# 数字学术，数字课堂

## 研究与教学的全新国际视野

大英图书馆Gale数字人文日会议文集

2019年5月2日

GALE
A Cengage Company

# Digital Scholarship, Digital Classrooms

## New International Perspectives on Research and Teaching

# 数字学术，数字课堂

## 研究与教学的全新国际视野

2019 年 5 月 2 日大英图书馆 Gale 数字人文日会议文集

GALE
A Cengage Company

GALE
A Cengage Company

数字学术，数字课堂：
研究与教学的全新国际视野

本书翻译自：
Digital Scholarship, Digital Classrooms: New International Perspectives on Research and Teaching

ISBN：978-0-4182-9146-7 (print)



更多产品信息或技术帮助，请与我们联系：
GaleChina@cengage.com



**圣智学习（北京）教育科技有限公司**

北京市海淀区科学院南路 2 号融科资讯中心 C 座南楼 707 室，邮编：100190
电话：+86 10 8343 5000
传真：+86 10 8286 2089
邮箱：GaleChina@cengage.com
**www.gale.com**

# 目录

前言，Seth Cayley .....vii

论原则和价值：挖掘 1785-2010 年伦敦
《泰晤士报》的保守言辞，Joris van Eijnatten .....1

对挖掘十九世纪新闻数据的基础设施的反思，
Julianne Nyhan、Tessa Hauswedell 和 Ulrich Tiedau .....27

人文数据分析教学，Ryan Cordell .....39

将数字人文引入本科生课堂：策略、方案及
利用 Gale 数字学术实验室的教学实践，Sarah L. Ketchley .....49

作者介绍 .....65

# 前言

**Seth Cayley，Gale 原始档案项目副总裁**

“创始神话”是技术行业不可或缺的一部分。每一家成功的公司都有自己的传奇故事，它如何无意中诞生或忽然间豁然开朗，亦或它的创始人如何住在集装箱中将其从零发展壮大。

创始神话著名的例子有史蒂夫·乔布斯在车库里成立了苹果公司、比尔·盖茨从哈佛退学创立了微软。将它们称之为“神话”并非因为他们是虚假的，而是说他们简化了复杂的历史，忽略了其他人和机构在其中起到的重要作用。

苹果公司的联合创始人史蒂夫·沃兹尼亚克自己也承认“车库比其他任何地方都更好的代表了我们，但我们并没有在那做设计……很少有超过两个人同时呆在车库里，而且大部分时间他们只是坐在那，没做什么有意义的事情。”[1] 同样，围绕比尔·盖茨的退学故事掩盖了他（及他的联合创始人保罗·艾伦）优越的出生背景，在美国最著名的一所私立高中上学，很幸运在此就接触到了计算机。

这些神话的结果是放大了天才个体（通常是白人男性）的理想典范作用。

数字人文也有自己的创始神话。1949 年前后，耶稣会信徒罗伯托·布萨（Jesuit Roberto Busa）联系到了 IBM 的 CEO 托马斯·沃森（Thomas J. Watson Sr.），说服他资助自己的《托马斯著作索引》（Index Thomisticus），这是布萨设计的一个工具，能让用户对圣托马斯·阿奎那（St. Thomas Aquinas）的全部作品开展文本分析。在布萨自己的叙述中，就像大卫和歌利亚的故事一样，谦恭的耶稣会教士说服了资本主义的大主教支持一个纯学术的研究项目。[2] 从这样一个匪夷所思的故事开始，“人文计算”运动诞生了。根据 Julianne Nyhan 的观察，这个弱势者的故事帮助数字人文界团结在一起，区别于主流的人文研究领域。[3]

并非要诋毁布萨无可置疑的成就，但布萨的创始神话确实存在一些问题。学者提托·奥兰迪（Tito Orlandi）指出，其他从事翻译和考古学研究的学者也应被认为是数字人文的先驱者。[4] 同样，布萨并非独自一人工作。即将完成的一项研究表明，他在 20 世纪 50 年代和 60 年代，依靠一队穿孔卡片操作员（女性）让他的愿景成为了现实。[5] 只着眼于布萨，我们忽略了团队合作对他取得成功所发挥的重要作用，也忽略了推动数字人文发展的国际学术环境。

2019 年的数字人文已经更加多元化，远超出了布萨创始神话所蕴含的意义。这一领域强调合作、跨越学科界线，并且具有真正的国际视野。在本文写作时，追溯历史报纸中全球信息网络的研究项目“海洋交流”（Oceanic Exchanges）有 39 位研究者参与，他们分别来自 6 个国家。[6] 近期的国际数字人文组织联盟大会在不断努力邀请来自南半球的参会者，2021 年该联盟在亚洲的第一次大会将在东京召开。[7]

正如 Nyhan 所说，布萨的神话可能是“有用的虚构”，[8] 但数字人文正在超越它的边界，融入主流。值得注意的是，该领域内大量的对话围绕如何开展数字人文教学。在过去的两年中，不断有论文发表和图书出版，探讨如何将数字人文整合到课堂与课程中。[9]

我个人从 2012 年开始涉足数字人文，当时我为 Dallas Liddle 主持的一个研究项目提供支持，这个项目对《泰晤士报》上数千篇的社论进行遥读。通过分析字数变化的规律，Dallas 的研究挑战了长

期以来人们认为维多利亚时代报纸发展缓慢的观点……无需他实际阅读任何一篇社论。[10] 这一研究项目也挑战了我对“历史”是什么的想象。文本可以作为数据来处理。我被吸引了。

自那时起，Gale 经历了文化的转变，让潜藏在 Gale 原始档案下的数据更能为全球范围内的这类研究项目所用。我们已经为多个数字人文项目提供了支持，看到了各种迥然不同的想法，例如都铎王朝的权利网络（Ruth Ahnert 主持 [11]）、十八世纪印刷商装饰图案（花形装饰图案，Hazel Wikinson 主持 [12]），并且追溯了十九世纪英语母语世界中新闻再版的方式（剪刀与粘贴，Melodee Beals 主持 [13]）。2018 年我们发布了 Gale 数字学术实验室（Gale Digital Scholar Lab），一个基于云存储的文本和数据研究环境，设计的初衷就是要降低各个层级上研究者开展数字人文研究的入门门槛。[14]

因此，此时正是一个大好时机，将学者、图书馆员和学生汇聚一堂，从来自全世界各地的演讲者那里了解数字人文的最新研究项目和研究趋势。《数字学术、数字课堂》精选 2019 年 5 月 2 日大英图书馆 Gale 数字人文日活动上展示的八篇论文。

这一活动分为四个环节：（1）文学与遥读，（2）计算机读新闻，（3）教室中的数字人文，和（4）院校的数字人文支持与基础设施。每一环节都旨在突出数字人文中的不同主题，前两个环节围绕最新研究项目，而后两个环节探讨教学问题及图书馆的作用。

在本卷中，你将会看到活动中展示的四篇论文。

在第一篇论文“论原则和价值：挖掘 1785-2010 年伦敦《泰晤士报》的保守言辞”中，荷兰乌特勒支大学的文化史教授 Joris van Eijnatten 呈现了一个教科书式的案例，如何使用 N 元语法和其他文本分析工具实现全新且有趣的发现。在他对道德语言和历史报纸的研究中，Joris 探讨了“《泰晤士报》数字典藏”中保守言辞的发展。他记录了在政治新闻报道中特定词组的出现和消失，例如从“保守主义原则”变为“保守主义价值”。将这些发现置于历史背景中，他注意到“从 20 世纪 60 年代后政治观点开始明确的蕴含在道德词语中，”而新的修辞方式则“出自现代化带来的紧张关系”，这种关系的形成因 20 世纪 60 年代的放任型社会而尤为明显。

不配备相应的基础设施，开展这类研究是不可能的。在“对挖掘十九世纪新闻数据的基础设施的反思”中，英国伦敦大学学院数字人文中心的副主任、数字信息研究副教授 Julianne Nyhan 探讨了她的团队尝试从零开始建立大型文本挖掘项目的经验。她强调，即便研究者已有大量的数据并且身处资源丰富的院校，例如伦敦大学学院，在能够开始分析数据前仍有一系列出乎意料的挑战。尽管有来自大学 IT 研究服务小组的支持，Julianne 仍指出“计算设施并非真正为支持人文研究而建立，”而且每天的成本高达 350 英镑，日复一日的数据加载耗尽了项目预算。她在论文结尾建议大学应当如何应对“数字转变”，以及公私合作可能发挥的作用。

Ryan Cordell 论文的主题是“人文数据分析教学”。Ryan 在美国西北大学同时开设研究生和本科生课程，在本文中深入反思了他自己的教学理念。[15] 用他的话说：“我作为教师的一个首要目标是反对‘数字原生代’的观念以及由此而产生的毁灭性的教育学后果……我相信未来几年中最迫切的学术问题将需要真正的跨学科视角。我并不是说学者自己需要涉足多个学科领域，而是一组学者能够贡献他们各自的专长，形成持续而坚固的合作体。我正是希望我的学生能够具备创新精神和无限活力去迎接这样的未来。实现这一点不仅仅需要建立在某些技术或工具基础上的教学法，也需要丰富的学术和技术想象力。”Ryan 的论文拓展了这些想法，探讨了我们不仅仅需要传授编程技巧，而是要传授思考数据的思维方式——用他的话说就是“编程式的思维”。

在课堂主题的另一个演讲中，美国华盛顿大学的埃及古物学家 Sarah Ketchley 探讨了她自己在数字人文入门课程中指导本科生的经验。Sarah 阐述了她所面对的学习目标和教学难题，以及她用于克服这些困难的方法学和技术解决方案。以她 2018 年秋季课程为研究案例，她解释了她的学生如何使用 Gale 数字学术实验室（Gale Digital Scholar Lab）整合他们的数据集，使用得到的分析结果建立令人印象深刻的数字展览。Sarah 在演讲最后反思了教学的成果以及数字人文教学如何帮助学生培育可用于大学以外生涯的技能：“这些学生能够清晰地看到，这些他们正在课堂上锻炼和使用的特殊技能与他们就业的市场竞争力密切相关。”

当天的其他演讲者包括：美国斯坦福文学实验室主任 Mark Algee-Hewitt，他概述了该实验室的“微体裁”项目，探讨小说家吸纳其他学科的方式，例如将哲学、历史和自然科学与叙事相结合；日本大阪大学语料库语言学副教授田畑智司，他阐述了自己如何使用计量文体学分析十八世纪和十九世纪文学作品，并向听众们揭示了查尔斯·狄更斯最喜欢用的词组是“他的手揣在兜里”（his hands in his pockets）；英国拉夫堡大学数字历史讲师 Melodee Beals，他探讨了对文化遗产和历史数据出乎预料的利用；以及澳大利亚悉尼大学访问服务主任 Lisa McIntosh，她阐述了她所在的大学图书馆正在如何发展基础设施以支持最新研究。尽管他们的论文没有发表在本卷中，但建议你自行搜索他们的作品。

感谢所有演讲者分享他们的研究项目和想法。

## 注释：

1. https://www.washingtonpost.com/news/morning-mix/wp/2014/12/05/steve-wozniak-refutes-apples-creation-story-the-garage-is-a-bit-of-a-myth/.
2. R. Busa, “The Annals of Humanities Computing: The Index Thomisticus,” Computers and the Humanities 14, no. 2 (1980), 84.
3. Julianne Nyhan and Andrew Flinn, Computation and the Humanities: Towards an Oral History of Digital Humanities (Springer International Publishing, 2016), 270.
4. Julianne Nyhan and Andrew Flinn, 79.
5. Julianne Nyhan and Melissa Terras, “Uncovering ‘Hidden’ Contributions to the History of Digital Humanities: The Index Thomisticus’ Female Keypunch Operators,” Digital Humanities 2017 (Montreal, QC, Canada). See also the authors’ paper presented at DH2017: https://discovery.ucl.ac.uk/id/eprint/10052279/9/Nyhan_DH2017.redacted.pdf.
6. Oceanic Exchanges Project Team. Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914. 2017. DOI 10.17605/OSF.IO/WA94S. Available at osf.io/wa94s.
7. List of ADHO conference locations available at http://adho.org/conference.
8. Julianne Nyhan and Andrew Flinn, 80.
9. For example, Claire Battershill and Shawna Ross, Using Digital Humanities in the Classroom: A Practical Introduction for Teachers, Lecturers and Students (Bloomsbury, 2017), and Anna Wing-bo Tso, editor, Digital Humanities and New Ways of Teaching (Springer, 2019).
10. Dallas Liddle, “Reflections on 20,000 Victorian Newspapers: ‘Distant Reading’ the Times using The Times Digital Archive,” Journal of Victorian Culture 17, no. 2 (June 2012), 230–237.
11. https://ahrc.ukri.org/research/case-study-archives/tudor-networks-of-power-1509-1603/.
12. https://fleuron.lib.cam.ac.uk/.
13. http://scissorsandpaste.net/.
14. https://www.gale.com/intl/primary-sources/digital-scholar-lab.
15. https://ryancordell.org/statements#teaching.

# 论原则和价值：挖掘 1785-2010 年伦敦《泰晤士报》的保守言辞

**Joris van Eijnatten**，荷兰乌特勒支大学文化史教授

j.vaneijnatten@uu.nl

**摘要**：本文探讨了 1785 年至 2010 年伦敦《泰晤士报》保守言辞的本质。报纸撰稿人有意采用了哪些语言表达可以被称之为“保守主义”的道德观念？通过检查常用词簇追溯历史上的这类道德语言，它们反过来又折射出这一研究中的“保守”思想。词簇更迭频繁，通常长度较短，非常适合于对较长的一段时期进行计算机辅助分析。方法学上，本文利用了两种经证实的、简易的文本挖掘技术：N 元语法（Ngrams，特别是二元语法）和词语嵌入。研究追踪了整个时期内的二元语法词组，其中最重要的是“conservative principles”（保守主义原则）、“conservative values”（保守主义价值）、“traditional values”（传统价值）和“permissive society”（放任型社会）。

“保守主义原则”一词被当做一种道德和政治表达，特别是在十九世纪和二十世纪前半叶。保守言辞的重大变化主要发生在 20 世纪 60 年代期间和之后。新的词语开始用于表达道德立场，特别是所谓放任型社会出现之后。但与此同时，识别一种特殊的保守言辞也变得越来越困难。首先，传统与现代之间的紧张关系是保守主义必定要强调的，也成为了更大范围内公共辩论的一部分。其次，所有的道德语言都变成了“道德说教”，事实上，很多的语言中都充满了像“价值”、“传统”这样的词语。最终形成了左翼和右翼言辞的融汇，意识形态差异的共存共生。

**关键词**：保守主义、道德语言、伦敦《泰晤士报》、数字历史、原则、价值

## 保守言辞

1997 年 9 月 25 日，迈克尔·波蒂略（Michael Portillo）——1995 年起任约翰·梅杰（John Major）政府国防大臣但刚刚失去议会议席的一位保守主义政治家，为《泰晤士报》撰写了一篇书评。他探讨了《保守主义死去了吗？》（Is Conservatism Dead?）一书，这本书是由两位保守主义知识分子——哲学家约翰格雷（John Gray）和下议院议员大卫威利茨（David Willetts）共同撰写的。文中波蒂略引用威利茨的话“布莱尔先生在刻意逢迎‘价值’，保守派是有‘原则’的”。[1] 显然，威利茨已经发现了保守党和工党（或至少他们中的布莱尔派）之间政治言辞的差异。左翼谈论价值，右翼谈论原则。威利茨是否真得曾经这样说仍存争议。但有趣的是，保守主义与“原则”一词之间的历史关联事实上的确存在。

《牛津英语词典》定义“原则”（principle）一词为“作为一种行为准则被采纳和承认的通用法则或规则，行为或实践的固定基础或根据，行动的基本动机或理由，特别是有意认可和遵循的理由。”[2]《牛津英语词典》的一个例句是极端保守主义者本杰明·迪斯雷利（Benjamin Disraeli）的小说《康宁斯比（年轻的一代）》（Coningsby，1844）中的一句话：“在我支持保守主义原则前……我仅仅想要知道这些原则的目标是要保存哪些东西。”[3]“原则”对于十九世纪的他们有着完全不同的感觉，[4]遵循一项原则意味着毫不含糊地秉持坚定的道德信念做事情。原则相互之间互为对照，让不同行为方式的可能结果更为明晰，一些行为可能是正确的，一些行为可能是错误的，为道德规范提供了一个起点。迪斯雷利在《康宁斯比（年轻的一代）》中使用词组“保守主义原则”（conservative principles）不下十三次，有一次将其与“让步原则”（concessionary principles）相对比。[5]后者是错误的，因为一个人不应向原则让步，让步原则导致的行为只能是不幸的。

本文探讨 1785 年至 2010 年之间《泰晤士报》对保守言辞的使用。[6]撰稿人有意采用了哪些语言表达可以被称之为“保守主义”的道德观念？[7]《泰晤士报》一直以来被人们认为是中间偏右立场，因此这一资源以刻板印象描述保守言辞的可能性较低。当然，刻板印象的情况也有发生，例如在 2001 年的一封读者来信中，一位读者讽刺地指出著名的保守主义者伊恩·邓肯·史密斯（Duncan Smith）不仅仅“‘绝不’使用欧元”而且“赞成体罚、绞刑和猎狐”。[8]本文也考察了保守言辞（与保守派或保守主义的言辞不同）作为一种“道德语言”，也就是在尝试论证社会和政治行为时的一种假定修辞形式，对什么是好的、什么是坏的做出基本判断。这种充满道德含义的语言的重要性很少被低估，因为他们影响着决策并鼓舞着行动，最终形成了我们所知的这个世界。

保守言辞是极有帮助的、通向过去语言的入口，但仅是众多入口其中之一。在后文中，我并非主要对保守主义本身感兴趣，或想要论证保守主义是统一的政治语言，即埃德蒙·伯克（Edmund Burke）两百多年来思想的传承。[9]我首先且最感兴趣的是构成历史上道德语言的词簇。这些词簇更迭频繁且常常长度较短，使得它们成为了大众文化基因或声音记忆。因此它们指向了更大的语义场，储存了有着标准内涵且随时间发生了复杂变化的思想与信仰。因为这些文化基因是更迭的，所以可以使用计算机追踪它们。本文充分利用了计算机辅助分析，但我主要着眼于内容而非研究方法。接下来的核心部分是两种经证实的、简易的数字技术：N 元语法（Ngrams，特别是二元语法）和词语嵌入（word embeddings）。同时，数字化分析结果被放在了对原有资料传统精读的背景之下。注释中简单概述了研究方法。

## 保守主义原则的衰落

“原则”一词在保守党历史上显然非常重要，我们可以将其视为保守言辞的第一个例子——显而易见的起点。为了呈现二元语法“conservative principles”（保守主义原则）在历史上的使用，我们可以简单利用这一二元语法在《泰晤士报》中出现的次数统计（图 1）。[10]

显然，这个词在该报两个多世纪的报道中持续被使用。考虑到该报倾向于报道政治事件，大部分的使用场景应当与政治新闻相关。我们可以通过比较《泰晤士报》与 1800 年至 2005 年的英国议会辩论来验证这一点（图 2）。

这一二元语法在上议院出现的次数多于在下议院，但出现规律与《泰晤士报》非常相似。十九世纪的议会议员特别青睐这个词，规律显示出一些起伏，但自 1850 年后呈稳定下降趋势。

