来的，本土社会学又是如何与性感的摩登女性图像建立关系的，广告又是如何将这种对现代性的另类幻化图像自然化的。[①] 这些学者的研究都为近代广告研

① Barlow, Tani. 2008. *Buying In: Advertising and the Sexy Modern Girl Icon in Shanghai in the* 1920*s and* 1930*s*. in *The Modern Girl Around the World*, *ed. The Modern Girl Around the World Group*, 288 - 316. Durham. N. C.: Duke University Press.

究提供了非常具有开拓意义的视角和路径。

然而，近代中国广告的丰富性、复杂性和多样性在提供很多有价值的学术问题的同时，也给研究工作造成了一些困难。这些问题和困难中，有的是当时就已经被认识并讨论、研究了的，比如广告的意义和功效问题；有的是一直遗留至今的，只不过在不同时期对其有不同的认识和理解，比如广告的策略问题；还有的在当时并非问题，但在如今却成为问题，比如广告与现代性问题、广告与民族主义以及广告与身份认同问题等。最后一种情况的出现，往往是因为在新的社会语境和条件中，出现了新的资料或研究方法，改变了整个研究的路径，产生了新的知识，打破了我们之前固有的认知框架。尤其是在当今数字时代语境之下，海量的文字和图像等历史资料及相关研究成果被数字化，为研究者提供了前所未有的丰富信息，这些资源在一定程度上已经开始改变学者的研究方式和路径。

以广告研究为例，普遍是将广告作为“图例”（illustration）来对历史进行补充性描述，尽管不少研究已经开始采用定量研究方法，但一般还是手工抽样统计，并没有进行较大规模的量化统计，更遑论对海量广告图片和内容做进一步的知识挖掘。而近年来，诸多广告/图像数据库的出现可以看作是一种从学者需求出发的新的探索。比如，梅嘉乐教授领衔建立的“晚清及民国早期中国妇女杂志”（“Chinese Women's Magazines in the Late Qing and Early Republican Period”）项目、安克强教授领衔建立的“虚拟上海”（Virtual Shanghai）项目、叶凯蒂（Catherine Vance Yeh）教授领衔开展的“中国小报”（“Chinese Entertainment Newspapers”）项目，都分别对近代中国的妇女杂志、小报、照片等历史材料进行了数字化存储、编目和建档，并对图片中的部分信息进行了标注。这些项目的发展和相关研究的展开都为以图片为对象的广告档案库的建设提供了大量可供借

鉴和学习的经验，[①] 也为下一步更深入地从文本和图像等多重信息入手对广告进行研究提供了条件。

当然，这并不是说有关广告的研究就一定要通过数字化、档案化的方式，运用数字技术或统计方法。正如凯瑟琳·海勒斯（N. Katherine Hayles）在《我们如何思考：转型的力量和数字技术》（*How We Think: Transforming Power and Digital Technologies*）中所说，数字技术对人文学科的渗透已经确实地改变了学者们思考的方式，在研究的范围、批评/生产理论、合作方式、数据库、编码和多模型的学术方法等诸领域都可以看到这一改变的过程，现在是时候去思考如何从数字人文的角度出发对广告及其相关研究进行改变了。[②]

一、广告的“蜉蝣性”与档案库化

尽管学者们已经意识到建立广告基本数据资源的重要性，但事实上，这个过程并不容易。尤其是有关数据资源的获取、存储、建档及展示，简单地说，就是如何对图片进行数字化、档案化，如何对公众开放。

就中国近现代报纸而言，广告基本数字资源的获取并不是非常的便利。这一方面是因为中国近现代报纸的数量非常庞大，其保存地也相对比较分散，从而造成学者在研究的时候，获取相应的资源并不是非常的方便；另一方面是因为尽管国内外很多机构，比如图书馆和一些商业公司

① 在这里先驱性的项目，也对CCAA意义很重要的数据库有两个，一个是“The William Blake Project”（http://www.blakearchive.org/），一个是上海图书馆的知识加工平台。

② Hayles, N. Katherine.. *How We Think: Transforming Power and Digital Technologies*. In *Understanding Digital Humanities*, ed. David M. Berry., 2012. p. 42 - 66.

都对报纸进行了数字化，但使用时大多是要收费的，并且这些数据库的建设主要针对的是报刊上的新闻及评论文章，对广告的内容加工和信息提炼并不是很充分，大部分都只有广告中的一行字，但是没有对图像进行进一步的分析。这意味着广告档案库化的第一步要从数据的数字化开始。

而对报纸的数字化，则必须要兼顾广告这个数据采集对象的特殊性，在此将之称为“蜉蝣性”。“蜉蝣”（ephemera）指只能短暂存活的昆虫，以隐喻“短暂的时间”。相比而言，ephemera 多了一层隐喻，可以指代短暂存在的印刷品，继而成为了一个独立的分类，即“日常生活中琐碎而短暂的文档”（the minor transient documents of everyday life）。而在中国的印刷史和出版史上并没有“ephemera”这样一个特定的分类，尽管从宋代开始就出现了宝卷、符箓、黄历、纸钞等具有蜉蝣性质的短期印刷品，但只被当作一个专属的类别来对待，甚至在中国印刷史中也没有作为一个专题来讨论。[①] 只有到了晚清，石印、铅印技术和机器的引进、造纸技术的改进、印刷行业新型组织形成、新型商品经济和市场发展以及大规模印刷品的出现，技术、资本、社会需求和市场应用由此在时空中接合，诸如报纸、海报、广告这样的短期印刷品才得以进入人们的日常生活，成为英文

① 在《中国印刷史》中，张秀民用“其它印刷品”来称呼所有不能被归为书的印刷品。Lucille Chia 也在 Printing for Profit: The Commercial Publishers of Jianyang, Fujian (11th - 17th Centuries) 中对这个现象进行了描述，指出宋以来的 300 年间，印刷书籍从少到多，但自始至终都缺少印刷蜉蝣——传单、黄历、卜辞、符咒等。这些材料绝大多数彻底消失了，仅有一些因为印在书中而被保留了下来，见该书第 145 页。

意义上的“蜉蝣”。①

从时间和技术的角度来看，一方面，广告是其所处的“现代”这一时期的流行叙事在线性时间上的图释；另一方面，正是这一历史时期使广告变得复杂，并具有普遍性和特殊性，前者体现在蜉蝣性质的广告（日常生活的短暂记录）是某一特殊时间的复杂图示的证据，是历史时刻的闪回；后者体现在这些广告于当代依然具有复杂含义，我们可以将这些广告置于当下的历史条件中进行讨论。②除了历史意义和理论意义外，“蜉蝣”还很好地说明了广告的物质特性，即“蜉蝣性”（ephemerality）。ephemerality 在《韦氏词典》中意指具有“具有蜉蝣性质的事物”，后来也被用于指称短期印刷物的短暂性特点。比如纽约大学保拉·麦克道尔（Paula McDowell）教授在《蛆与昆虫：在 18 世纪英国书写中建构“蜉蝣”分类》（“Of Grubs and Other Insects: Constructing the Categories of ‘Ephemera’ and ‘Literature’ in Eighteenth-Century British Writing”）一文中，以 ephemerality 对 18 世纪歌谣印刷物的特性加以描述。③ 与歌谣单页一样，报纸广告作为短期印刷物广告，就物理特性上而言，不仅存在时间更短（以日为出版周期），而且印刷工艺也较为粗糙，更多地考虑的是其经济价值而非保存价值，因此与空间和时间的关系更为紧密。而且，其内部信息的组织形式也并没有书籍那样的内在空

① 参见 Reed, A. Christopher. *Gutenberg in Shanghai: Chinese Print Capitalism 1876-1937*. Vancouver & Toronto: UBC Press。

② Barlow, Tani. *Advertising Ephemera and the Angel of History*. *positions: asia critique*, 2012, 20: 111-158. p. 112.

③ Murphy, Kevin & O'Driscoll, Sally ed., *Studies in Ephemera: Text and Image in in Eighteenth-Century Print*, 2013-01-30. (Kindle Location 730). Bucknell University Press. Kindle Edition.