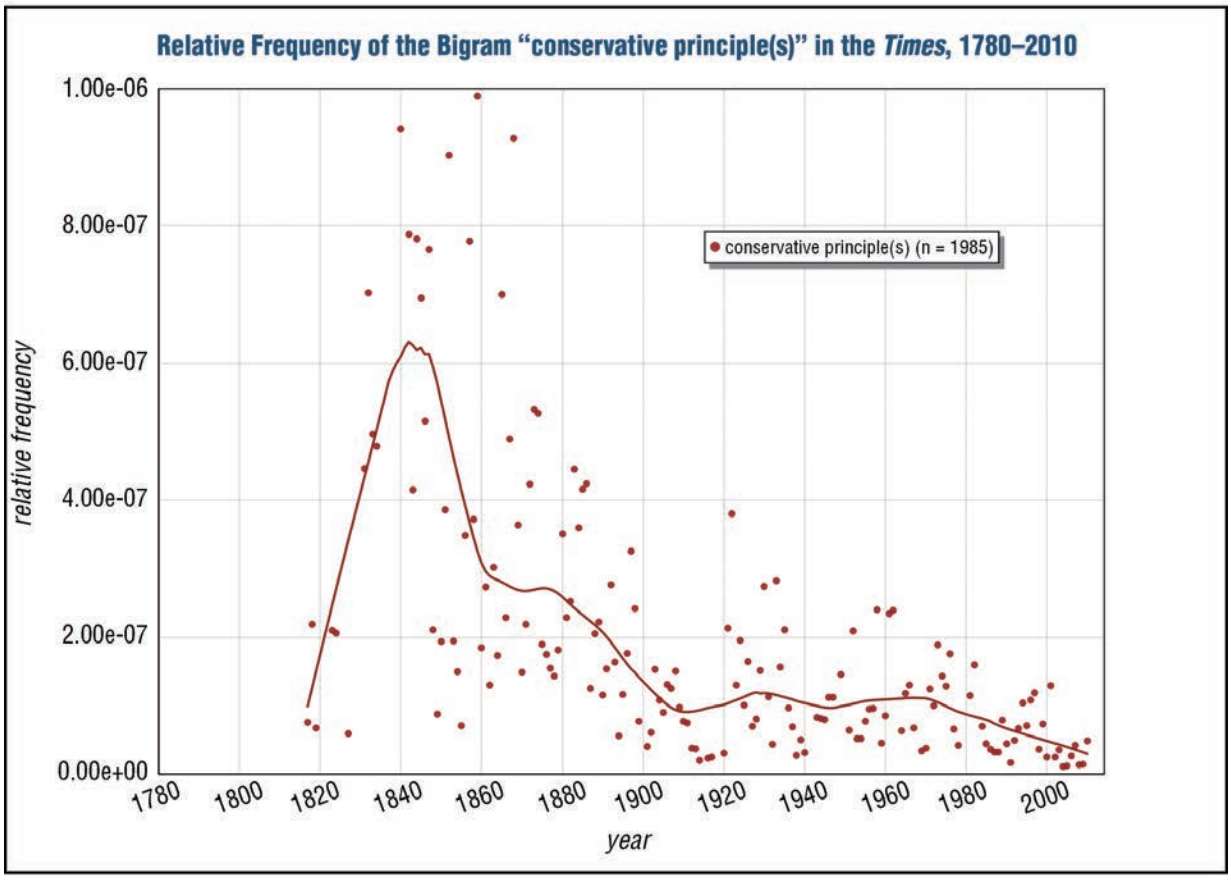



*图 1. 1780-2010 年《泰晤士报》中二元语法“conservative principles”（保守主义原则）的相对出现频率，n=1,985。相对出现频率是每年出现次数与当年词语总数的相对数值。曲线是根据局部多项式拟合函数（loess）计算的“平滑”曲线，便于我们更好的识别规律。*

那么《泰晤士报》在 1785 年至 2010 年间提及 1,780 多次的“保守主义原则”到底指什么？总体规律在一定程度上与基本法的通过有关。图 1 和图 2 中的曲线高点或多或少对应于 1832 年、1867 年和 1884 年的《改革法案》以及 1870 年的《教育法案》和 1919 年的所谓蒙太古 - 切姆斯福德改革（Montagu-Chelmsford Reforms）。换言之，当自由改革提上日程时，保守主义原则就会重被提起。1820 年至 1850 年间这一二元语法较高的出现频率是由于对参加议会议员竞选的候选人的报道。他们可靠的品性常常在讲台致辞和晚宴演讲中被肯定，演讲者保证候选人严格遵守“保守主义原则”。这类新闻报道在十九世纪后半叶逐渐消失。

“保守主义原则”一词最常用在政治语境中。仅在少数情况下，它在《牛津英语词典》中“产生或决定特定结果的基本要素、力量或法则”的含义被用在文化和社会语境中。[11] 例如在 1907 年，人类活动被描述为“两种互补原则的结果”：模仿——保守主义原则，以及创造——进步主义原则。[12] 在 1930 年，牛津大学历史学家及知名保守主义思想家基思·法伊林（Keith Feiling）尝试将保守主义意识形态归结为某种形而上学。根据法伊林的说法，保守主义原则的基础是一种连续性，体现在“持续的实体力量”以及“内在的精神品质”。因为国家力量的形式各有不同，所以“不同的品质而非整体的体积，才是国家安康的保守主义标准”。保存这种品质差异的一个方法是通过保存阶级，因此必须保

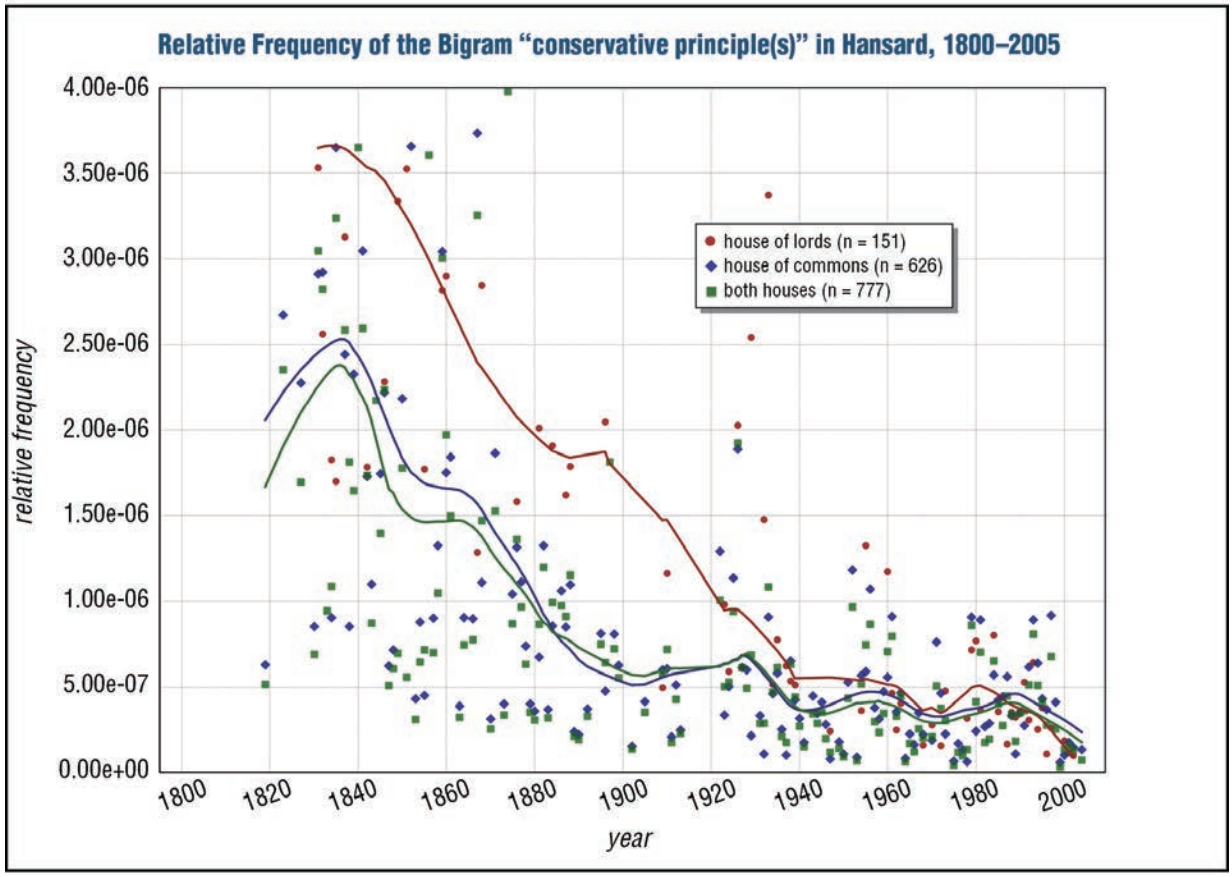



*图 2. 1800-2005 年《英国国会议事录》（Hansard）中二元语法“conservative principle(s)”（保守主义原则）的相对出现频率，n=777。相对出现频率是每年出现次数与当年词语总数的相对数值。*

护阶级“不被普遍的民主所取代和消灭”。尽管法伊林对差异和不平等的偏爱，他仍强调了保证所有人福祉的家长式责任。保守主义区别于社会主义的地方并非对集体的关心，而是它的宗旨，并非平等而是“生命的持久价值”。[13]

1785 至 2010 年，这一二元语法在《泰晤士报》中的最严格政治用途大致可分为三个阶段。1830 年以前，“保守主义原则”一词不常出现，而出现时则是指革命思想和行动语境下的外交政策。1823 年的新闻报道称（法国的）“合法性保守主义原则”和（西班牙的）“社会秩序的保守主义原则”。[14] 这样的表述显示出早期保守主义思想的中立和反对革命的主旨。的确，“法律和秩序”的思想仍然是核心的保守主义信条。它基本上排斥了所有提出快速、激进变革的意识形态，从法国大革命和十九世纪自由改革到共产主义和工党社会主义。

第二阶段（1830-1950 年）也在一定程度上与改革和改革运动更加负相关，比与革命的负相关性更甚。一位评论家简化了保守言辞，定义其原则为仅仅是“否定任何全面的改革”。[15] 对当时保守言辞的更多评判是将这些原则总结在四个相关标题下：王位（throne）、宪法（constitutional）、教会（church）和帝国（empire）。法律和秩序是核心信条，通常处于光荣革命（1699-1689 年）期间达成的宪法协议的庇护之下，而光荣革命正是大部分保守派想要保存的。在 1841 年，《泰晤士报》将“保

守主义的主要原则”保护起来，形容它是“对古老宪法权利制衡的保存”。这一原则的目的是“维护所有阶级和构成立法权的社会各界的完整与行动自由，各自在自己的界线内，任何一个阶级都不会受到其他阶级的侵犯，也不会令整个阶级与人民大众相脱离。”[16]

这个时代被称之为本杰明·迪斯雷利的“一国保守主义”（one-nation conservatism）——这个词组在 20 世纪 70 年代由《泰晤士报》无意间开始定期使用。[17] 这是一种合作国家的思想，其中国王、贵族和平民保持着完美的平衡，每一片领地行使自己的权利和特权，造福整个国家。论据有时利用了对历史或国家身份的参照。这一“古老的宪法原则及对其的实践”基于“众多个体的共同智慧”。[18] 它保存了“法律、机构、用法、惯例和风俗的构成体系，推动塑造和形成了英国人的特性”。[19] 立宪主义者和法学家共同向其中加入了更多的保守主义原则，即教会和国家的联合以及维护大英帝国的完整性。

在所有这些中宗教发挥了重要的作用。一位教士在 1838 年伯明翰的一次保守派晚宴中提出，社会最佳的保存方式是教育和启蒙穷人，传授他们“道德责任的课业，具备健全愉悦的精神，一种基督教精神而非奴性的服从”。这将让不幸的人们“作为有价值的社会一员站在分配给他们的位置上，成为基督教会的感恩之子”。[20] 这样的家长作风是在女性参政权受限制的情况下由代表表达意见的必然结果。因此一位评论家将保守主义原则定义为“授予某个坐在王位上的人绝对的权利”以及“几百个特权家庭……拥有几乎同样的绝对权力”。[21] 当然，十九世纪保守派真正的意思是“不同阶级和不同利益群体的公正代表”，而非将“对国家的控制给予数量居多的人群”。[22]

在第一次世界大战之后，《泰晤士报》中的保守主义原则特别对立于社会主义和工党。在保守言辞中“同志关系的保守主义原则和所有阶级共同的利益”与“社会主义者想要将同志关系的表达限定在一个阶级中”形成对照。[23] 一群保守主义“死硬派”将帝国包含在他们相当标准的原则列表中，大英帝国在全世界的存在让英国确保“与我们相比欠发达的人们享受到正义和文明政府的无价礼物”。[24] 当时较为重要的保守派领袖斯坦利·鲍德温（Stanley Baldwin）也同样重述了四条“永恒原则”：宗教、王位和宪法、帝国及人民福祉，即迪斯雷利曾多次主张的原则（现在由樱草会提倡）。[25]

第三阶段（1950-2010 年）见证了二元语法“conservative principles”（保守主义原则）使用的衰落，表明政治领袖正在寻找一些方式，赋予这一表达新的含义。温斯顿·丘吉尔仍然非常传统地讨论了“供需法则的自由和灵活运用”，认为这一原则应当与另一项原则相平衡，即“同情和帮助那些无论是因为年龄、疾病还是不幸而无法跟上社会前进步伐的人们”。[26] 但在爱德华·希斯 1965 年的“英国权利优先”项目中，变革和改革完全被主观化，家长作风表面看来被放在了一边。大英帝国正在衰落，而社会阶层已经变得不太重要。在这种情况下，认识到“英国在现代世界中的真实地位”将是非常明智的。现在被认为是保守主义原则的思想是“人类个体”是“变革、现代化和改革的主要动力”。[27]

20 世纪 70 和 80 年代个人主义仍风头正劲，但也出现了朝向文化保守主义的重新定位。随着新右派的出现，二元语法“conservative principles”（保守主义原则）的含义开始弱化，不再是政治团结的呼声。很明显，“里根主义”（和撒切尔主义）为首的新自由主义“强有力的国际领导力、自由贸易、缩编政府、降低说率、撤销管制和文化保守主义”现在都被认为是一套“新保守主义原则”。[28] 然而，这段引述来自于 20 世纪 90 年代，此时里根和撒切尔夫人都已经退出政坛。总体上，这些岁月见证了美国道德多数派的兴起以及老布什治下右翼政治的转变。老布什提出的“赋能美国”（Empower America）项目被视为替代了帕特·布坎南（Pat Buchanan）倡导的“孤立主义者、保护主义者的‘古老保守主义’”。[29]

时代变了，原则远去了。20 世纪 90 年代约翰·梅杰试图复兴传统的“一国保守主义”，而迈克尔·波蒂略建议保守党领袖用“同情心和宽容心”“重新凝聚”保守党。[30] 这是撒切尔新自由主义败落的明显信号。读者来信证实了一些保守主义思维的读者对撒切尔政治抱持的疑问。“长久以来保守党都被自由市场信条所劫持，受到美国大学享受终生职位的学者们的深刻影响”，一位《泰晤士报》的读者写到。他嘲讽一种经济学说显然更多归因于“十九世纪的自由主义，而非任何已知的保守主义原则”。保守党所需的是“富于同情心的、灵活且包容的政治，保持最佳的保守主义传统”。[31] 如果《泰晤士报》中日益混淆的词语用法还可用于参照的话，20 世纪 90 年代的保守主义者们需要一种新的措辞。

## 从原则到价值

直到第二次世界大战，二元语法“conservative principle”（保守主义原则）一直是非常常见的措辞，为保守主义发言的人们偏爱这个词，保守主义的评论家也认可这种表达方式。但是否保守主义作家有其他的方式阐明基本的信仰并表达出类似的道德意图？是否有与二元语法“保守主义原则”相近的词组正在被使用，其中表示意识形态（例如“保守主义”）的形容词用于限定表示基本事物的名词（例

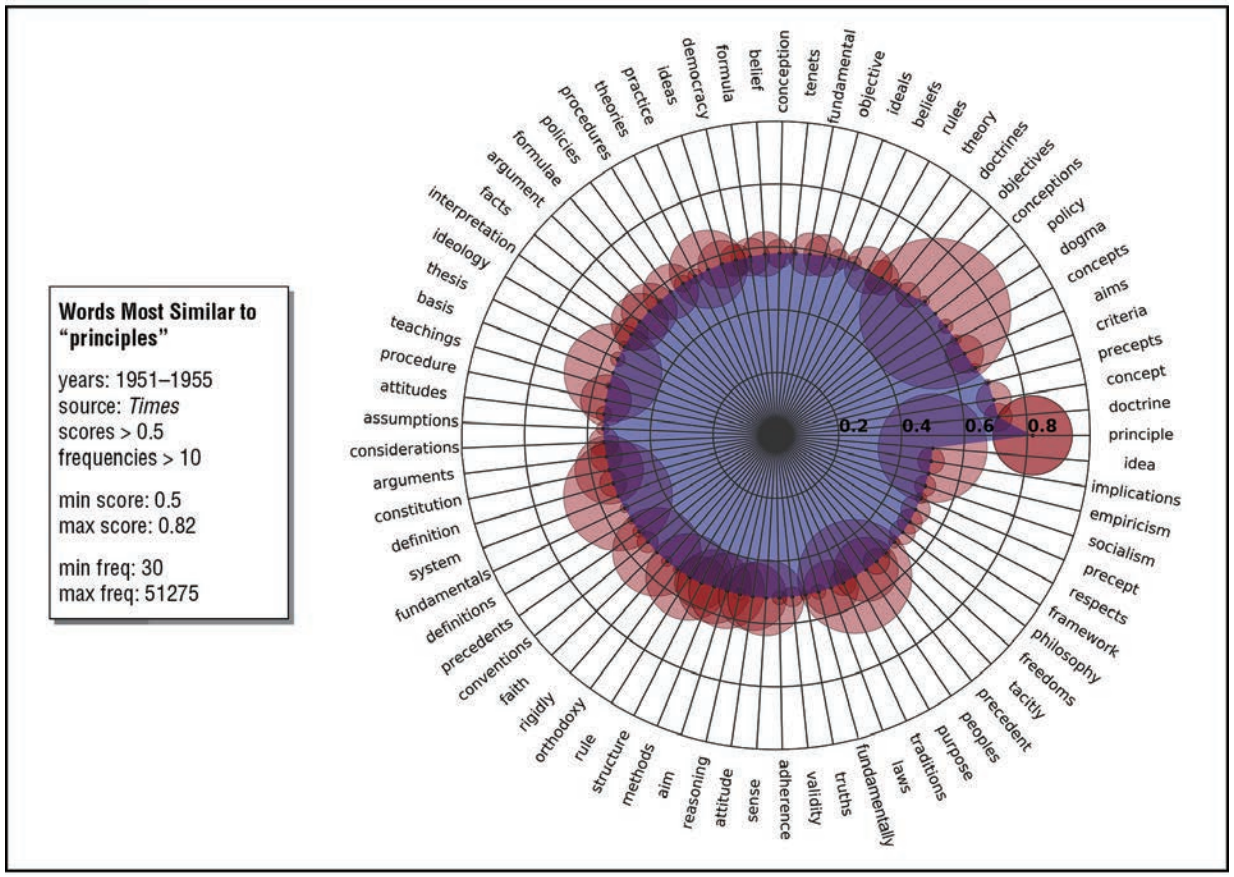



*图 3. 与“principles”（原则）最相近的词语，基于《泰晤士报》（1951-1955 年）的一元语法嵌入。径向图表明每个词相对于“原则”一词的相似度分值，所有数值等于或大于 0.5，分数越高，相似度越大。红色圆圈代表每个词在所探讨的年代内在数据集中出现的次数，出现次数小于 30 的词并未包含在内。*

如“原则”）？我们可能预期“conservative beliefs”（保守主义信仰）或“conservative tenets”（保守主义信条）都曾被使用，但在我们的研究中是否的确如此呢？

理想上，我们应当能够利用二元语法嵌入自动探测到类似于“保守主义原则”这样的词组。[32] 然而 2001 年至 2005 年“保守主义”最常出现的同义词是“winning elections”（赢得选举），而二元语法“hard-line minority”（强硬的少数派）和“tasted blood”（尝到甜头）紧随其后。尽管明显与政治相关，这些相当不具特定性的结果表明，二元语法“保守主义原则”仅仅是出现的次数不够多，使得词语嵌入无法得到有意义的结果。1876-1880 年（“保守主义原则”二元语法最常出现的时期）与“保守主义原则”最相近的二元语法则更有意义，出现了“liberal principle”（自由原则）和“political convictions”（政治信仰）这样的词组。后者似乎更有希望，但“conservative convictions”（保守

**Top Twenty Words Most Similar to “principles”**

| 1851-1855 | 1901-1905 | 1951-1955 | 2001-2005 |
|---|---|---|---|
| doctrine | doctrine | doctrine | tenet |
| basis | policy | concept | doctrine |
| theory | idea | precept | belief |
| truth | proposition | criterion | concept |
| dogma | notion | aim | value |
| system | theory | dogma | norm |
| policy | tenet | policy | framework |
| opinion | axiom | conception | philosophy |
| tendency | method | objective | ideology |
| tenet | conception | theory | ideal |
| maxim | formula | rule | ethos |
| notion | ideal | belief | constitution |
| dictate | belief | ideal | policy |
| interest | shibboleth | fundamental | precept |
| Idea | dictate | tenet | concept |
| consideration | tradition | formula | approach |
| wisdom | dogma | idea | objective |
| reasoning | usage | practice | purpose |
| rules | consideration | procedure | rule |
| assumption | teaching | argument | notion |