间性，而是更为分散，这也为我们建立广告档案库、设计广告的元数据标准提供了非常重要的理论依据和出发点。那么广告的“蜉蝣性”是如何具体体现的呢？

首先，广告的“蜉蝣性”体现在从一开始它就注定是一种商业经济的短暂伴生物，在其原初时间点上并不具备长期保存的价值，而中国印刷传统一向对短期印刷品不重视，所以民国报纸的保存现状难以尽如人意。比如，以“中国商业报纸广告数据库”存档的五份报纸为例，它们虽然都有残缺，但相比较而言，《申报》《大公报》和《盛京时报》因为是全国性大报，所以保存状况比较好，并且在多个图书馆都藏有其缩微胶卷版；而《汉口中西报》和《越华报》作为地方性报纸，则没有这么幸运，经过努力寻找，项目组仅在上海图书馆和北京大学图书馆找到 1907—1935 年的《汉口中西报》，且每年都不全。在广东中山图书馆找到部分《越华报》，其馆藏时间为 1927—1950 年，选取时间为 1927—1938 年，亦是每年都不全。可见，报纸广告档案库难以做到基础数据完整和全面，无法保证时间序列上的连续性。

其次，广告的“蜉蝣性”体现在广告对于媒介、时空和外部信息的高度依赖性，即广告非常依赖于其媒介的物质性，是断裂的、跳跃式的，因此广告内容本身所提供的信息并不完整，需要与广告之外的信息产生关联，构成意义的网络。本质上，报纸广告的内容是外向性的、参考性的，也是超文本性的。以报纸广告为例，所有报纸广告同时也是该报的一部分，因此选择报纸广告图片的时候要考虑到报纸的选择标准，包括报纸的出版地、发行范围（地区还是全国）、报纸的性质（商业、政论还是政治）、发行量（日发行量需要达到本地同类报纸中的较高水平）和历史地位（在当时社会的影响力）等，这些会影响广告的性质以及后期的研究结

果。与此同时，还要考虑到广告的信息完整性以及广告与报纸的关联性。因此，尽管数字化的对象是广告，但还是有必要数字化广告所在的整个版面，而且这部分信息也应成为数据库元数据的一部分。

最后，广告的蜉蝣性还体现在物质性和文本性的问题上。从数字化到建档，广告既是物质对象也是数字文件。那么在数字化存档时，是否要同时保留两者？又该如何处理其物质性和文本性问题呢？在蜉蝣研究中，研究者普遍认为蜉蝣是一种时间胶囊，是伊丽莎白·艾森斯坦所说的现代欧洲早期“印刷革命”中的“文本证据”（proof-text）①，与历史有一种本质的或者近似天然的联系，即使是因为它是被无意识地保存下来的，这更增加了它的珍稀性，提升了它的价值，使之变得更有意义。但在运用蜉蝣材料进行研究的学者那里，蜉蝣作为文本的价值已经高于其本身所具有的物质性价值。尤其是在数字时代，数字化蜉蝣材料已经成为一种“惯例”，学者、图书馆、档案馆都采用数字化方式保存蜉蝣材料。与此同时，数字蜉蝣（digital ephemera）的出现，比如电子邮件、Youtube 上面的各种视频、blog 上面的各种帖子、网上保存的各种图片以及存储在“云端”的各种文件等，都已经重新定义了蜉蝣的概念。就数字广告的存档而言，如果我们认为原真性与广告蜉蝣产生的那个历史时刻的独一无二性有关，与广告的物质性有关，那么广告的商品属性、大规模印刷的技术特性和原始资源的不可接触，已经使得存档意义上的原真性不复存在。那么，我们在讨论广告（更准确地说是数字广告图片）档案库时，就不可避免地要关注广告的文本性（信息）而非物质性（媒介）。在此议题上，数字人文先驱杰

① Randall, David. *Review: Recent Studies in Print Culture: News, Propaganda, and Ephemera*. *Huntington Library Quarterly*, 2004. 67: 457 - 472.