*表 1. 与“principles”（原则）最相近的 20 个词，基于四个时期的词语嵌入，复数形式造成的重复已经排除。*

**Number of Occurrences of Most Similar Words Preceded by "conservative" for All Years (1785–2005)**

| bigram | occurrence | range |
|---|---|---|
| conservative policy | 5407 | 1830-2010 |
| conservative interest | 2340 | 1830-2010 |
| conservative basis | 1257 | 1840-2010 |
| conservative element | 778 | 1845-2010 |
| conservative opinion | 684 | 1835-2010 |
| conservative approach | 368 | 1930-2010 |
| conservative attitude | 366 | 1880-2010 |
| conservative rule | 352 | 1840-2010 |
| conservative tradition | 234 | 1850-2010 |
| conservative tendency | 192 | 1830-2010 |
| conservative value | 216 | 1910-2010 |
| conservative method | 155 | 1885-2010 |
| conservative philosophy | 151 | 1930-2010 |
| conservative aim | 118 | 1880-2010 |
| conservative idea | 117 | 1855-2010 |
| conservative legislation | 91 | 1840-2010 |
| conservative system | 87 | 1830-2010 |
| conservative assumption | 80 | 1930-2010 |
| conservative doctrine | 80 | 1830-2010 |
| conservative ideal | 78 | 1890-2010 |

| bigrams | occurrence | range |
|---|---|---|
| conservative law | 70 | 1830-2010 |
| conservative convention | 60 | 1865-2010 |
| conservative faith | 56 | 1880-2010 |
| conservative argument | 53 | 1885-2010 |
| conservative belief | 49 | 1925-2010 |
| conservative standard | 39 | 1830-2010 |
| conservative practice | 35 | 1930-2005 |
| conservative ideology | 30 | 1970-2010 |
| conservative standpoint | 25 | 1885-2010 |
| conservative notion | 13 | 1840-2005 |
| conservative dogma | 11 | 1870-1995 |
| conservative objective | 11 | 1970-2005 |
| conservative orthodoxy | 10 | 1980-2010 |
| conservative concept | 6 | 1980-2000 |
| conservative definition | 6 | 1970-2010 |
| conservative interpretation | 6 | 1970-2005 |
| conservative theory | 6 | 1970-1990 |
| conservative constitution | 5 | 1975-2005 |
| conservative structure | 5 | 1970-1990 |
| conservative tenet | 5 | 1840-1870 |

*表 2. 所有年代（1785-2005 年）以“conservative”（保守主义）开头的最相似词语的出现次数，标注了它们出现的大致年代。*

主义信仰）这样的二元语法仅在整个语料库中出现了十几次。

一元语法“principles”（原则）出现的频率高得多（1785 年至 2010 年出现 300,000 多次），因此更有意义的做法应当是使用一元语法嵌入找到哪些词与“原则”含义相近。为了简短一些，我们将仅考察在四个较短的时期内前 100 个最为相似的二元语法，四个时期之间相隔半个世纪：1851-1855 年、1901-1905 年、1951-1955 年和 2001-2005 年。第三个时期（1951-1955 年）的结果显示在图 3 中。

表 1 显示出每个时期的前二十个同义词。

这些词语中是否有哪些用在保守主义言辞的语境中呢？检查以“conservative”（保守主义）作为限定形容词的二元语法是检验这一点的做直接方法。通过将每一时期内前 100 个最相近词语（总共 400 个词）中的每一个与形容词“conservative”相组合，然后去除所有的重复项（包括每个单词的复数形式），我们就得到了表 2 中的结果。

前十个二元语法并没有像“principle”（原则）一词那样传递出强烈的道德价值含义。的确传递

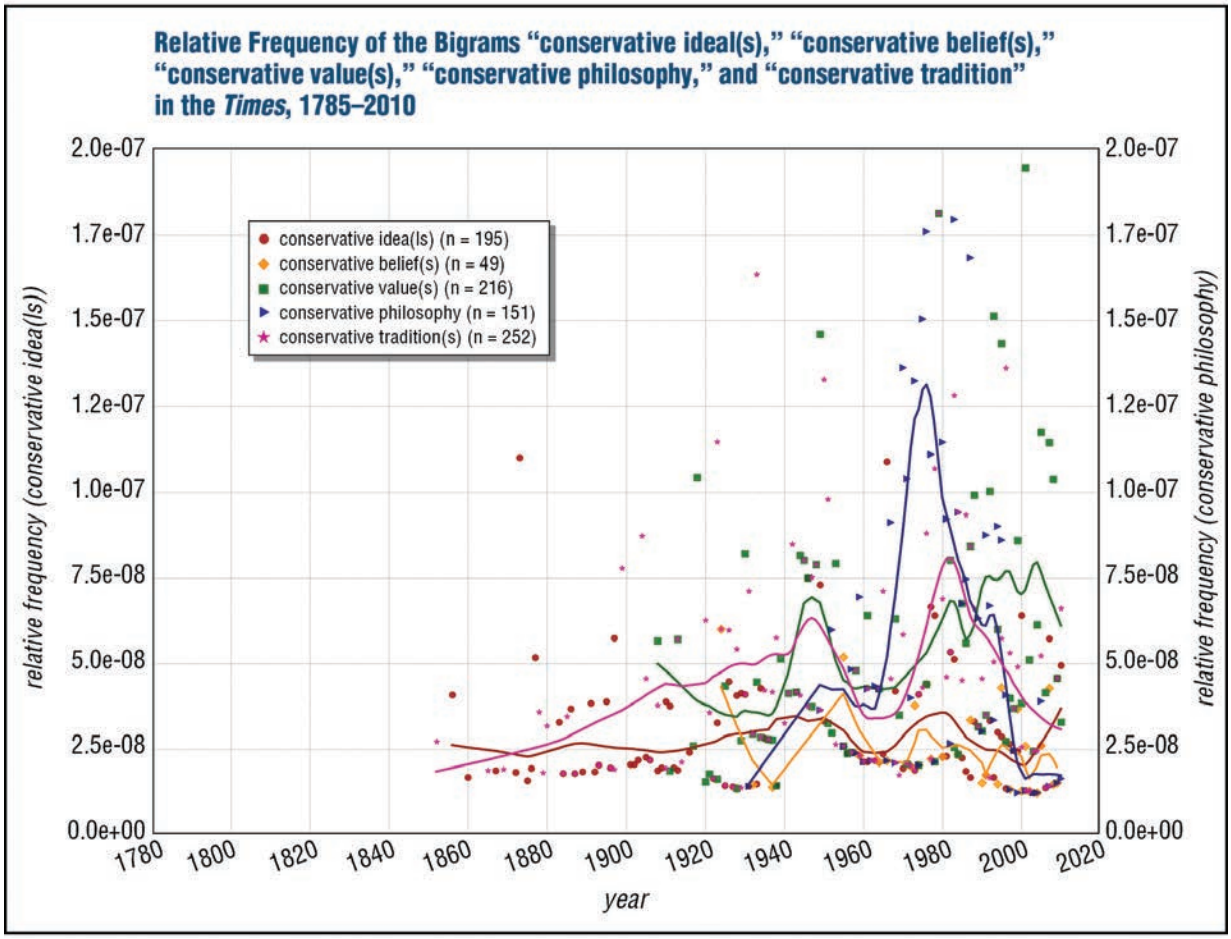



*图 4. 二元语法“conservative idea(s)”（保守主义思想，n=195）、“conservative belief(s)”（保守主义信仰，49）、“conservative value(s)”（保守主义价值，216）、“conservative philosophy”（保守主义哲学，151）和“conservative tradition(s)”（保守主义传统，252）在 1785-2010 年《泰晤士报》中的相对出现频率。*

出这一含义且在较长时间段内出现频率也较多的二元语法是“conservative tradition(s)”（保守主义传统）、“conservative value(s)”（保守主义价值）、“conservative philosophy”（保守主义哲学）、“conservative idea(s)”（保守主义思想）和“conservative belief(s)”（保守主义信仰）。保守主义传统、保守主义思想和保守主义理想在十九世纪已经出现（图 4），而蕴含更多道德意味的二元语法则大部分发生在二十世纪。

“beliefs”（信仰）一词并不十分突出，但有趣的是，“philosophy”（哲学）在大约 1970 年之后变得相当流行，而“values”（价值）则是在 1980 年后流行。后者可能是与“保守主义原则”的道德意味最接近的二元语法。因此我们转而将“conservative value(s)”（保守主义价值）作为一个历史印记，它在保守主义原则几乎消失后成为较常见的保守言辞。

## 保守主义价值的兴起

“价值”（values）一词在十九世纪末逐渐开始被有规律的使用，此时一元语法“principles”（原则）的使用已经开始消退（见图 5）。

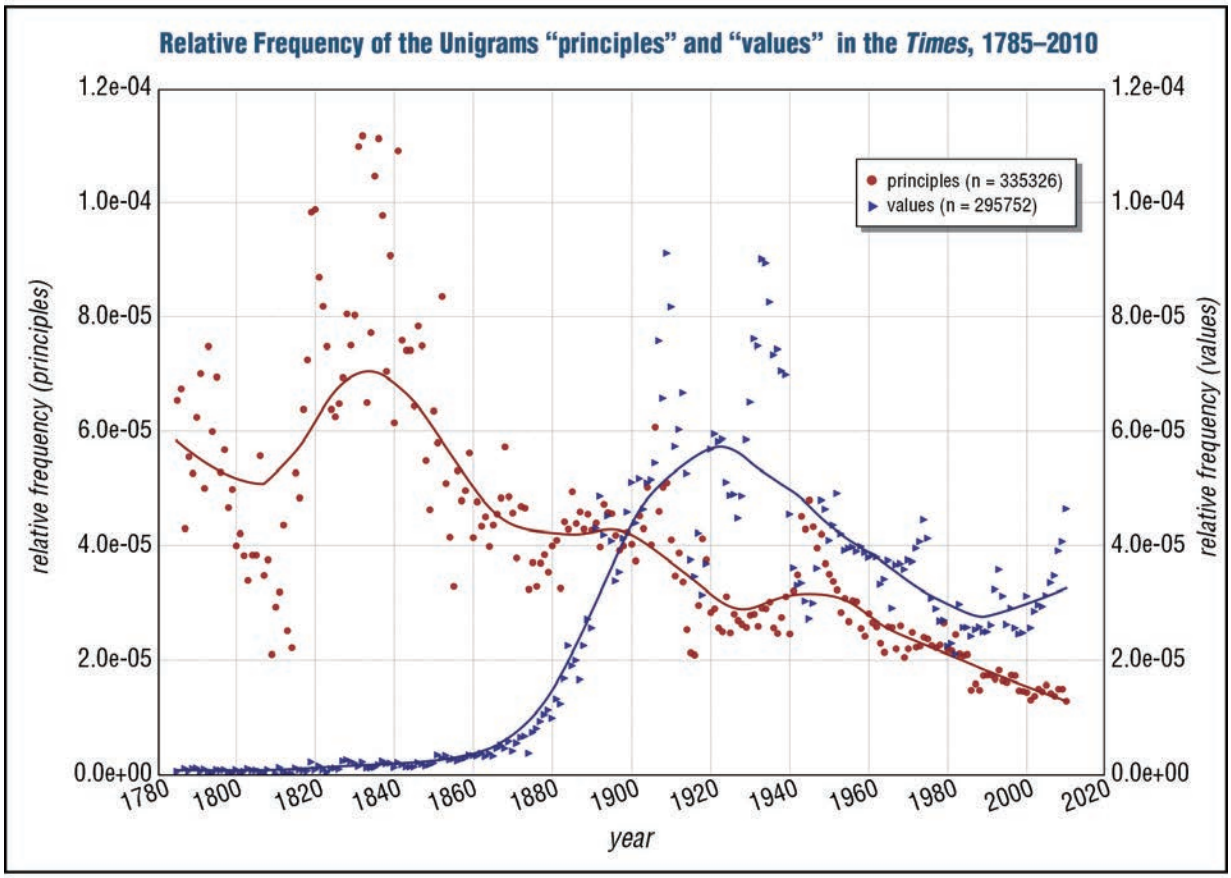



*图 5. 1785-2010 年《泰晤士报》中一元语法“principles”（原则，n=335,326）和“values”（价值，n=295,752）的相对出现频率。*

然而，作为二元语法的一个元素，价值主要但并非完全与经济、金融和贸易相关联，例如“关盘价值”（closing values）或“市场价值”（market values）。《牛津英语词典》定义价值的道德而非经济或货币的含义为“一个人或社会团体秉持的原则或道德标准，大众接受的或个人认为的对生命中珍贵和重要事物的判断”，起源于美国。[33] 但它在英国英语中已经成为一个极为普通的词。根据威利茨的说法，它不是一个仅用于左翼言辞中的词语。

二元语法“conservative values”（保守主义价值）最早的相关用法起源于 1961 年一次“保守主义大会”的议程记录。“我希望大会让保守主义价值接受检验”，哈罗德·麦克米伦（Harold Macmillan）首相时期的外交大臣霍姆勋爵（Lord Home）在新闻报道中说，“并且看到这些价值为人民指出一条清晰的道路”。[34] 战前的保守主义精神的确如此，他所提及的价值意味着所有的行动应当“都出自基督教的原则”，“充满了对君王、宪法和法律的忠诚”以及“受到服务全体人民之愿望的鼓舞”。

与保守主义原则一样，保守主义价值多半都指代政治价值，保守党拥护者支持的那些价值，或保守主义政治哲学家剖析的那些价值。最后一任香港总督彭定康（Chris Patten）正是在 1982 年发起有关保守主义价值深度讨论的人（如果我们追随《泰晤士报》的报道），在他的一篇文章中论及“为什么撒切尔夫人应当加入真正的保守党”以及“回到传统保守主义的价值上”。他用在这一论述中的词语包括一些保守主义常用词，例如比例和平衡，暗含实用主义意味的价值一词在彭定康眼中也意味着群体。他认为，人是“社会动物，仅在组成比他自己更大的团体——例如他的国家、他的教会或他的家庭之后才能达到真正的高度。”[35]

玛格丽特·撒切尔（Margaret Thatcher）将在六年之后回应这一定位，尽管她将家庭放大为保守主义价值的核心。而且，这一特殊的群体显然是由女性掌控的领域。“对于家庭”，她在 1988 年保守主义女性大会上的一次著名演讲中说，“是社会的组成元素。它是幼儿园、学校、医院、休闲中心、避难所和休息的地方。它囊括了整个社会。它塑造了我们的信仰。它让我们为余生做好准备。而女性负责运营它”。[36] 然而，不单单是女性，“传统”家庭都成为了保守言辞的核心。在 1980 年至 2010 年的报纸新闻报道中，二元语法“保守主义价值”被用在三种语境中，都指向要将“传统”约束在保守主义范畴内的趋势。

第一种语境涉及更年轻一代的“后放任型社会”，他们不再觉得“激进的抗议”有意思。相反，他们偏爱保守主义的价值，例如“老式爱国主义、宗教信仰以及支持‘法律和秩序’政策”；或者，他们对婚姻和同性问题采取了更为保守的立场。[37] 这样的评价将会被证明是不正确的，但在 20 世纪 80 年代确实相当常见的。第二种语境是美国，在此里根和“新权利运动”正在努力“恢复”保守主义价值并清除“堕胎、亲吻、色情作品的罪恶，扭转歧视与制度化的无神论”。[38]

第三种语境是英国政治。在撒切尔夫人 1988 年关于女性是经实践考验的管理者演讲中，她提到了自立、个人责任、睦邻友好和慷慨都是“保守主义价值”。除了她所赞成的新自由主义，[39] 她的想法仍受大众欢迎。不时有人会重提“自由、国家、家庭和责任的旧保守主义价值”，正是在这一基础上，保守党政治家基思·约瑟夫（Keith Joseph）构建了 20 世纪 70 年代的撒切尔政策。[40] 然而约翰·格雷很快终结了撒切尔主义。在一篇多少带有预言性质的文章《为什么保守党应投票给工党》中，他哀叹了保守党摈弃传统保守主义悠久目标的方式，“英国社会各阶层的培育以及民间机构的延续”。撒切尔的新自由主义仅仅提高了“普通民众日常生活的风险和不确定性”。不幸的是，保守主义现在正在嘲笑“英国文化中根深蒂固的公平和正派的准则”。撒切尔主义导致了“分裂和道德败坏的社会，犯罪成风，家庭生活缺失或破碎”。[41]

在随后的几十年中，没有一位保守党政治家，无论他或她支持或反对撒切尔夫人，敢不提及自己如何看待传统家庭（和始终如一的核心家庭）。“我们相信家庭”，戴维·卡梅伦（David Cameron）一次次的重申。但根据记者在 2005 年的观察，工党如今也相信家庭。卡梅伦列出了他所有的保守主义价值——个人责任、降低税率、卫生与教育高标准、有限政府、国家主权——传统家庭对左翼和右翼同样具有吸引力。显然，价值还不是替代原则的最佳选择。

## 道德说教政治

所有这些有关“保守主义价值”的保守主义言论即便不是彻头彻尾的说教，显然也是高度规范化的。这就对这些词语所处的更广泛语言环境提出了问题：价值是否与原则出现在不同的语境中？同样，

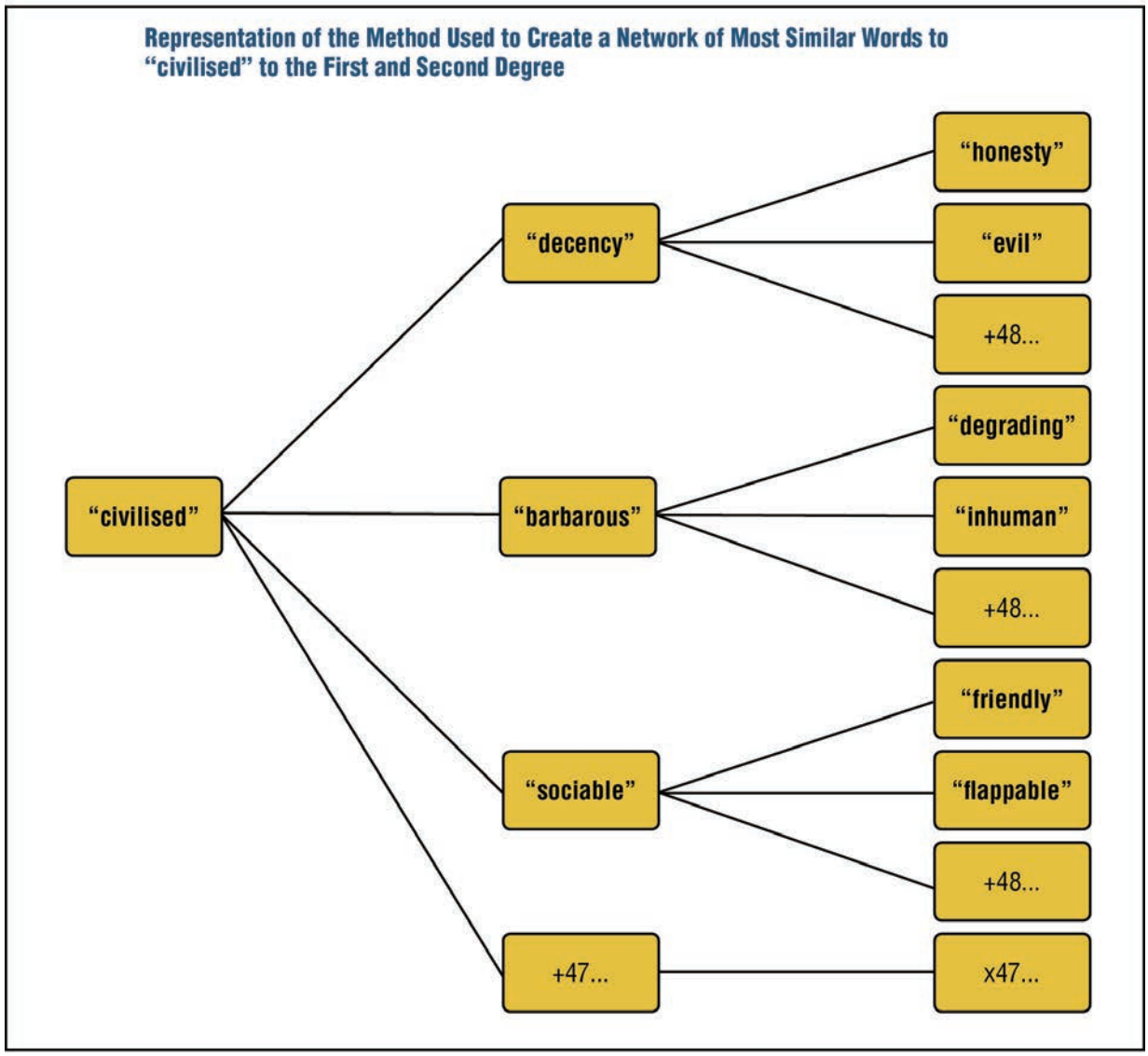



*图 6. 创建第一级和第二级最相近词语网络图所用方法的图示，基于“文明的”一词举例。*

我们可以使用词语嵌入找到 1785 年至 2010 年之间存在的与“保守”（conservative）有意义的关联。因为在《泰晤士报》中“保守”一词本身几乎总是围绕政治（“保守派”和“保守党”极为常见），我们需要找到更多值得考察的词语。这些词语需要在整个时期内具有相对稳定的含义，以便于比较。

我选择了九个词语：”conservative”（保守的，中性词）、“constitutional”（宪法的，与保

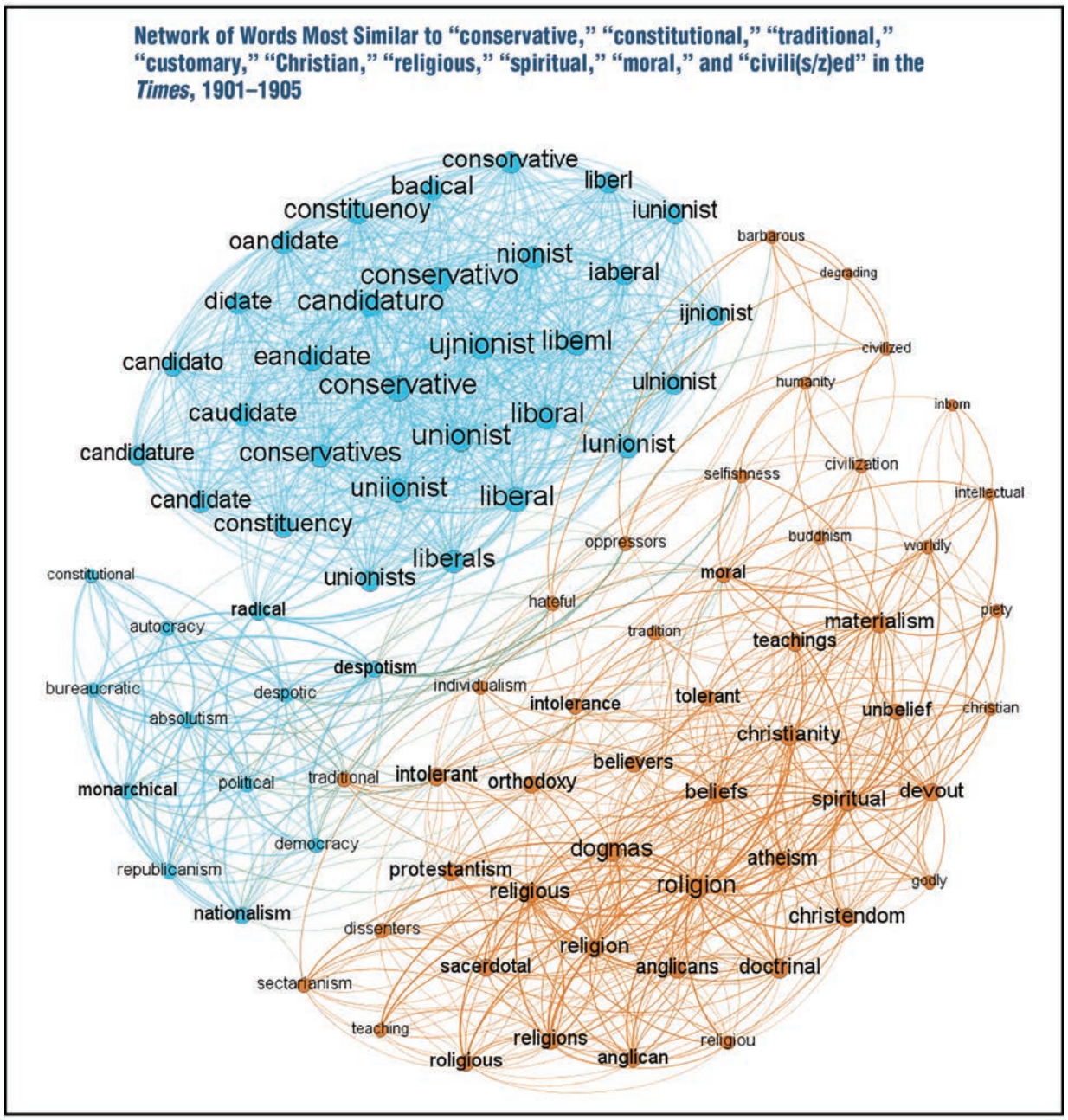



*图 7. 第一级和第二级与“conservative”（保守）、“constitutional”（宪法的）、“traditional”（传统的）、“customary”（习惯的）、“Christian”（基督教）、“religious”（宗教的）、“spiritual”（精神的）、“moral”（道德）和“civili(s/z)ed”（文明的）最相近词语的网络图。基于 1901-1905 年《泰晤士报》中所有词语的一元语法嵌入（无 OCR 修正）。*

守主义相关的政治词语）、“traditional”（传统的）和“customary”（习惯的，与形容词保守的含义相关联）、“Christian”（基督教）、“religious”（宗教的）、“spiritual”（精神的）和“moral”（道德，指保守主义作家经常会提及的宗教和世俗道德的来源），以及“civilised”（文明的，道德品质的标志，另一个我们预期保守主义者会青睐的词）。[42] 对于每一个词在 1901-1905 年、1951-1955 年和

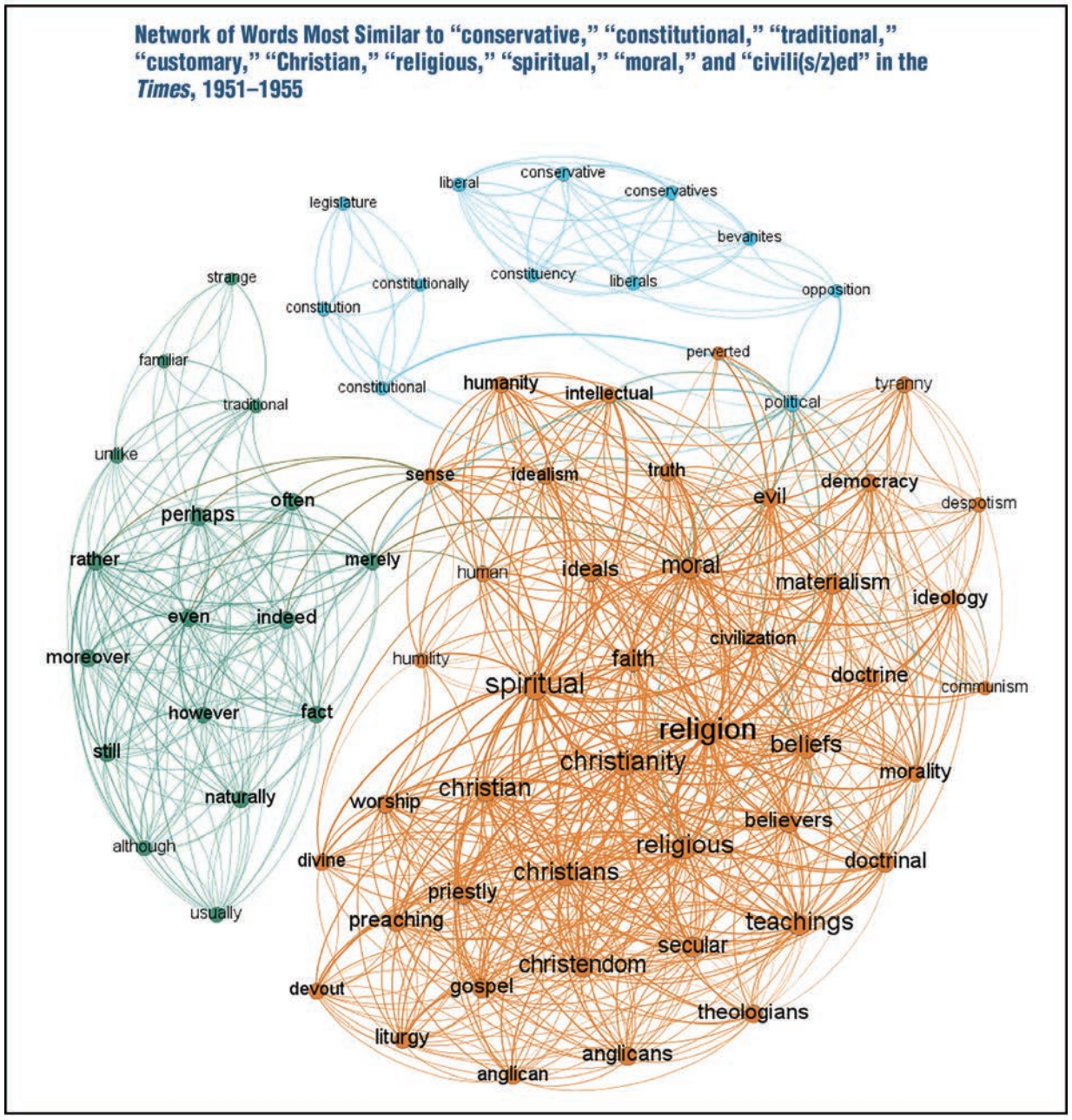



*图 8. 第一级和第二级与“conservative”（保守）、“constitutional”（宪法的）、“traditional”（传统的）、“customary”（习惯的）、“Christian”（基督教）、“religious”（宗教的）、“spiritual”（精神的）、“moral”（道德）和“civili(s/z)ed”（文明的）最相近词语的网络图。基于 1951-1955 年《泰晤士报》中所有词语的一元语法嵌入（无 OCR 修正）。*

2001-2005 年每一个时间段内，我生成了前 50 个最相近的词，这 50 个词的每一个我又计算了另外前 50 个最相近的词（图 6）。

最终得到的共计 22,500（9 x 50 x 50）个词中，独立单词的数量约 5,000 至 7,500 个。这些数据随后被可视化为网络图，每个独立的单词都作为一个节点，节点之间的连线代表“相似度分值”（表示两个词之间的相似程度）。

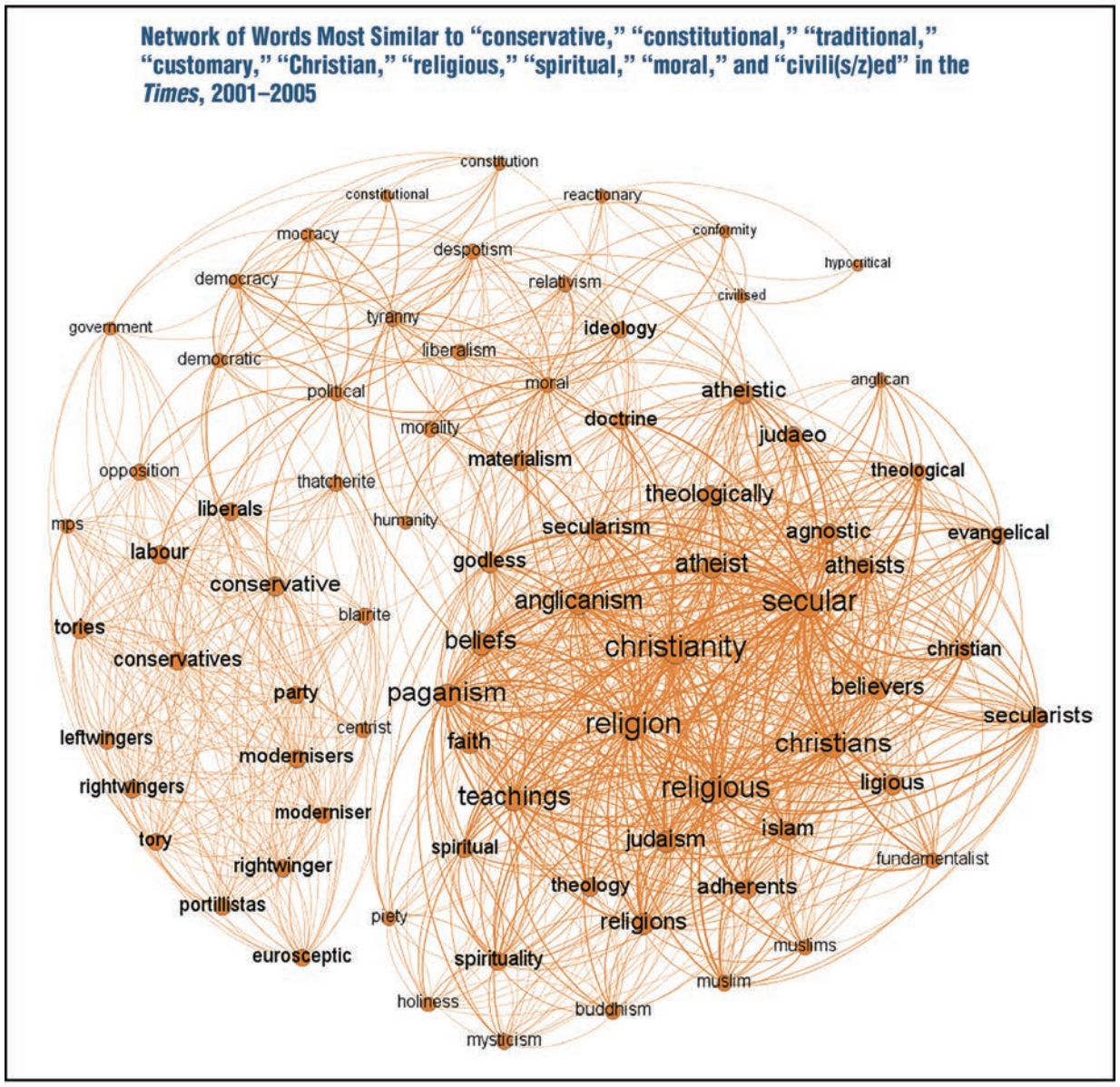



*图 9. 第一级和第二级与“conservative”（保守）、“constitutional”（宪法的）、“traditional”（传统的）、“customary”（习惯的）、“Christian”（基督教）、“religious”（宗教的）、“spiritual”（精神的）、“moral”（道德）和“civili(s/z)ed”（文明的）最相近词语的网络图。基于 2001-2005 年《泰晤士报》中所有词语的一元语法嵌入（无 OCR 修正）。*

通过强制将网络图自动聚类为三组，我们可以看到大致的规律。[43] 在第一个时间段（1901-1905 年，图 7），两个词簇清晰可见，按他们各自的语义判断，明显为“政治”（蓝色）和“文明”或“道德”（橙色）。半个世纪后（1951-1955 年，图 8），可看到三个词簇。我们同样看到一个政治词簇和一个文明词簇，而暗绿色的词簇中的词语与“传统的”或“习惯的”两词更相近。第三个时间段（2001-2005 年，图 9）则出人意料，同样的算法和同样的设置仅识别出一个词簇，根据大部分词语的语义，我认为它关乎“文明”。

这种方法并不精确，需要更深入的探讨，但我们可以得出一个显而易见的结论。规律表明我们也许可以将历史上的“道德说教”（moralisation）称之为政治：在 1955 年至 2001 年间“政治”（political）与“文明”（civilisational）之间的语义关联变得更加强烈。到目前为止我们已经描述过的从保守主义原则到保守主义价值的过渡也符合这个规律。这表明，如果我们想要彻底探讨保守言辞的性质，我们的视线必须超越“保守主义价值”，我们应当检查承载价值的词组，其中有政治意味的限定词（例如“保守的”）不一定会出现。检验这类词组出现次数的一个合乎逻辑的选择是仅查看二元语法“traditional values”（传统价值），因为对传统的保存正是政治保守主义想要做到的，而价值现在则处在政治言论的核心位置。这一二元语法的确是有意义的。1950 年以后，“保守主义价值”出现 165 次的情况下，

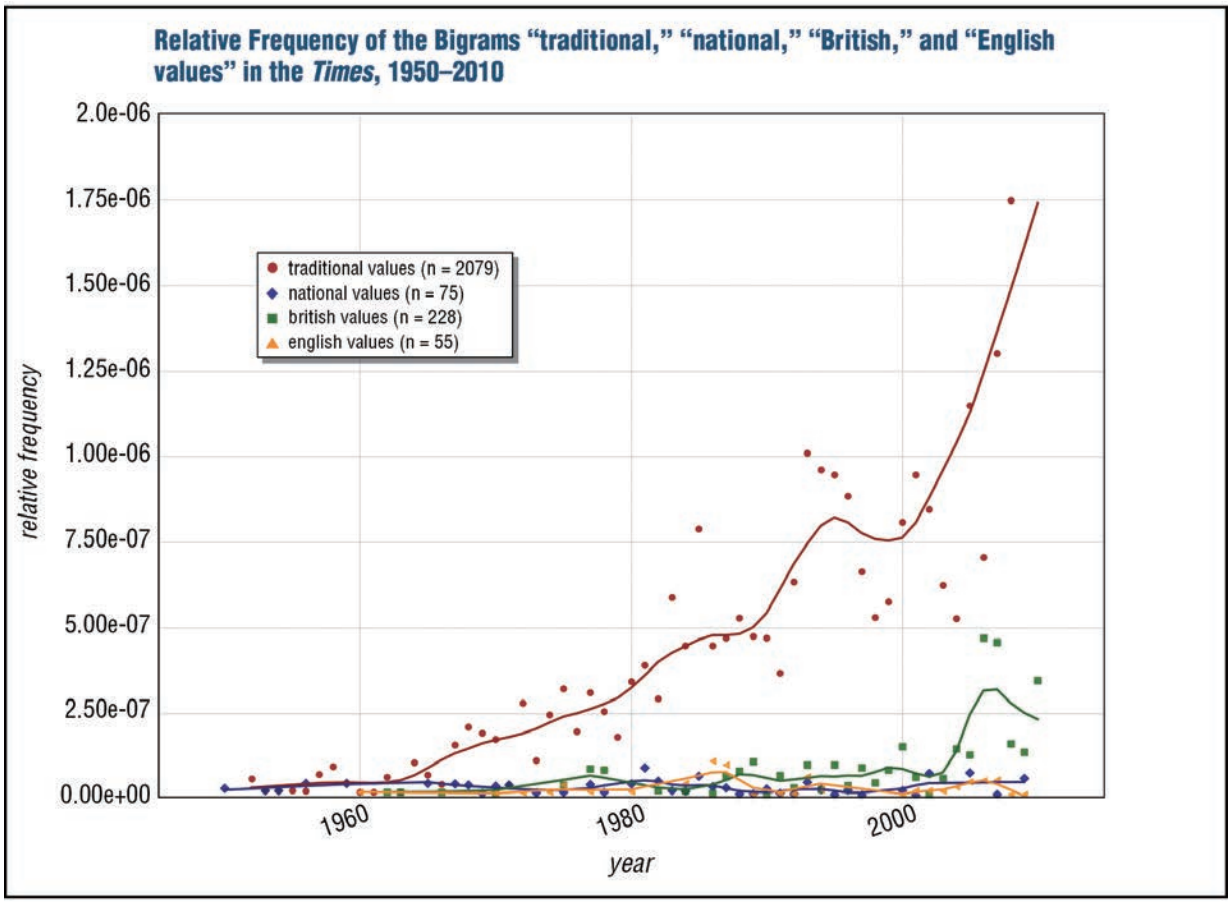



*图 10. 1950-2010 年《泰晤士报》中二元语法“traditional values”（传统价值，n=2,095）、“national values”（国家价值，412）、“British values”（英国价值，252）和“English values”（英国人价值，61）的相对出现频率。*

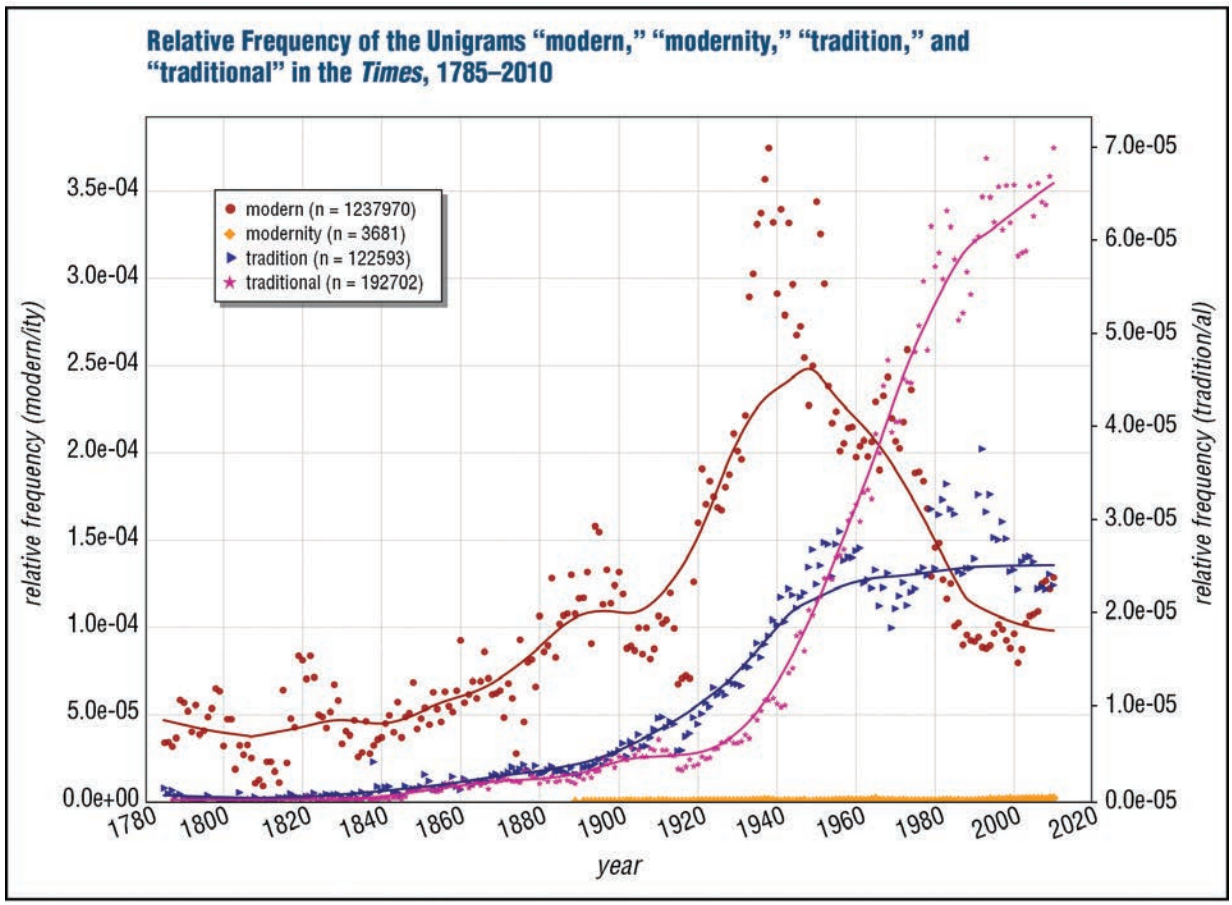



*图 11. 1785-2010 年《泰晤士报》中一元语法“modern”（现代的）、“modernity”（现代化）、“tradition”（传统）和“traditional”（传统的）的相对出现频率。注意 Y 轴的数量级不同。*