罗米·麦克甘（Jerome Mcgann）教授有关文本性的研究[①]以及 The William Blake 项目数据库[②]和 Rosette Archive 数据库提供了很重要的参考。对于建立一个广告档案库而言，如何数字化对象、如何设计元数据标准、如何展示图片都是重点。

二、“中国商业广告数据库”的数字化与元数据标准

以中国商业广告数据库（CCAA）为例。在数字化过程中，因为考虑到报纸的脆弱性，而藏有纸质版的图书馆都不愿意对其进行数字化，再加上购买原版报纸的难度和巨额费用，所以 CCAA 采取了从缩微胶卷转数字图片的方式。这就又一次受到了缩微胶卷物质性的限制，因为这批报纸的缩微胶卷基本上都是 20 世纪七八十年代制作的，而传统胶卷的质量自然无法与现在的数字技术转拍或者扫描制作的胶卷相比，尽管我们尽量都是从馆藏母带转存数字图片，但依然会看到很多在之前制作时就不尽人意的图片效果。此外，缩微胶卷转数字图片的过程中，还有一个物质条件需要考虑，就是缩微胶卷扫描仪。扫描仪的质量、先进程度和老化程度将直接影响到数字图片的数据完整性与清晰度。在物质条件限制之外，还需要考虑人为因素，比如必须与藏有缩微胶卷的图书馆达成协议，这就涉及人际关系和资金投入问题。此外，还要考虑从缩微胶卷人工转为数字图片的时候，人眼的观看效果与所持标准的差异以及显示器分辨率差异和色差，

---

① 参见 Mcgann，Jeorome. *Radiant Textuality：Literature after the World Wide Web*. Palgrave。Mcgann 教授提出文本应该被理解为由语义部分和图像意义部分共同组成同时当文本文档以语义学的（semantical）和文献学的（bibliographical）两种方式被编码了以后，事实上也就是被以图表的（graphical）方式标记了。

② 参见 http：//www. blakearchive. org/。

这些都为数字化广告图片造成了困难。

最终，CCAA 选择了《申报》《大公报》《盛京时报》《汉口中西报》和《越华报》5 份报纸，分别在上海图书馆、国家图书馆、广东中山图书馆、北京大学图书馆和华盛顿大学西雅图分校亚洲图书馆等地对近 12 万张、不小于 300 dpi 的广告版面图片进行了数字化，并建设了标准平台，对图片进行存储和标注。项目组还花费了大量时间，对档案库的元数据标准进行了三次抽样、测试和完善，元数据从最初的 6 项增加到目前的 44 项。他们先从《申报》和《大公报》进行了部分抽样，设计了元数据框架，然后批量数字化《汉口中西报》，再将元数据框架应用于 2 700 多张数字图片，进行测试和调整。接着对《汉口中西报》（近 10 000 张）和《越华报》（4 719 张）进行了完整的标注，最终确定了方案。在此过程中，项目组先后与上海图书馆、莱斯大学图书馆的元数据管理员、程序员反复讨论，参考了大量已有的文献和数据库，就其中的问题咨询了中国和美国广告学、历史学、新闻学和信息学的学者，并开发了专用于图片标注的工作平台。当然，就目前的方案而言，还有很多地方可以进一步扩展，特别是有关图像的“主题词”部分，现有主题词并不能涵盖广告图像中的所有信息，而这部分对于广告图像的研究有着非常重要的意义。①

目前，CCAA 的元数据标准主要基于 Dublin Core 数据结构，其中包含了“描述性数据”（descriptive data）“语境信息数据”（contextual information data）“文献数据”（bibliographical data）和“技术数据”（technical data）。“描述性数据”主要将对图像的文字和图片内容进行记

① 这部分内容非常值得参考“The William Blake Archive”的元数据分类方案，其中包括了图像中人物的“动作”“表情”“服饰”的元数据。

录，“语境信息数据”主要针对的是商品信息和广告公司信息，“文献数据”则主要是指和报纸相关的信息，而“技术数据”则主要是指数字化过程中与图片有关的数据。通过这种定制化元数据框架的方式，CCAA 试图在最大程度上使广告图像及其相关信息得以用数据的方式保存，并以数据库方式和结构呈现给读者。

数字档案化广告图片仅仅是从数字人文角度开展近代中国广告研究的第一步。事实上，真正的研究还尚未开始。但这样的基础性工作也必须尽快进行，才可能及时保存、保护历史材料，并且从新的视角对历史问题进行再审视，发现新的问题，发掘新的知识。而这一步要求学者们能建立一种基于共享、合作的知识生产方式，目前很多非常成功的数字人文项目已经提供了很好的经验，希望能有更多的学者加入其中。