“传统价值”的出现飞速上升，总计 2,095 次（图 10）。

有一些出人意料的是，与国家身份相关的二元语法（“国家”、“英国的”和“英国人的”价值）都远远落后。让我们来全面的审视二元语法“传统价值”。

## 传统与现代

二元语法“traditional values”（传统价值）出现在《泰晤士报》的哪些语境中？它在二十世纪上半叶形成，例如在 1934 年，位于英国多塞特郡的布莱恩斯顿学校校长索罗尔德·科德（Thorold Coade，1896-1963）将教育形容为“将过去的文化传递到新生一代的手中”。他认为这种方法意味着性格的塑造，“传统价值的长存”、自律的学习、有组织的游戏和学长制。[44] 无论科德投票给保守党的行为是否直接与此相关，关键的一点是他呼吁的价值可以被称为保守主义价值。而这种对传统事物的呼吁是与现代化相对立的，或与现代化形成了一种相当紧张的关系。

有时候这种摩擦会直白地用意识形态类词语表达。据报道，阿道夫·希特勒曾在 1935 年说，布尔

什维克主义为其无阶级社会的理论牺牲了“数百万人类及无数文化和传统价值”，但取得的进步却甚小。[45] 但更多的情况下二元语法“传统价值”被用来说明非信条的问题。二十世纪是前所未有的时代，快速的变革可能是陈词滥调的说法，但它的确见证了很多人传统生活方式的一去不返。大部分人对这种变化感觉不适，无论他们是否是明确的保守主义者（见图 11）。

1970 年被公认为“致力于复兴帝国统治” 、彻底的保守主义者、日本小说家三岛由纪夫甚至因传统价值的丧失而切腹自杀。[46] 《泰晤士报》中这类关于传统与现代的言论在 20 世纪 50 年代特别盛行。例如，巴基斯坦的新宪法需要和《古兰经》中的世俗法保持和谐一致。《泰晤士报》指出：“调和传统价值与现代需求的任务是复杂且微妙的”。[47]

现代与传统以多种方式发生关联。这种联系通常发生在社会或国家层面，例如在巴基斯坦宪法中，有时现代特指西方社会。如关于东南亚的一篇报道所述，传统价值“被大量的西方产物扫荡一空”。[48] 无论传统和现代的戏码被看作是正面的或负面的或中立的，它都在世界各地上演，从赞比亚、斯里兰卡、加拿大和撒丁岛，到扎伊尔、伊朗和摩洛哥。[49] 坦桑尼亚的乌贾马社会主义并不代表对传统价值的否定，而更多的是传统价值的转型：“团结一致、集体所有和工作分担”简单地变成了“民主、平等分享资源及合作”。[50] 例外的国家显然是智利，它在 1970 年仍保持了完整的传统价值。“这些价值支撑着宪法和有序的政府，尊重法律、人与人之间的礼貌与好客、教育，以及密切关心所有智利的事情”。[51]

商业价值是一项重要的国内话题。1965 年《泰晤士报》警醒了所有的经理人。技术变革正在引发“工业领域基于科学的新文化演变，与传统价值直接冲突”。然而，人们的自然反应（特别是英国人的反应）是“更坚定的固守他原有的且曾经卓有成效的风俗和习惯”。[52] 从经济的角度讲，传统价值成就了英国的伟大。但老派的品质——“很高的技术技能以及绝对的正直和公正”，仅在英国商业实现了现代化之后才能在今天仍是一项资产。[53]

另一个显而易见的主题是大都市成为了现代盘踞的地方，相反农村则成为传统价值的堡垒。法国西部倾向于支持罗马天主教和保守主义，证明“乡下生活和观点较为平淡，更接地气，更显示出内在的稳定性。”。[54] 性别角色是《泰晤士报》中的另一条线索。法国女性主义者伊芙琳·苏莱罗特（Evelyne Sullerot）认为女性对于公民生活的参与刻画出男性担忧的传统问题。“在技术日益发达的社会中，令男性安心的是将女性作为连接他们与传统价值的链条，女性像他们的母亲和祖母那样围着锅台转”。[55]

当性别与对乡村生活的敬畏相结合时，乡愁将会战胜解放。“令我们感到鼓舞的是”，一位（男性）艺术评论家写到，“想到传统价值在一些地方得以保存”。波兰的马佐斯歌舞团是这一事业的拥趸。它表现出“所有的老派美德”，“腼腆而漂亮”的女性穿着蓬松的衣裙，小心翼翼地不露出衬裙跳着玛祖卡舞，身边的男性“活泼而体贴”，手里拿着农具。[56] 在一定程度上，这样生机勃勃的景象看似是永恒的：传统总是好的，而乡愁总是畅销品，即便有时价格昂贵。英国坎布里亚郡凯斯维克的一家旅店提供了重返传统价值的高价服务，包括刚刚“熨平的早报、晚餐着装规范、老式的六便士硬币小费和原版《笨拙》杂志”。[57] 在德文郡西德茅斯的另一家旅店，客人可以享受到“乡村绿地上蟋蟀的低鸣，乡间漫步或草地上的传统英式下午茶”。[58] 同时，野餐篮人士的礼品篮子庆祝着“现代世界中的传统价值”。[59] 广告中传统价值一词的盛行是社会变化的信号。

到 20 世纪 80 年代早期，传统价值已经从一项责任变成了一项资产。从分类广告中我们明显可以看到它们开始被随意地使用。寄宿学校宣扬它们对传统价值的坚守。[60] 汽车生产商召唤高风亮节的过去。一位评论人赞美 200 系列是“结合了传统价值与不仅一点点气派的一辆罗孚汽车”。[61] 建

筑公司 Mowlem 成功混合“传统价值和技能与对当今挑战的远见卓识”。[62] 英属马恩岛则以“强调友好与独立传统价值的一种生活方式”（以及低很多的税率）吸引着公司。[63] 继邮局和 1997 年正值辉煌的斯萨克酒庄之后，地产（Hillier Parker）、法律（Walker, Smith & Way，Reynolds Porter Chamberlain）、住宅（Beazer）、医药（Bayer）、游艇制造（Westerly）和咨询（ADTI Consulting、Connaught）等领域内的公司都纷纷展示它们调和传统与现代的能力，直到 2010 年之后仍是这些公司的一个独特卖点。

在传统与现代的争论中，任何人都可能站在任何立场上，它主要是关于社会 - 文化变革的争论。因此什么时候传统与现代之间的摩擦突变成了保守主义和进步主义之间的冲突呢？《泰晤士报》中的一些文章尝试进行不带感情色彩的分析，结果指向了意识形态的根本转变。在西方民粹主义再次出现以及无社会归属感的中产阶级开始抗议活动之前很久，激进自由主义记者内斯塔·维恩·埃利斯（Nesta Wyn Ellis）就曾警告“法西斯主义”就潜伏在周围。“传统价值破碎导致社会错位”的感觉鼓动着“爱国主义与种族身份”，特别是在“小企业主、商人和职员等看似受困的中产阶级中间，这些人的地位和安全因此而岌岌可危”。[64] 但在重视传统价值的人们的头脑中，主要的元凶并非移民或全球化，而是抗议的一代。六十年代做派的现代化是动摇传统价值的罪魁祸首。

加拿大总督乔治·凡尼尔（Georges Vanier）和他的妻子保琳（Pauline）在 1964 年创立了加拿大家庭联盟（后来的凡尼尔家庭学院），以重温对社会至关重要的传统价值和精神遗产。[65] 然而一旦所有这些都被打乱，一旦“长发、短衫和下流娱乐节目”带来了大规模的“对传统价值的践踏”，[66] 从意大利到英国的右翼评论员和政治家都开始亲自攀越障碍。有时评论极为微妙。《泰晤士报》的保守派编辑威廉·里兹 - 莫格（William Rees-Mogg），因责备英国司法体系对“贾格尔先生”贩毒的严厉指控而令读者印象深刻。“谁在小题大做？”社论标题写道：“如果我们要让一起案件成为合理的英国传统价值与新快乐主义之间矛盾的象征，那么我们必须保证合理的传统价值中包含了宽容与平等”。[67] 但社会在变化，而且在飞速变化。

## 反对放任型社会

在 20 世纪 60 年代就已经有自封的“传统与道德卫士”在谴责“放任型社会”，这是与权威始料不及的正面冲突，是对“健康社会建筑其上的精神现实”的蓄意破坏。为了防止传统价值的永久失落，“我们都应当知道对主的敬畏是智慧的开端”。[68] 毫不意外，保守派拒绝屈从恣意自我满足的思想常常表达为宗教言辞。英国国教主教积极地加入了这一辩论，尽管其中几位，例如拉塞尔·巴里（Russell Barry）明智地采取了中立立场。“目前公众舆论的态度认为绝对的道德规范是对个人自由的侵犯”。如果进步派削弱传统价值，道德家们会”将绝对道德规范等同于强硬的、不变的道德规则”。[69]

无论是宗教还是其他类型的道德家们始终都会存在。宗教组织也纷纷登场，想要保存他们认为已经失落的东西。Pro Fide，一个在英国罗马天主教教会中反对“进步”影响的宗教组织，重新主张传统价值。[70] 基督教团结会“发誓保卫英式生活的传统价值”，[71] 家庭团结联合会坚持“传统价值和文化”，实际上是谴责计划生育和离婚。[72] 当艾滋病疫情变得异常严峻时，一个保守主义但不切实际的派别中的人士（男性和女性）举行了一天的室外活动。正如《泰晤士报》的一位读者所提出的：“长期关系中纯洁和忠诚的传统价值以及一生一位伴侣的高度理想状态应当被积极推崇”。[73] 柏拉图、亚里士多德、西塞罗和托马斯·阿奎那精神中纯洁的友谊以及“英国修道院院长艾尔雷德”，是另一封担心传统价值丧失的读者来信中提出的艾滋病解决方案。[74]

一位曼彻斯特大学的教授在 1973 年评价到，“意识形态之战”已经在保守主义激进分子和马克思主义进步人士之间爆发。前者相信过去的高雅文化和传统价值，而后者认为所有文化都是等价值的，但更青睐“简单的”文化。[75] 这场战争持续了一段时间。当英国童子军领袖、前授勋军官迈克尔·沃尔什（Michael Walsh）在 1982 年提出“回归传统价值”（更多的参加人数、更出色的运动员精神、更有礼貌以及更多的野外冒险而不仅仅是擦拭奖杯）可能会令英国童子军总会受益时，争议出现了。[76] “放任型社会”一词充满了道德暗示，包括与性相关的暗示，在 20 世纪 80 年代和 90 年代则特别牵扯到从保守主义视角会引出负面评论的意识形态（见表 3）。

政治家们很快就开始尝试利用涉及传统价值的言论。在 20 世纪 70 年代的美国，乔治麦戈文（George McGovern）通过提出“复兴美国传统价值”将对激进主义的指控转向民主党的使命。[77] “美国中产阶级”的传统价值在其他地方被视为是“爱国主义、宗教虔诚、独立、乐观和自强不息”，[78] 而尼克松支持者的传统价值则是“法律与秩序、个人品德、冷静节制、自立和爱国主义”。[79] 最终吸引了每个人注意力的右翼现象是罗纳德·里根，他用“上帝、国家、家庭和传统价值”项目征服了美国。[80]

这也是美国传统价值联合会的鼎盛时期，获得了许多美国的保守主义福音派的认可。为了击退同性恋、色情文学和其他各种类型的放纵，该会的主席宣布“我们必须让基督徒遍布各个联邦机构”。[81] 这是在 1983 年，大约十年后，该联合会发起了反对克林顿“反上帝议程”的宣传活动，支持电视福音布道者帕特·罗伯森（Pat Robertson）。他定义女性主义为“一种社会主义的、反家庭政治运动，鼓励女性离开他们的丈夫、杀死他们的孩子、练习巫术、破坏资本主义、变成女同性恋”。[82] 这种措辞甚至是大西洋彼岸伦敦最强硬的保守派也难以忍受，特别是在该联合会将注意力转向《哈利·波特》（其中的“巫师相信堕胎是一种神圣的行为”）之后。[83]

美国人转向了在西欧引起共鸣的道德正确，西欧同样见证了道德说教性质的保守言辞的兴起。德国的基督教民主领袖赫尔穆特·科尔（Helmut Kohl）开始强调应当适度奖励女性留在家中，因为家庭是“传统价值和道德品质”的堡垒。[84] 在法国，右翼政治家雷蒙德·巴尔（Raymond Barre）赞扬劳动、家庭、祖国的传统价值。[85] 而在意大利，右翼联盟青睐基督教社会、“国家优先”以及语言和服装中蕴含的传统价值。[86] 在英国，当上议院通过一项教育法案的修正案时，有关“体面、健康传统价值”的保守言辞在 20 世纪 80 年代重获新生。其中规定了性教育应当灌输到年轻人的“道德意识和家庭生活价值观中”。[87] 不久之后，撒切尔夫人启动了她的“清除英国十字军”运动，以恢复“公平、政治、诚实和谦逊的传统价值”，并消灭“社会主义的错误价值观”。[88]

当然，英国保守派（相比美国共和党）的问题在于他们对宗教的诉求在社会上的基础极其薄弱。曾几何时英国国教被称之为“祈祷中的保守党”，到 2004 年传统宗教价值观的坚定维护者仅剩下罗马天主教。[89] 一位记者在 1991 年观察到：

> 大部分的保守党大会成员——尽管他们大声疾呼着传统价值，以及大部分的内阁成员可能都是无神论者。他们对英国赞美诗、《公祷书》和《圣经》的怀旧之感非常美好、值得尊敬，但从宗教意义上讲是完全空洞的。无神论保守党摆脱这一困境的简单方法是坚持基督教是我们社会中最重要的历史和文化力量且必须代代相传。（……）他们渴望在民众普遍认定宗教并非真实的氛围中复兴宗教这一社会团结的力量。[90]

这真是一针见血。这也是人们对自己所处世界向道德说教政治转变的担忧：如果我们不再相信基督教，至少我们可以将其作为一种文化力量。只有那些已经转变思想的人能够泰然自若的接受威尔士

**Top Twenty Phrases Most Similar to the Bigram "permissive society" in the *Times*, 1961–2010**

| | 1961-1965 | 1966-1970 | 1971-1975 | 1976-1980 | 1981-1985 | 1986-1990 | 1991-1995 | 1996-2000 | 2001-2005 | 2006-2010 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | foreignness | racial prejudice | permissiveness | uncaring | materialist | feminism | puritans | social revolution | insidious | freedom choice |
| 2 | exotically | misguided | morality | mentality | holocaust | permissiveness | feminism | religious differences | institutionalised | public discourse |
| 3 | scenes characters | intolerance | nazism | bigotry | telling phrase | intolerance | liberalism | sexual matters | road serfdom | totalitarian regimes |
| 4 | opera goer | bigotry | homosexuality | neuroses | sexual freedom | materialism | stalinism | secularism | irrational self | moral superiority |
| 5 | inaccessibility | immoral | death wish | atavistic | christian beliefs | liberalism | oppressor | calvinism | dismisses | whole notion |
| 6 | characteristic | morals | materialistic | taboos | modern architecture | anarchism | reactionary | sexual liberation | trade unionism | completely justified |
| 7 | altogether absent | hypocritical | fanaticism | restlessness | lust power | homosexuality | reactionaries | turmoils | lies root | proper regulation |
| 8 | difficult detect | permissive | perversion | inarticulate | liberal democracy | elitism | moralists | solidarity union | miscegenation | cultural relativism |
| 9 | regional characteristics | shameful | perverted | world weary | detestation | marxism | traditionalists | permissiveness | simply repeat | incapable |
| 10 | appreciate true | dogmatic | elitist | fatalism | agnosticism | family values | fascism | western culture | islamophobia | sense fairness |
| 11 | form expression | permissiveness | revolting | callousness | grossest | hypocrisy | puritanism | xenophobia | jobbery | persuasions |
| 12 | habits mind | ignorance | atheist | repressed | moralism | egalitarianism | [camille] paglia | atmosphere fear | totalitarian state | become burden |
| 13 | antecedent | misunderstood | womanhood | work ethic | sentimentality | authoritarianism | heretics | matriarchy | religious toleration | presumption |
| 14 | affluent society | revulsion | totalitarianism | egotism | tibetan people | passionately | sense shame | rastafarianism | woolly minded | restriction |
| 15 | avoid becoming | hypocrisy | neurosis | miseries | israel diaspora | reactionary | cruelties | apprehensions | superstition | permissiveness |
| 16 | great religions | facts life | revulsion | inbred | spiritualism | evils | triumphalist | provisionals | occultism | confusing |
| 17 | primordial | self-righteous | blasphemy | xenophobia | iconoclasm | bigotry | peasantry | anti-american | preoccupation | weak vulnerable |
| 18 | abrupt changes | homosexuality | sexuality | human spirit | semitism | selfishness | homosexuality | feminism | immigration crime | context explanation |
| 19 | horror films | ignorant | bigoted | craving | hardly blamed | tyranny | bad loser | radicalisation | strangled birth | opaquely |
| 20 | becomes aware | vindictive | self-righteous | hatreds | irreligious | feminists | marxists | sectarianism | tim luckhurst | imaginative ways |

*表 3. 与二元语法“permissive society”（放任型社会）最相似的前 20 个词组（去除停止词），基于 1961-2010 年《泰晤士报》每五年时间段的二元语法嵌入。与现代化相关的单词 / 词组标注为红色，与意识形态有关的词语（大部分词尾是“ism”和“ist”）标注为蓝色，与性文化相关的词语标注为绿色，所有其他词语为黑色。Camille Paglia（列在 1991-1995 一栏中）是主流女性主义性思想评论家，Tim Luckhurst（2001-2005 年一栏）作为《泰晤士报》专栏作家撰写了有关性教育的文章。还可以看到（2001-2005 年一栏中）弗雷德里希·哈耶克的《通往奴役之路》（1944 年）。*

亲王的期待“我们从祖先那里继承的文明价值的存亡取决于我们心中发自内心的神圣感的存亡”。[91]

不仅是保守派独有的传统价值如此，工党亦是如此。对于托尼·本恩（Tony Benn）而言，它们包括坚持“英国人民的民主权利以及经协商的社会主义主张”。[92] 对于工联主义者比尔·瑟斯（Bill Sirs）而言，它们相当于“民主、宽容、公平心和理解”。[93] 然而这些对保守言辞表达目的和意图方式的早期挪用与 20 世纪 90 年代左翼对保守主义道德语言厚颜无耻的劫持仍相距甚远，就是在此时，“新工党”开始将传统价值“置于现代语境之下”。[94]

## 平凡化传统

1993 年《泰晤士报》报道了约翰·梅杰在布莱克浦的著名“回到民生”（Back to Basics）演讲。这位首相发出了鼓舞人心、复兴主义的呼声：“传统价值、家庭责任、语法和拼写及无犯罪街道”。报道这篇新闻的记者观察到这次演讲的确打动了人们的心弦，尽管它“之所以著名是因为其语气和风格，而不是其主旨”。它不仅仅是对 20 世纪 60 年代和放任型社会的抗议。“它更是对过往岁月的祈求，那是《图画邮报》和伊林戏剧的世界，萨里队总是赢得郡板球冠军的年代，黑白电影而非彩色大片的时代”。[95] 保守党右派为之欢欣鼓舞。“基本价值”、五十年代风格已被提上国家议程，尽管一位专栏作家明智地预言，重新“强调责任与义务而不是权利”将主要是修辞意义上的。[96]

传统价值，包括“自律和忠诚”[97] 无疑在议程当中，但并不是以梅杰预期的方式。其中存在两个问题。第一个问题是梅杰的内阁很快就不得不应对保守党内部一系列的通奸和各种性犯罪指控。第二个问题是传统已经比过去更为社会接受。例如《好管家》（Good Housekeeping）并非一本普通的杂志，而是宣称男性沙文主义和女性牺牲品已经是过去的事情、未来取决于女性的“新传统主义者”。[98] 传统显然已不再是保守派的独家特权。

这是托尼·布莱尔登台的时刻。1994 年他发布了一项声明，“高瞻远瞩但不轻易承诺”，他在其中展现了他的雄心“维护工党的传统价值但将它们放在现代语境之中”。[99] 一年后《泰晤士报》转载了布莱尔另一次演讲的节选：

> 家庭是很重要的，因为我们正是在家庭中学会了自我尊重和尊重他人。正是在家庭中我们第一次感受到自由的界线和责任的根源来自何处。家庭是狭隘自私的对立面。从家庭出发，我们构建出更宽广的社会。我相信我们有道德义务去帮助那些生活境遇每况愈下、体弱多病或失业的人们。这是左翼人士的传统价值所在，理智且实用地运用于现代世界，展现它的力量。它能够为国家塑造新的德育目标。[100]

如果这段引述中的“左翼人士”替换为“右翼人士”，货真价实的保守党会发现这段文字完全是赏心悦目的。工党用保守党自己的立场击败了保守党。

胜利并未持续很长时间。“英国中产阶级努力、奋斗，直截了当地呼吁”更好的生活，他们可能变得离不开传统价值，但却没有兴趣聆听政客们对传统价值夸夸其谈。[101] 诚然，布莱尔的措辞无可挑剔。《泰晤士报》常将托尼·布莱尔描述为极有天赋的演说家，他的演讲总能让所有听众找到真正的归属感，但应当归属于什么却往往不得而知。分析布莱尔语言的专家发现他明显偏爱“进行显而易见的对比，用‘和’或‘但是’连接对比双方，传递出的含义好像它们是共同发生的而不是替代发生的”。利用两面讨好的修辞方式，布莱尔能够将包罗万象的信息压缩在仅有 839 个单词的演讲中，包含无数

成对的对比：

- 野心和同情心
- 头脑和心灵
- 投入并且也改变
- 社会公正和机会
- 联合商业但追求社会公正
- 不同但更合理
- 继续但保证
- 惩罚但提供善后
- 快速但稳妥

同时，布莱尔（及他的演讲写手）通过连结分歧融合左翼和右翼语言的这种能力的一个著名例子仍是工党对两个对立面的升华“现代语境下的传统价值”。[102]

一位评论员（从前的马克思主义者米克·休姆）注意到表面上的右翼言辞如今已被知识精英们摈弃，因此“传统价值的战士已经在文化相对主义、差异和身份政治面前撤退”。“西方文明已经变成一种肮脏的表达”。[103] 但适用于知识分子的不一定适用于整个社会。尽管“媒体圈子里自由至上，普通人”仍坚守着传统价值。整个社会公然地陶醉于社会价值，以至于一些二十一世纪的保守派将复兴这些传统的希望寄托于“信仰、旗帜和家庭”。[104]

在《泰晤士报》中，二元语法“traditional values”（传统价值，以及必然伴随而来的“家庭价值”这个词在 1980 年至 2010 年之间非常普遍，值得进行单独的分析）如今被所有想要表达对生活“保守”展望的人挪用。约会广告很好地说明了这种对传统价值平凡化的倾向。在 1990 年前后这些广告开始提及传统价值。“这将会改变你的生活”，刊登广告的是一位“有着体育爱好的乡绅”，正在寻找一位“受过传统价值培养和教育的年轻女性”。一位“坏脾气的怪老头”寻找一位“顽强的、善于察言观色的寡妇”，中年、聪明、良好教育、良好背景且具备传统价值。[105] 不久之后，因为保守思想的单身人士对灵魂伴侣的憧憬，对传统价值的引用达到了峰值（图 9）。这个词成为了日常用语的一部分。

由于《泰晤士报》的读者可能相当大部分是精英人才，所以除了我们所能预期到的，很难分辨在这些提及传统价值的广告中存在什么样的特定规律。刊登广告的男性和女性分享了他们的高雅爱好，例如音乐、戏剧、艺术、旅行、烹饪和乡村活动。他们十分重视外貌。他们期待未来的另一半是有教养的、优雅的、成功的、专业的、身心健全的、善交际的、幽默的和聪明的。有时人们介绍自己是基督教徒、希腊东正教徒或其他宗教信仰，但通常寻找伴侣的人并不说明自己的意识形态。一个很值得注意的例外是一位“有魅力的作家”，一位牛津大学毕业的女士寻找“一位聪明、有文化和传统价值的男士”，她期待他“和善、对政治感兴趣且是彻底的右翼”。[106] 这位问题女士显然在寻找配偶上发生了一些困难，她在 2007 年开始刊登广告时 39 岁，奇怪的是仅两年之后就变成了 44 岁，然后直到 2010 年始终 44 岁。

## 保守主义之外

追踪几条保守言辞的线索，从“保守主义原则”到“保守主义价值”再到“传统价值”，我尝试勾画出我们可能称之为“保守”的道德语言的一隅。这一研究得到了一些有趣的发现。特定的词组显

然遵循特定的历史轨迹。原则被价值替代。政治选择在 20 世纪 60 年代之后开始明确地表达为道德词语：政治言辞似乎从道德角度被人们解读着。左翼和右翼之间意识形态的差异结束了这一切，这种紧张关系在 20 世纪 80 年代因大众言论中保守言辞的同化而得以升华。整个英国社会都惊人地转向传统，或者至少转向了推动传统价值的通用保守言辞。

到 2010 年，这种表面上的保守言辞实际上已经成为主流。阅读《泰晤士报》，我们会看到仅有小范围的政府和国家身份讨论可以进一步发展为能够称之为“保守主义”的政治观点，而即便这些观点也不会总是被认为是“右翼”。事实上，自 20 世纪 60 年代之后，保守主义思想的人们就发现很难找到一种措辞方式能够替代不证自明的“保守主义原则”言论，但它的吸引力不如另一个事实，即这些新的措辞方式出现自与现代的紧张关系中，而这种关系则主要是由“六十年代”激发的。然而，随着时间的推移，放任自身也变成了一个相对概念，因此到底哪些需要被保存仍不清晰。但这种与现代的角力不仅仅适用于保守主义，也应适用于所有脱胎于“现代”这类问题的道德语言。

我们倾向于按意识形态分类道德语言，例如“启蒙的”、“自由的”、“基督教”、“民族主义”或“保守主义”，但这些简单的标签更有利于订购礼物而不是理解过去，反之亦然。我想要明确的是，数字历史技术帮助我们找到最终形成道德语言的不断变化中的词簇。本文的关注焦点是原则和价值，既有保守主义的也有传统的，这让我们触及了这些变化领域的复杂性。挑战在于将它们分解为构成它们的规范修辞元素和动态修辞元素。这将让我们能够领会一个不同的、更为复杂、模棱两可和多样化的过去，帮助我们在似乎已经失去承载的当下想象不同的未来。

## 注释：

1. “Right foot forward, march,” Times (Sept. 25, 1997), 38. All references to the Times are through The Times Digital Archive (https://www.gale.com/intl/c/the-timesdigital-archive) and accessed in April 2019.
2. “principle, n.,” OED Online (Oxford, Oxford University Press), accessed April 29, 2019.
3. The reference is to Disraeli, Coningsby, or The New Generation, Vol. III, Book viii, Chapter iii.
4. This does not just apply to the English language. Cf. for example the Dutch word for principles, beginselen, which is still in use but has a nineteenth-century flavour; see “beginsel” in http://gtb.ivdnt.org/, accessed April 8, 2019.
5. Disraeli, Coningsby, Vol. I, Book iii, Chapter v.
6. For this research the full dataset of the Times was used, made available by Gale as XML files, which were then converted to CSV for further processing. No further OCR correction was done.
7. For a history of conservative ideas, see E. H. H. Green, Ideologies of Conservatism. Conservative Political Ideas in the Twentieth Century (Oxford: Oxford University Press, 2002).
8. “Conservative Values Past, Present and for the Future,” Times (July 23, 2001), 15.
9. The recent literature includes: Emily Jones, Edmund Burke and the Invention of Modern Conservatism, 1830–1914. An Intellectual History (Oxford: Oxford University Press, 2017); Richard Bourke, “What Is Conservatism? History, Ideology and Party,” European Journal of Political Theory 17 (2018), 449–75.
10. N-grams were calculated per year on the basis of all newspapers. Only n-grams with a frequency higher than 4 over the whole corpus (1784–2010) were taken into consideration.
11. See note 2.
12. “British Association,” Times (Aug 2, 1907), 9 (on physiological inquiry in London).
13. “Conservatism,” Times (May 9, 1930), 10, a review of Feiling’s What Is Conservatism? (1930). On Feiling: Reba N. Soffer, History, Historians, and Conservatism in Britain and America: From the Great War to Thatcher and Reagan (Oxford: Oxford University Press, 2009), 86–109.
14. “Private Correspondence,” Times (Feb. 13, 1823), 3; “Despatch from the Count de Nesselrode,” Times (Jan. 20, 1823), 3.
15. “House of Commons,” Times (Nov. 21, 1837), 2.
16. “London, Tuesday, August 3, 1841,” Times (Aug. 3, 1841), 4.
17. The trigrams “one nation,” “tory,” “tories,” “toryism,” “conservatism,” “conservative(s),” and “party”

occur 250 times from 1971 onwards. In "Why Tories Are Unpopular," Times (Jan. 12, 1996), 1, Thatcher was reported castigating John Major's pro-European politics as "no-nation Conservatism." On the onenation tradition, see Robert Walsha, "The One Nation Group and One Nation Conservatism, 1950–2002," Contemporary British History 17, no. 2, 69–120; David Seawright, The British Conservative Party and One Nation Politics (New York: Continuum, 2010); Stephen Evans, "The Not So Odd Couple: Margaret Thatcher and One Nation Conservatism," Contemporary British History, 23, no. 1 (March 2009), 101–21.

18. "The British Constitution," Times (Mar. 20, 1832), 3.
19. "The Peel Banquet," Times (May 14, 1838), 5.
20. "Conservative Dinner at Birmingham," Times (Dec. 20, 1838), 3.
21. "Mr. Bright at Birmingham," Times (Apr. 25, 1859), 11.
22. "Election Intelligence," Times (Mar. 19, 1867), 12.
23. "Equal Sacrifice in Industry," Times (July 2, 1925), 13. On the rejection of socialism, see G. C. Webber, The Ideology of the British Right, 1918–1939 (New York: St. Martin's Press, 1986).
24. "Manifesto by 'Die-Hards,'" Times (Mar. 8, 1922), 14. See N. C. Fleming, "Diehard Conservatism, Mass Democracy, and Indian Constitutional Reform, c. 1918–35," Parliamentary History 32, no. 2 (2013), 337–60.
25. "Mr. Baldwin on Defence," Times (May 2, 1936), 8. On the Primrose League, Mitzi Auchterlonie, Conservative Suffragists: The Women's Vote and the Tory Party (New York: Tauris, 2007).
26. "Prime Minister on Defence of Europe," Times (Oct. 12, 1953), 11.
27. "The Conservative Path for 'Putting Britain Right Ahead,'" (Oct. 7, 1965), 18. See Peter Dorey and Mark Garnett, "'The Weaker-Willed, the Craven-Hearted': The Decline of One Nation Conservatism," Global Discourse: An Interdisciplinary Journal of Current Affairs and Applied Contemporary Thought 5, no. 1 (2015), 69–91.
28. "Republicans Already Staking Claims for 1996 Nomination," Times (Jan. 13, 1993), 11.
29. "Right's Knight Issues Challenge to Bush," Times (Dec. 11, 1991), 8.
30. "Right Foot Forward, March," Times (Sept. 25, 1997), 38.
31. "Christian Conservatism," Times (Dec. 7, 1998), 23.
32. Unigram and bigram embeddings were created mostly per five-year period in all cases with a minimum of 100 million tokens per period after stop word removal, on the basis of the following article genres in the Times: "Law," "News," "Sport," "Letters to the Editor," "News in Brief," "Reviews," "Editorials/ Leaders," "Business and Finance," "Politics and Parliament," "Arts and Entertainment," "Feature Articles (aka Opinion)," and "Obituaries." Gensim's word2vec (https://radimrehurek.com/gensim/models/word2vec.html) was used with the following settings: size=160, window=10, iter=12, min_count=3, workers=3.
33. "value, n.," OED Online (Oxford: Oxford University Press), accessed April 29, 2019.
34. "Soviet Convinced of Risks, Says Lord Home," Times (Oct. 12, 1961), 6.
35. "Why Mrs Thatcher Should Join the Real Tories," Times (Oct. 5. 1982), 10.
36. "We Are Only in Our Third Term, and a Woman's Work Is Never Done," Times (May 26, 1988), 12.
37. "The New Young Embrace the 'Older' Values," Times (Nov. 26, 1982), 14; "European Teenagers Value Money above Personal Happiness," Times (May 11, 1989), 3.
38. "The Right Hopes It Can at Last Court Favour," Times (June 21, 1986), 5.
39. Cf. "MPs in New Group Are 'Steeped in Traditional Tory Values,' says Pym," Times (May 15, 1985), 4, on the "Conservative Centre Forward" group.
40. "I Will Not Make the Same Mistake," Times (July 4, 2001), 12.
41. "Why Tories Should Vote Labour," Times (June 24, 1994), 16; based on Gray's The Undoing of Conservatism (1994).
42. Joris van Eijnatten and Ruben Ros, "The Eurocentric Fallacy: A Digital Approach to the Rise of Modernity, Civilization and Europe," International Journal for History, Culture and Modernity (2019).
43. The networks were generated in Gephi: filtered on degree > 75; the modularity was set to 3 clusters; no OCR corrections were done. The network layout is based on the Fruchterman-Reingold algorithm.
44. "Co-Education at the Top," Times (Oct. 26, 1934), 11.
45. "Herr Hitler's Thirteen Points," Times (May 22, 1935), 16.
46. "Japan Keeps the Mask on Mishima," Times (June 19, 1985), 12.
47. "Pakistan's New Shape," Times (Dec. 29, 1952), 7.
48. "Problems of the New Burma," Times (Aug. 22, 1955), 7.
49. "An Alienated Elite," Times (Nov. 18, 1968), VII; "Elite Point Tasks Ahead," Times (Jan. 8, 1969), II; "Indians on the March—A Crucial Choice," Times (Feb. 26, 1969), VII; "Collapse of Farm Economy," Times (Apr. 27, 1970), 26; "Mobutu—A Personal Portrait," Times (Nov. 24, 1972), III; "Conditions in Iran," Times (Sept. 22, 1978), 15; "Moroccans Fast in Daylight but Feast by Night," Times (June 14, 1983), 5.
50. "Letter from Dar es Salaam," Times (Mar. 4, 1975),

XIV.

51. "Order in the Midst of Strife," Times (Aug. 29, 1970), I.
52. "Managers Must Be Ready for Revolutionary Changes," Times (Sept. 8, 1965), 7.
53. "Consultants: A Major Force in Engineering," Times (Aug. 9, 1971), 14.
54. "Traditional West Undergoing Process of Change," Times (Nov. 7, 1958), 11.
55. "France's Feminine Feminist," Times (May 14, 1969), 11.
56. "Mazowsze," Times (July 23, 1976), 11.
57. "Necks Please," Times (Jan. 18, 1994), 18.
58. "The Victoria Hotel," Times (May 26, 1994), 2.
59. "Hamperpeople," Times (May 21, 2001), 19.
60. "Courses and Seminars Review 1981," Times (Aug. 17, 1981), 20 (on Buckswood Grange).
61. "Austin Rover," Times (Jan. 8, 1985), 7.
62. "John Mowlem & Company Plc," Times (Apr. 19, 1985), 17.
63. "Isle of Man," Times (Feb. 21, 1986), 23.
64. "Disturbing Signs that Fascism Could Be Just round the Corner in Britain," Times (June 1, 1977), 7.
65. "Protecting Canada Family Life," Times (June 10, 1964), 14.
66. "Italians Fear Growth of Gangster Violence," Times (Sept. 27, 1967), 5.
67. "Who Breaks a Butterfly on a Wheel?," Times (July 1, 1967), 11.
68. "Slippery Slope of a Permissive Society," Times (Dec. 27, 1969), 7.
69. "The Paradox of Christian Freedom," Times (July 8, 1972), 16.
70. "Ten Priests Guilty of Heresy, Group Says," Times (Mar. 19, 1971), 2.
71. "The Need for Better Research into Family Life," Times (Mar. 28, 1977), 14.
72. "Irish Divided by Divorce," Times (Apr. 30, 1986), 11.
73. "Fight against AIDS," Times (Dec. 8, 1986), 13.
74. "AIDS and Morals," Times (Jan. 28, 1987), 17.
75. "Parents May Win Right to Appoint School Staff," Times (Jan. 8, 1973), 4.
76. "Threat to Grant Regrettable," Times (Dec. 3, 1982), 4.
77. "Democrats Given Blunt Warning," Times (Aug. 10, 1972), 5.
78. "Voices that Have Become an American Institution," Times (Aug. 29, 1973), 14.
79. "End of a Presidency," Times (Aug. 9, 1974), 17.
80. "Election 84," Times (Nov. 2, 1984), 7.
81. "Moral Majority Awaits Its Greatest Triumph," Times (Nov. 5, 1984), 7.
82. "Evangelicals Rally to Beat the 'Sinful,'" Times (Oct. 19, 1992), 11.
83. "Spin Cycle," Times (Aug. 23, 2001), 14.
84. "CDU Offers the Traditional Remedies," Times (Feb. 8, 1983), 5.
85. "Jovial Defender of Tradition," Times (Mar. 12, 1986), 40.
86. "Italy Moves to the Right in Local Elections," Times (Apr. 18, 2000), 21.
87. "Putting Love into Sex Education," Times (June 4, 1986), 13.
88. "Thatcher Champions the Family," Times (May 26, 1988), 1 and 24.
89. "Is the Pope a Tory?," Times (Apr. 3, 2004), 6.
90. "Finding God in the Classroom," Times (Aug. 5, 1992), 10.
91. "Prince Blames 'God of Technology' for the Loss of Values," Times (Apr. 30, 1997), 6.
92. "Socialism by Consent Pledge in Statement," Times (Apr. 3, 1981), 2.
93. "Steel Union Faces Call for Review of Affiliation," Times (Apr. 13, 1981), 2.
94. "Blair Says Tradition Is Bedrock of Change," Times (June 8, 2000), 4. Blair's list of values: "Respect for the old for what it has still to teach; respect for others, honour, self-discipline, duty, obligation, the essential decency of the British character."
95. "A Party Eager for that Old-Time Religion," Times (Oct. 9, 1993), 8.
96. "How It Looks," Times (Jan. 1, 1994), 16.
97. "Major Fights to Salvage Back-to-Basics," Times (Jan. 7, 1994), 1.
98. "Housewife, Cinderella or Superstar?," Times (July 10, 1991), 12.
99. "Blair Stakes His Claim to Downing Street," Times (June 24, 1994), 8.
100. "Is Labour the True Heir to Thatcher?," Times (July 17, 1995), 17.
101. "No 10 Told to Cut the Spin as Poll Lead Dwindles," Times (June 12, 2000), 10.
102. "He Won't, They Can't, and We Know It," Times (June 16, 2001), 22. The analysis was performed by Encarta World English Dictionary.
103. "What Do They Know of England Who Only England Loathe?" Times (Oct. 1, 2001), 16.
104. "Tory Right Rallies around Faith, Flag and the Family," Times (July 25, 2005), 24.
105. "Announcements & Personal," Times (June 2, 1990), 13; "Announcements," Times (Sept. 1, 1990), 13.
106. "Multiple Classified Advertising Items," Times (Mar. 21, 2008), 74.

# 对挖掘十九世纪新闻数据的基础设施的反思

**Julianne Nyhan**，英国伦敦大学学院数字信息研究副教授
j.nyhan@ucl.ac.uk

**Tessa Hauswedell**，英国伦敦大学学院研究助理
t.hauswedell@ucl.ac.uk

**Ulrich Tiedau**，英国伦敦大学学院荷兰语副教授
u.tiedau@ucl.ac.uk

**摘要：**在此研究中我们比较和对照了我们（作为历史学家以及数字人文和信息研究者）利用大学提供的高性能计算设施开展大规模历史数据挖掘的经验及利用外部云存储平台和工具挖掘同样数据的经验。特别是，我们反思了我们在近期两大跨国数字人文项目中的经验：“不对称的相遇：引用 1815-1992 年欧洲文化的数字人文方法”，该项目由欧洲人文研究基金资助（2013-2016 年）；以及“海洋交流：在 1814-1914 年历史报纸馆藏中追踪全球信息网络”，该项目通过 2016 挖掘数据挑战项目社科人文跨大西洋合作小组资助（2017-2019 年）。作为这两个项目的一部分，我们尝试挖掘从 Gale 的文本数据挖掘（TDM）硬盘安装到伦敦大学学院高性能计算设施上的十九世纪历史报纸 OCR 文本。我们将其与我们后来通过 Gale 数字学术实验室（Gale Digital Scholar Lab）开展的类似挖掘工作进行了比较和对照。我们将我们的经验和观察置于对人文大数据分析支持设施的更广泛探讨和建议的语境之下，并讨论了两种方法的优缺点。我们的讨论围绕之前提及的设施问题，并反思了开展这类研究时人员自身的经验，代表了（数字）人文领域众多研究者研究步骤的变化。最后，我们以探讨公共和私营领域的研究投入结束，这些投入是支持进一步发展以及促进对大型数字档案的访问及批判性分析所必需的。

**关键词：**数字设施、文本挖掘、历史报纸典藏、高性能计算、重要文化遗产、数字人文、《泰晤士报》数字典藏、伦敦《泰晤士报》

## 前言

在此次演讲中，我们将比较和对照作为历史学家、数字人文学者和信息研究者，我们尝试利用院校的高性能计算设备挖掘大型、数字化历史报纸典藏的经验以及我们利用外部、云存储工具挖掘同样数据的经验。[1]

我们进行这些反思的背景是我们近期参与的两大跨国研究项目。“不对称的相遇：引用 1815-1992 年欧洲文化的数字人文方法”，该项目由欧洲人文研究基金资助（2013-2016 年），合作方来自英国、荷兰和德国，由乌特勒支大学协调。该项目利用数字技术对大型数字化报纸和杂志典藏开展纵向分析，并考察在 1815 年至 1992 年之间欧洲身份的文化层面发生了怎样的变化（AsymEnc, n.d.）。我们也利用了我们至今在“海洋交流：在 1814-1914 年历史报纸馆藏中追踪全球信息网络”（OcEx）项目中获得的经验，该项目由挖掘数据挑战项目跨大西洋合作平台资助（2017-2019 年）。这一项目汇集了芬兰、德国、墨西哥、荷兰、英国和美国的研究者，由西北大学协调。它“考察了跨国家和语言界线的信息流动规律，并与在大型数字化报纸典藏中的研究相关联”（Oceanic Exchanges, n.d.）。

在此次演讲中，根据一些文献，例如 Susan Leigh Star 和 Karen Ruhleder 的文章（1996 年）和美国学术团体联合会的资料（2006 年），我们定义基础设施为包含一套复杂的动态系统，同时也受其影响，不仅仅包含物理结构，也包含社会、文化和院校流程和背景。正如 Wolfgang Kaltenbrunner 所论述的：

> 基础设施的出现是在全球通行的标准和局地背景之间的矛盾以及自动化的技术程序与人类行为者执行的任务之间存在的矛盾可以被成功解决的情况之下。这一基础设施定义的重要结果在于，它是循序渐进地发展的，并非创造出的，而是逐步形成的。（Kalterbrunner 2015，211）

其他数字研究基础设施的定义使用了“数字生态系统”的类比，它提供“围绕社区构建的服务”（Blanke, Kristel, Romary 2015）。在我们这种情况下，我们关注的社区需求是人文学界的需求，特别是历史和传统研究领域的需求。相应的，在本文中，我们没有限定我们的观察对象仅为我们用于开展研究的计算设施，也反思了院校、技术、法律、社会文化和劳工组织等因素，这些因素有时助力我们的工作，有时阻碍我们的工作，但都是在人文研究的语境中需要考虑的。陈述了这些因素之后，我们最后反思了不断进步的、用以对文化遗产资源和文物进行大规模数据挖掘的基础设施，以及它们可能的未来发展方向。

首先从描述我们寻求计算机分析“《泰晤士报》数字典藏”的经验开始，这一典藏是 200 多年伦敦《泰晤士报》的在线、全文扫描复本，“包含 1785 年至 2010 年这一报纸每一期的每一页，是最受赞誉的十八世纪、十九世纪和二十世纪新闻报道来源之一”。（Gale, n.d.）

## 获取对“《泰晤士报》数字典藏”的挖掘访问权

在 20 世纪 90 年代末各国图书馆和商业公司启动报纸数字化项目（例如 Gooding 2014 历史项目）之后，我们已经看到数字化报纸典藏的创建和全球访问呈指数级增长。大量的历史报纸已经由公共机构（例如美国国会图书馆）、商业公司（例如 Gale/Cengage）和公私合作（例如大英图书馆参与的英国报纸典藏）数字化。尽管覆盖范围存在很多断档（Hobbs 2013，Milligan 2013），以及诸如 OCR 转录质量的差异这类问题（Smith and Cordell 2018，Tanner et al 2009），现有的国家图书馆、商业公司以及某些情况下的公私合作典藏已经为研究者和其他人群提供了丰富的资料可供钻研（见 Milligan 2019）。通过提供这些资料的平台，研究者可以进行简单和高级关键词检索。一些人更感兴趣进行较为复杂的检索或探讨令这些数字符号的结构性或其他组成元素（和关于它们的信息）可机器读取的元数据，他们面前有着各种各样的可能性。欧盟数字图书馆（Europeana）等数据提供方提供的 API [2] 和

数据转储[3]服务，允许研究者制作开放许可数据和查询的副本，或根据他们各自的需求转化数据。对于想要利用许可数据和元数据开展工作的人们，Gale 和 ProQuest 都为已采购产品的用户提供文本和数据挖掘硬盘（Gale 此项服务的背景资料参见 Fyfe 2016）。

挖掘“《泰晤士报》数字典藏”是我们参与项目的重要组成部分。为了完成这一工作，我们必须获取对典藏内含全文数据和元数据的访问。虽然我们的大学图书馆多年来一直订阅“《泰晤士报》数字典藏”，但我们仅能以标准检索的形式访问这一典藏。2014 年与圣智旗下 Gale 公司员工的一次偶遇，使得我们了解到通过外部硬盘，也就是文本和数据挖掘硬盘或 TDM 硬盘，我们可以获取《泰晤士报》数据的副本。硬盘上有近 2 TB 的“《泰晤士报》数字典藏”，时间跨度从 1785 年至 2010 年（并可扩展到 2013 年），一个 XML 文件为一期报纸，每一页有一个 TIFF 图像。我们为硬盘支付了三位数的适中价格，费用的产生主要是因为 Gale 公司为了让我们能够进行文本和数据挖掘而重新设置数据格式所做的工作，以及将硬盘递送到伦敦大学学院图书馆，以便我们访问。按照我们的理解，许可协议实际上视硬盘为伦敦大学学院图书馆已订阅 Gale 公司产品在订阅访问期间的服务延伸，这解释了合理、适中的收费。Gale 授权伦敦大学学院“免版税、非独家、不可转让 [ 原文如此 ]、不可转发许可的全球访问权，受协议中条款和条件的约束……允许授权用户仅为非商业目的开展文本和数据挖掘活动之用访问数据”（Gale 许可协议）。

在继续回忆我们采取的下一步行动，即加载数据并设法挖掘之前，我们将暂停一下，反思这种方式获取数据的优缺点。首先是优点：通过数字化项目创建封闭内容从而货币化文化遗产的风险担忧近年来已经被提出且很具说服力（例如 Prescott 2014，Darnton 2010）。例如，《智者委员会的新文艺复兴报告》（New Renaissance Report of the Comite´ des Sages）提醒人们，如果公共、私营及公私合作数字化之间的平衡被打破，就很可能会形成一个数字黑暗时代（Comite´ des Sages 2010）。然而，报告中清晰指出且大量的经验表明，互惠互利的公私合作方式将获益良多。除了这种方式在院校层面和馆藏层面上的益处，作为需要可靠十九世纪报纸文本和元数据以开展工作的个别研究者而言，我们发现与 Gale 公司合作的过程是高效且直接的。作为大学里的研究者，我们的本能是尝试通过大学基础设施或其他公共机构保证对我们所需数据的访问。私营提供商能够推动对资源的第一层级访问，提醒了我们开展这项研究所需基础设施的分散性。同时也必须注意到，我们享有足够的特权，能够获得采用这一路线所需的经费。也就是说，这种开展研究的方式存在某些限制条件。其次是缺点：我们是在一项多国合作研究项目中寻求使用 TDM 硬盘，但我们签署的许可协议是按国别划分的，并且与一所大学的收入来源相关联——让许可协议范围超出国家界线被证明存在更多的问题。因此，我们的工作方案无法扩大到跨国范围，即便研究项目本身是多国合作的。

## 数据加载与查询

无论如何，我们设法获得了“《泰晤士报》数字典藏”的全文数据和相关元数据，我们在伦敦大学学院已经做好了利用它开展研究的准备。与伦敦大学学院研究 IT 服务中心（RITS）建立联系也是在伦敦大学学院数字人文中心组织的一次数字科考中的意外收获。[4] 伦敦大学学院 RITS[5] 那时常常与开展大量计算研究、涉及大量数据的学科合作，迫切需要向这些学科以外的院系证明自己的价值。他们提出的目标是“帮助研究者超越理工科边界”（UCL 2017），那时他们刚刚首次与人文项目合作：“实现对大英图书馆与伦敦大学学院合作大规模数字化典藏的复杂分析”，该项目作为 JISC 研究数据源（JISC Research Data Spring）的一部分得到资助（Terras 2015）。伦敦大学学院研究计算平台服务

是伦敦大学学院研究 IT 服务中心的服务项目之一，提供“支持……伦敦大学学院的计算密集型研究，提供高性能和高计算产量的专业平台”。[6] 他们建立了三个计算集群（Legion、Myriad 和 Grace），向伦敦大学学院研究者开放，使用时无需为运行脚本或存储所耗费的时间付费。除此之外，服务中心对其他研究项目收取每天 350 英镑的费用，例如编写和运行自定义的文本挖掘脚本。毫无疑问，在研究 IT 服务中心同事的帮助之下，我们才能够在 Legion 上加载“《泰晤士报》数字典藏”的数据并运行查询，Legion 是“主机位于伦敦大学学院布鲁姆斯伯里数据中心的一个多用途计算集群。”[7]

第一步是将数据上传到伦敦大学学院研究数据服务中心的基于规则的集成数据系统（IRODS）中的一个保护账户中（参见 Wilson 2017），在此分配给我们 5 TB 的空间。但几乎从一开始，我们就明显看出这一系统并非专为支持人文数据和研究而建立。与以计算为基础的学科常常查询的大批量、高度结构化的数据（例如 Stevens 2013）不同，我们的数据如前所述包含无数的 XML 小文件和 TIFF 文件，以多重路径的形式分组。因此上传数据需要数天的时间。上传完成后，研究 IT 服务中心必须重新组合数据，以便我们能够将每一期报纸的 XML 文件按月、季度或年合并在一起。我们理解这一做法的必要性不仅仅是用于支持我们想要对数据开展的纵向分析，也是从计算的角度出发。此外还必须为 XML 概要添加一个解析器类，好处是从 JISC 研究数据源“实现复杂分析”项目（如前所述）中获取的代码可以在此重新使用。

数据最终安装好后，就可以重新使用代码和其他特别编写的 R 和 Python 脚本挖掘数据了。然而，我们仍要指出，计算设施并非真正为支持人文研究而设立。例如，在 Legion 上建立参考数据集背后的想法是它们一旦存储后不应发生变化。正如 David Smith 和 Ryan Cordell 指出的，“十九世纪报纸中错误单词的比例超过 40%，其他语言和更早期报纸上的错误率甚至更高”（2018，5）。在我们的许可协议允许的情况下，我们可以对 OCR 转录文本进行一些后期修订，或者甚至采用自动和手动修订相结合的方式，但重新安装数据所需的时间是令我们打消这些念头的重要因素。

继续到我们试图查询这些文献的经验，尽管我们非常感激伦敦大学学院研究 IT 服务中心，但我们发现我们使用大学内高性能计算（HPC）开展文本挖掘的努力从多个角度来讲都是问题重重的。我们将会呈现的例子仍然与我们想要开展研究的文化遗传数据集的特殊性有关，与我们想要从中找到答案的问题相关。如前所述，数字化的报纸文本结构相对较为混乱，OCR 错误更多。当根据特定语言规则处理多语言文献资源时，这一点特别具有挑战性，从语言学的角度对人文研究者和研究软件工程师都是挑战，需要多种计算语言学的技能。甚至对于单一语言文本的简单查询都可能涉及不可预料的因素，因为这样的查询可能返回不可控数量的结果或无法返回任何有意义的结果，预设被证明错误，需要更进一步的优化或拓宽检索参数。

例如，我们想要使用“《泰晤士报》数字典藏”数据硬盘开展的一项关于十九世纪报纸如何报道女性移民的研究。我们特别关注这些报纸如何讨论和报道那时女性向英国殖民地的移民，目的地例如南非、新西兰和澳大利亚。在我们最初的查询中，我们检索了包含女性移民社团名称的文章，这些社团都在十九世纪后半叶运营。在这段时期，移民逐渐得到了慈善团体的帮助、援助和管理，而慈善团体将其视为自己的使命，运送女性到海外，充实已成为英国殖民地的国家的人口（Constantine 1991, 95，Bush 1994）。我们最初的查询集中于绘制出在 1850 年至 1914 年之间报纸文章中这些移民社团被提及的次数。为了实现这一目的，我们调研并提取了移民社团的名称清单，并将此发送给了与我们合作这一研究项目的软件工程师。由于这些社团名称的长度（例如，“英国女性移民协会”）以及 OCR 质量的差异，每一个名称都根据各自的词密度被手动分配了一个固定的“容错度”。我们的目标，


*图 1. 图表显示出 1850 年至 1914 年“《泰晤士报》数字典藏”数据集中词组“Society for Promoting the Employment of Women”（促进女性就业协会）的标准化出现次数。*

一方面是避免结果的低度匹配，而另一方面是确保我们在这个过程中没有引入过多的误报。然而，当我们收到第一张绘制的曲线时，每个移民社团的结果之间存在巨大的差异。

“Society for Promoting the Employment of Women”（促进女性就业协会）的结果表明这一社团在几十年间一直被报纸文章提及和讨论。而“Female Middle Class Emigration Society”（中产阶级女性移民协会）的曲线呈现出它在 20 世纪 80 年代中期几乎不曾被提及。然而，关于这一协会的现有历史研究表明它在 20 世纪 80 年代之后非常活跃（Chilton 2007）。但这为什么没有表现在曲线上呢？很快我们就明白我们必须重新优化和调整“容错度”的设置，以产生更多可靠的、有代表性的结果。然而，没有对每一年每一期“《泰晤士报》数字典藏”OCR 质量的清晰了解，显然这个过程需要对这大约 25 个移民社团的列表迭代运算多次。

许多需要处理噪音数据的数字人文项目都强调，研究过程应当包含迭代阶段，调整清理方式或优化数据、创建和修改查询条件，以及进一步优化研究的预设假说（参见 Schöch 2013）。但要记得这项工作每天的花费是 350 英镑。尽管整个项目的研究资金充裕（高达六位数），但刨去院校收取的日常管理费用，项目经费在多个项目合作方之间平分后，资源仍是很紧张的。事实的确如此，现实很快表明如果要采取我们理想中的迭代方法，花销可能超出项目经费。因此，我们以这种方式开展研究的经验是在试验和错误中前进。当 TDM 硬盘送达时，我们感觉充满机遇的世界正在打开，拥有了数个

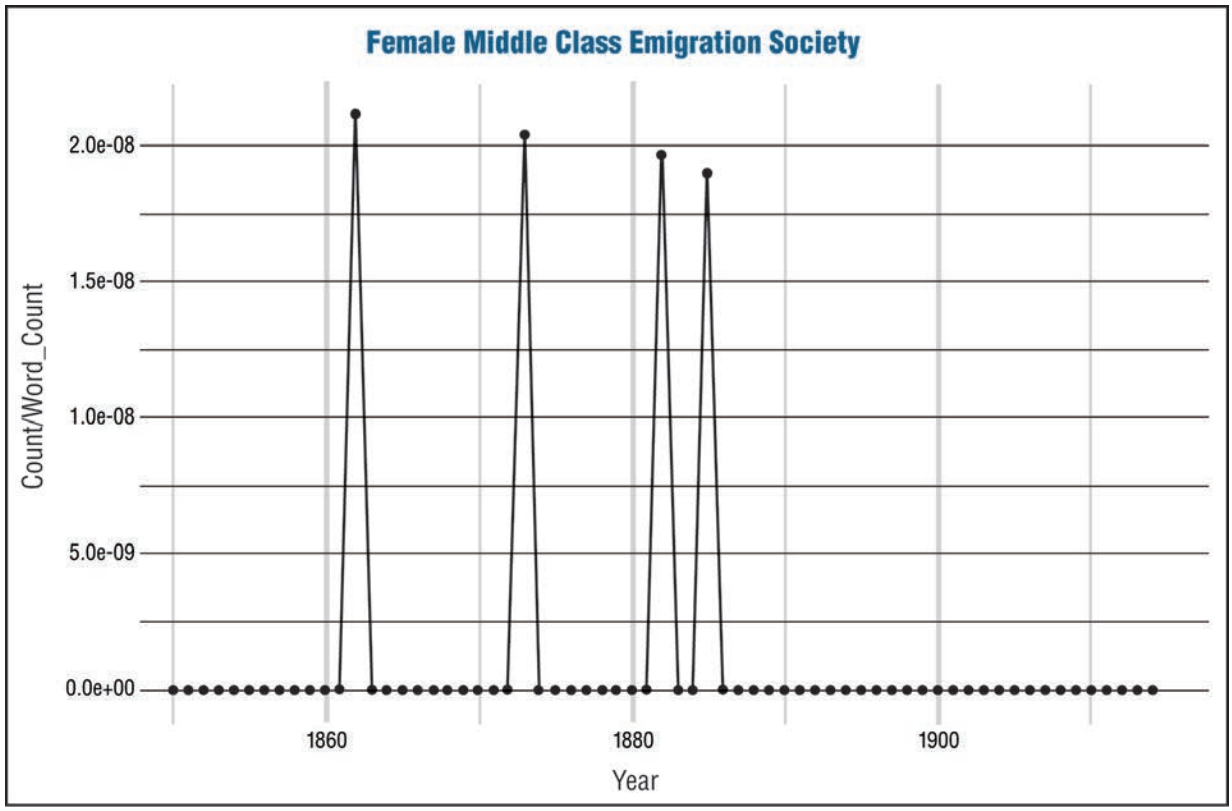



*图 2. 图表显示出 1850 年至 1910 年词组“Female Middle Class Emigration Society”（中产阶级女性移民协会）的标准化出现次数。*

TB 的内容。然而，正如我们总结过的，应对随后出现的院校、后勤、财务、官僚主义等障碍花费了我们大量的时间，而这些都无法充分体现在项目的发表成果中。

## Gale 数字学术实验室

那么，从多个角度看，像“Gale 数字学术实验室”这样的云存储工具提供了前进一大步的动力。以上描述的困难，例如保证和协商对下层数据的访问、安装数据、思考数据的独特性或限制因素以及如何在计算中给予相应的考虑、对数据进行查询以及尝试开展迭代运算整合不断出现的新见解，都在不同程度上得以解决（假定你所在院校订阅了恰当的产品）。

Gale 数字学术实验室（Gale Digital Scholar Lab，DSL）提供了高度实用性的界面，可以对数据应用多个主要的文本挖掘工具：聚类、N 元语法、情感分析和主题建模。平台也支持命名实体识别和词性标注，且添加了标签的数据可以下载后进行更多变换，例如网络分析。实验室平台使用方法简单直接，因此吸引了探索性的研究和有趣的文本探讨。它以普及这些数字方法的形式出现，降低了数字学术研究的入门门槛，特别是对于不熟悉文本挖掘的学生而言。“Gale 数字学术实验室”提供的有关这些工具的解释是简洁而实用的，不会给更为随意和经验不足的用户造成过大的压力。因此，这一平台适用于教学目的。我们也注意到“Gale 数字学术实验室”带来的更多可能性，它可以帮助将数字研

究手段进一步嵌入到一些人的学术研究中，没有这一平台工具这些人可能不会采用类似的研究方法。从这个意义上讲，“Gale 数字学术实验室”为提高数字人文在各个学科中的可见度做出了贡献。

此外，“Gale 数字学术实验室”也在一定程度上满足了更资深用户和研究者的需求。我们注意到，实验室平台的功能包括创建自定义内容集以及个性化设置包含相关文章子集的文件夹，管理和组织形式都简单且直观。这些个性化语料库的检索历史都可以保存。在取得实验室平台的试用访问权限时，我们曾对它仅提供对授权文献资料的访问感到担心，并关注实验室平台是否会加入可同时分析开放内容和授权内容的新功能。但按我们的理解，平台将很可能会稍晚增加外部的开放许可资源（当然，一些人不愿看到在这类专门平台上使用开放许可的文献资料，但这是另一个问题）。

而且，因为一次文献原文图像和 OCR 文本并排显示，用户可以直接对比二者，评价 OCR 与原稿相比的质量。此外，用户能够获取详细的元数据信息和 OCR 置信度评级，为如何解释和处理他们得到的结果提供一些指引。

关于 OCR 置信度的信息通常不会出现在标准的数字化报纸检索界面上，但好消息是“Gale 数字学术实验室”将这一信息整合其中，帮助研究者更好理解存在问题的典藏文献。我们相信，这种对转录文本质量和数据出处的公开透明，在培育对数字资源的批判性使用中至关重要，我们希望未来更多的内容提供商能考虑加入这类功能。

而且，导出表格数据的功能便于更资深的用户管理他们的结果并将数据用于其他可视化程序或仅仅只是记录他们的工作流程。这对于想要发表他们的结果以及需要可重复和可重现他们的研究发现的研究者非常重要。同样，按照现在的情况，当研究者使用标准检索界面时，通常很难按他们需要的方式记录工作流程。因此“Gale 数字学术实验室”在某种程度上帮助学者实现这些目标，因为它可以实现研究的透明度，呈现出研究者如何获得文本挖掘分析结果的过程。

无论怎样，我们并非要暗示这是 Gale 的主张，像 Gale 数字学术实验室这样的云存储工具的存在并不能免除使用高性能计算或其他基础设施开展深入文本挖掘研究的需要。

为了较为详细的探讨我们使用“Gale 数字学术实验室”的经验，我们将回到之前关于女性移民社团的例子中。我们对“《泰晤士报》数字典藏”进行了全文检索（用户也可以选择在文档标题或出版地点等字段中检索），限定时间为 1850 年至 1914 年。我们进一步将检索的内容类型限定为“报纸和期刊”。与词组“Society for Promoting the Employment of Women”（促进女性就业协会）完全匹配的结果数量是 54（不完全匹配的话是 58）。这与我们从“《泰晤士报》数字典藏”硬盘上获得的结果（162 个匹配结果）存在明显的差异。造成这一差异的原因显然是实验室平台设置的容错率与我们手工对“《泰晤士报》数字典藏”硬盘设置的容错率不同。我们无法改变“Gale 数字学术实验室”的容错率，但在检索中我们可以使用正则表达式。这为我们提供了一种强制调整实验室平台容错率的方法，但这种方法并不适用于数量较多的查询。

对于返回的结果，我们发现实验室平台的界面和对元数据的呈现有助于我们快速概览这 54 条出现检索词的结果中包含哪些类型的文献资料。我们能够快速看到检索结果中存在的文档类型：文章（34 篇）、广告（10 篇）和读者来信（10 篇）。涉及的主题也分别列出：共计 11 个主题，前三个主题是工作女性（9 篇）、协会（5 篇）、商会（1 篇）。用户可以分别查看每个主题下的文章。结果中还返回了一个插图作品，具体而言是一个表格。出现检索词的文章的时间跨度是从 1860 年到 1913 年。OCR 的置信度水平参差不齐，从高于 90% 到低至 40%。在将文档加入到一个内容集中之后，我们看到这些文献的作者

共有 10 位。在探讨这些结果时，我们发现我们可以使用实验室平台生成这些数据的可视化图形，可以为我们读取数据提供一些有趣的方式。例如，我们可以生成从 1860 年到 1914 年文章情感的图表，可能是进一步研究的方向（见图 3）。

然而，实验室平台目前没有提供路线，能够让我们按研究所需的方式绘制数据。实际上，在“Gale 数字学术实验室”中，我们无法写下具体的问题，因此实验室平台对我们想要开展的研究用处有限。也就是说，我们显然可以想象这样的情景，研究者使用实验室平台探讨和批判性评价他们的数据集，而这一过程将辅助和提升后期使用更昂贵、更定制化、基于高性能计算的研究方法分析他们的数据。

对于实验室平台，我们也要指出目前它无法实现一种工具的输出结果作为另一种工具的数据输入，形成串联的数字工作流程。实现这一目的需要的是高性能计算设备上的自定义脚本，正如我校研究 IT 服务中心所提供的（或亚马逊网络服务这样的商业方案）。同时，“Gale 数字学术实验室”目前对协同工作的支持较少。因此，在某些情况下，例如基础课程的教学和个人使用，它会非常有用，但目前并不适合较大规模且分散在世界各地的研究团队开展的长期文本挖掘工作。这些项目将需要专为较大团体账户设置的功能，各个团队成员均可以存储和访问研究结果，而每个人也都可能需要一些自定义的功能。

当然，我们也知道没有哪个工具能够满足每一类型研究者的需求，更不要说每一类型研究团队的

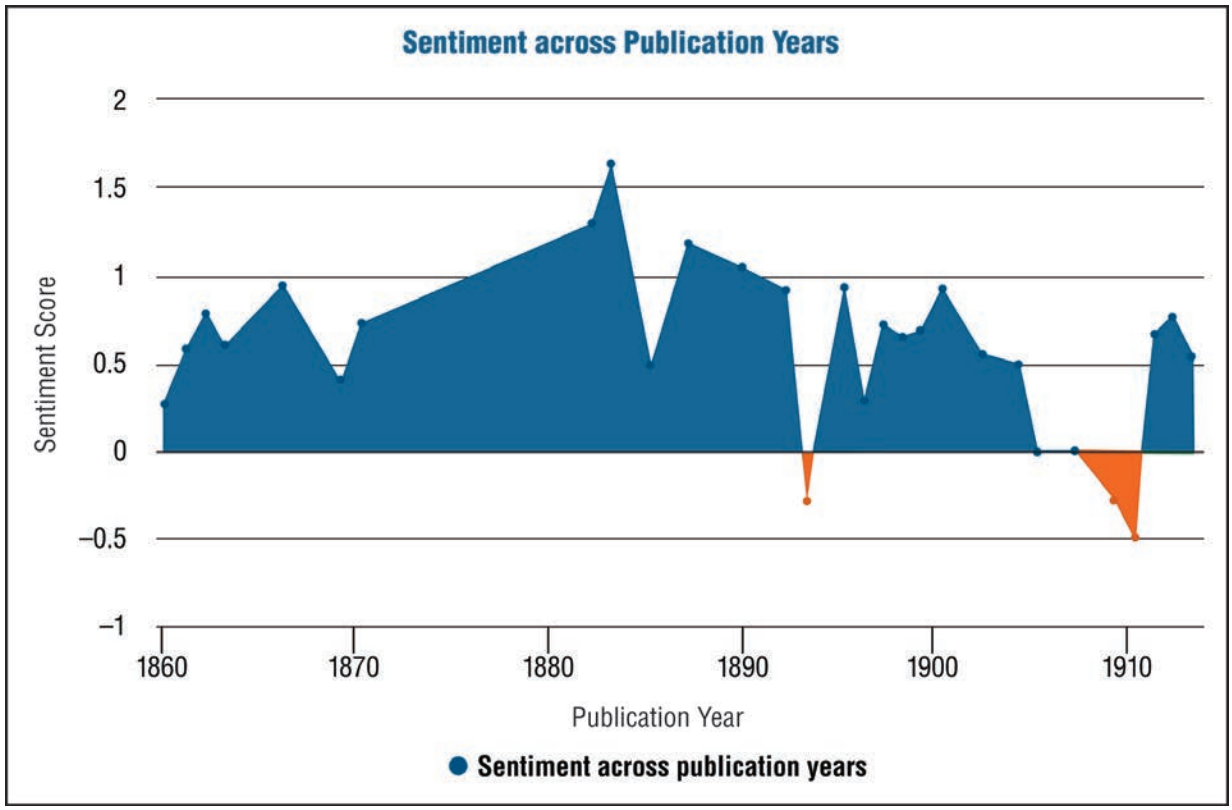



*图 3. 图表显示出 1860 年至 1914 年发表文章的情感值。*

需求。因此，事实上很可能是这样，基于云存储的平台将作为一名研究者能够信任的媒介，在此按照严格且可靠的标准进行“标准”文本挖掘工作。对于绝大部分的用户，这已足够，而需要自定义脚本的复杂操作则会继续在其他设施上开展。

## 结论

最后我们将对我们寻求使用大学高性能计算设施挖掘大型、商业数字化十九世纪历史报纸典藏的经验做总体的反思。我们显然不是最早提出在尝试开展这类工作的过程中遇到重重困难的研究者（在这种情况下显然是更大规模的泛欧洲项目，例如 CLARIN 行动 [Wynne 2013]，以及开源的网络工具 Voyant[Rockwell and Sinclair 2016]）。然而，即便如此，“数字人文研究需求已经变得更为复杂，因为人文学家与许多其他领域的学者一样，比以往更可能需要应对大型数据带来的研究问题”（Dombrowski and Lippincott 2018），支持这类计算密集型人文研究工作所需的基础设施发展形势并不稳定。我们要做些什么呢？本文曾提及的“实现大规模数据分析”项目提出的建议如下：

1. 投资提高研究软件工程师的能力，以部署和维护开放许可的大型数字典藏，覆盖美术馆、图书馆、档案馆和博物馆（GLAM）领域，支持艺术、人文、社会科学和历史学的研究

2. 投资培训图书馆员，与人文教师合作开展这些初步的查询，以支持对所收集数据子集的研究工作，并记录和管理生成的代码和获得的数据（Terras et al. 2018, 463）。

这些建议很有帮助，我们希望它们能够被采纳。无论如何，我们认为，对于我们的经验中以及这些建议中所谈及的计算密集型人文研究工作，另一个重要方面是要更加关注如何拓宽研究策略和预见性的视野。

伦敦国王学院的数字实验室罕见地在专门的部门长期聘用了研究软件工程师，他们都具备清晰的专业发展路径，许多拥有丰富的人文和数字人文研究经验以及与人文和数字人文研究者合作的经验，能够为文化遗产和艺术及人文资源进行研究软件的建模与执行。[8] 而在其他地方，这项投资则相当的不均衡。因此，是的，在这个例子中需要研究软件工程师的参与以开展对开放数据集的研究，但在更多的情况下，随着人文研究逐渐向数字化方向发展，研究软件工程师（以及他们所参与和实现的基础设施）的作用将变得至关重要且涉及面更加广泛。目前，人文研究者很难在大学院校中发现研究软件工程相关的职位，考虑到大学中这类职位往往不清晰的职业路线以及研究软件工程师能够在大学以外得到更好的工作机会，这样的职位的流动性可能会很高。这一点对于涉及其中的个人而言是令人遗憾的，从管理限定期限经费的角度也是困难的。

因此，我们相信，我们在尝试挖掘大型数字化开放或封闭文化遗产典藏过程中遇到的种种困难，代表了许多大学面临的困境，如何充分地响应数字化转变带来的更大范围的投资要求。回到 Kaltenbrunner 对基础设施的讨论以及这一话题中涉及到的全球和局部问题的纵横交错，显然我们在局部环境中面临的问题也相应影响着全球的研究者。从更大范围的对话和语境中看，在尝试挖掘大型数据集时面临的困难可能是全球性的，人文学者在努力为其做出贡献时遇到了阻碍。研究工作中的困难不仅仅限制了人文学者需要大型资源的问题，也降低了人文学科的存在感，甚至可能阻碍大学之间及其与商业公司和创意产业等外部参与者之间发展合作关系的计划。这些困难也限制了我们在更大范围的数字化产业中建立和从事数字文化遗产经济项目的机会。

最后回到“Gale 数字学术实验室”，它最令我们关注的方面之一可能就是启发我们思考数字人文领域的创新与开拓，以及与行业合作的可能性——当然也存在不利之处，以找到较为少见、超出大学范围的方式推动开展研究所需基础设施的重大进步，例如在对数字文化遗产数据进行计算密集型分析的研究中。在这种环境下实现的互惠合作可能会产生新的方法，让大学相信研究软件工程师这类职位的价值，继续投资支持人文领域数字研究基础设施的必要性，以及数字人文在多个方面的积极影响。通过公私合作对文化遗产文献资料的数字化，公私双方都获益良多，了解了共同工作的益处与不足。我们必须继续吸取这些经验教训，同时仔细思考我们对大型、开放和许可文化遗传典藏数据开展的数字化研究，还需要哪些类型的基础设施和访问权限。

## 注释：

1. The views that we present in this paper are our personal views as individual researchers; they should not be taken to express the views or opinions of University College London.
2. https://pro.europeana.eu/resources/apis [accessed May 1, 2019].
3. https://pro.europeana.eu/post/experimental-text-dumpsfrom-europeana-newspapers [accessed May 1, 2019].
4. https://www.ucl.ac.uk/digital-humanities/.
5. https://www.ucl.ac.uk/research-it-services/ [accessed May 1, 2019].
6. https://www.ucl.ac.uk/isd/services/research-it-services.
7. https://wiki.rc.ucl.ac.uk/wiki/Cluster_Computing.
8. See https://www.kdl.kcl.ac.uk/.

## 参考文献：


American Council of Learned Societies. 2006. Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. Accessed June 13, 2009. https://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf.

Anderson, Sheila, and Tobias Blanke. 2012. “Taking the Long View: From e-Science Humanities to Humanities Digital Ecosystems.” Historical Social Research 37, no.3: 147–64.

AsymEnc (website). Accessed May 1, 2019. http://asyenc.eu.

Blanke, Tobias, Conny Kristel, and Laurent Romary. 2015. “Crowds for Clouds: Recent Trends in Humanities Research Infrastructures.” Accessed July 8, 2019. http://arxiv.org/abs/1601.00533.

Bush, Julia. 1994. “‘The Right Sort of Woman’: Female Emigrators and Emigration to the British Empire, 1890–1910.” Women’s History Review 3, no. 3: 385–409.

Chilton, Lisa. 2007. Agents of Empire: British Female Migration to Canada and Australia, 1860s–1930. Toronto: University of Toronto Press.

Comite´ des Sages. 2010. The New Renaissance. Report of the ‘Comite´ Des Sages.’ Reflection Group on Bringing Europe’s Cultural Heritage Online. https://ec.europa.eu/digital-single-market/sites/digital-genda/files/final_report_cds_0.pdf.

Constantine, S. 1991. “Empire Migration and Social Reform 1880–1950.” In Migrants, Emigrants, and Immigrants: A Social History of Migration, ed. By Colin G. Pooley and Ian Whyte. New York: Routledge, pp. 62–87.

Darnton, Robert. 2010. The Case for Books: Past, Present, and Future. New York: PublicAffairs.

Dombrowski, Quinn, and Joan Lippincott. 2018. “Moving Ahead with Support for Digital Humanities.” EDUCAUSE Review, March 12, 2018. Accessed July 8, 2019. https://er.educause.edu/articles/2018/3/moving-ahead-with-support-for-digital-humanities.

Fyfe, Paul. 2016. “An Archaeology of Victorian Newspapers.” Victorian Periodicals Review 49, no. 4: 546–77. https://doi.org/10.1353/vpr.2016.0039.

Gale. “The Times Digital Archives 1785–2013.” https://www.gale.com/intl/c/the-times-digital-archive.

Gooding, Paul. 2014. “Search All about It: A Mixed Methods Study into the Impact of Large-Scale Newspaper Digitisation.” PhD diss., University College London.

Hobbs, Andrew. 2013. “The Deleterious Dominance of The Times in Nineteenth-Century Scholarship.” Journal of Victorian Culture 18, no. 4: 472–97. https://doi.org/10.1080/13555502.2013.854519.

Kaltenbrunner, Wolfgang. 2015. “Scholarly Labour and

Digital Collaboration in Literary Studies." Social Epistemology 29, no. 2: 207–33. https://doi.org/10.1080/02691728.2014.907834.

Milligan, Ian. 2019. History in the Age of Abundance?: How the Web Is Transforming Historical Research. Montreal: McGill-Queen's University Press.

Milligan, Ian. 2013. "Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010." The Canadian Historical Review 94, no. 4: 540–69.

Oceanic Exchanges (website). Accessed May 1, 2019. https://oceanicexchanges.org.

Prescott, Andrew. 2014. "I'd Rather Be a Librarian: A Response to Tim Hitchcock, 'Confronting the Digital.'" Cultural and Social History 11, no. 3: 335–41. https://doi.org/10.2752/147800414X13983595303192.

Rockwell, Geoffrey, and Ste´fan Sinclair. 2016. Hermeneutica: Computer-Assisted Interpretation in the Humanities. Cambridge, MA: The MIT Press.

Scho¨ch, Christof. 2013. "Big? Smart? Clean? Messy? Data in the Humanities." Journal of Digital Humanities 2, no. 3 (Summer 2103). Accessed June 10, 2019. http://journalofdigitalhumanities.org/2-3/bigsmart-clean-messy-data-in-the-humanities/.

Smith, David A., and Ryan Cordell. 2018. A Research Agenda for Historical and Multilingual Optical Character Recognition. Boston: Northeastern University; NULab; The Andrew W. Mellon Foundation. Accessed June 8, 2019. https://ocr. northeastern.edu/report/.

Star, Susan Leigh, and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." Information Systems Research 7, no. 1 (March 1996): 111–37. https://pdfs.semanticscholar.org/9cfc/d2dfe7927451f2c39617e6ac0aa499fd2edb.pdf.

Stevens, Hallam. 2013. Life Out of Sequence: A Data-Driven History of Bioinformatics. Chicago: University of Chicago Press.

Tanner, Simon, Trevor Mun˜ oz, and Pich Hemy Ros. 2009. "Measuring Mass Text Digitization Quality and Usefulness." D-Lib Magazine 15, no. 7/8 (July/August 2009). https://doi.org/10.1045/july2009-munoz.

Terras, Melissa M. 2011. "The Rise of Digitization: An Overview." In Digitisation Perspectives, edited by Ruth Rikowski, 3–20. Rotterdam: Sense Publishers.

Terras, Melissa M. 2015. "Bluclobber, or: Enabling Complex Analysis of Large Scale Digital

Collections." UCLDH Blog. Accessed May 7, 2015. https://blogs.ucl.ac.uk/dh/2015/05/07/bluclobber-orenabling-complex-analysis-of-large-scale-digitalcollections/.

Terras, Melissa, James Baker, James Hetherington, David Beavan, Martin Zaltz Austwick, et al. 2018. "Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High-Performance Computing, and Transforming Access to British Library Digital Collections." Digital Scholarship in the Humanities 33, no. 2: 456–66. https://doi.org/10.1093/llc/fqx020.

UCL. 2017. "Partners in Time: HPC Opens New Horizons for Humanities Research." Research IT Services, October 24, 2017. Accessed 8 July 2019. https://www.ucl.ac.uk/research-it-services/news/2017/oct/partners-time-hpc-opens-newhorizons-humanities-research.

Wilson, James A. J. 2017. "The Research Data Storage Service at UCL – A LEARN Case Study." In LEARN Toolkit of Best Practice for Research Data Management, Leaders Activating Research Networks (LEARN), pp. 78-81. https://doi.org/10.14324/000.learn.00.

Wynne, Martin. 2013. "The Role of CLARIN in Digital Transformations in the Humanities." International Journal of Humanities and Arts Computing 7, no. 1–2: 89–104. https://doi.org/10.3366/ijhac.2013.0083.

# 人文数字分析的教学

**Ryan Cordell**，美国西北大学英语副教授
r.cordell@northeastern.edu

**摘要：**这篇论文主张“数据人文主义”，与计算机科学和数据科学有着明显的区别。向人文专业的学生仅仅传授数据科学的技能会错失我们的研究领域和我们的学生为更大范围内的数据和文化探讨做出独特贡献的机会。构想这一与众不同的数据人文主义的最重要原因是希望走出校园、迈入广阔天地的学生能够反过来改变它，而不是固守在现有的框架之内。这一数据人文主义应当是探索性的、往复前进和对话性质的，可以认为它的目标是将学者的注意力引导至数字化典藏之内和之外出其不意的地方，同时提出有关这些典藏及其在研究中缺席的新问题。本论文的最后提出了四项具体的建议，如何在本科生和研究生课堂中培育人文视角的数据处理。这种教学工作包括从创造性出发，利用人文主义而非预先打包的数据开展教学，研究语料库的视野需超出计算方法的范围，以及数据处理思维方式的培养并不仅限于某一特定的方法。

**关键词：**数据、计算分析、编程、教学法、数据科学、数字人文

## 前言：人文学科数据

Lorraine Daston 和 Peter Galison 在 2017 年出版的《客观性》（Objectivity）一书中，尝试追溯客观性作为一种科学家们的概念、理想和道德框架，在十九世纪出现之后的发展轨迹。这一著作主要围绕在这一时期内科学家形象的变化。Daston 和 Galison 认为，在十八世纪和十九世纪早期，科学理想是“自然的真相”，其中某些自然样本特别有用，因为它们呈现出并帮助构建了理想型：并非独独这片叶子，而是这一类型的叶子。在这种体系下，科学插图并没有尝试去重现单个、不完美的样本，而是从多个样本中归纳总结，并描绘出一个理想型。

因为摄影的出现，“自然的真相”框架发生了一些变化，推动科学家转向新的机器客观性理想。在早期关于插图与照片各自优点的辩论中，插图被标榜为优越于相对较为原始的照片，因为绘画和雕刻技术能呈现比十九世纪模糊不清的照片更多的细节。然而，照片在十九世纪逐渐占据了科学图片这一领域，因为它被人为涂改的可能性更低，对艺术家（或实际上的科学家）想象力的依赖性更低。正如 Daston 和 Galison 所解释的，

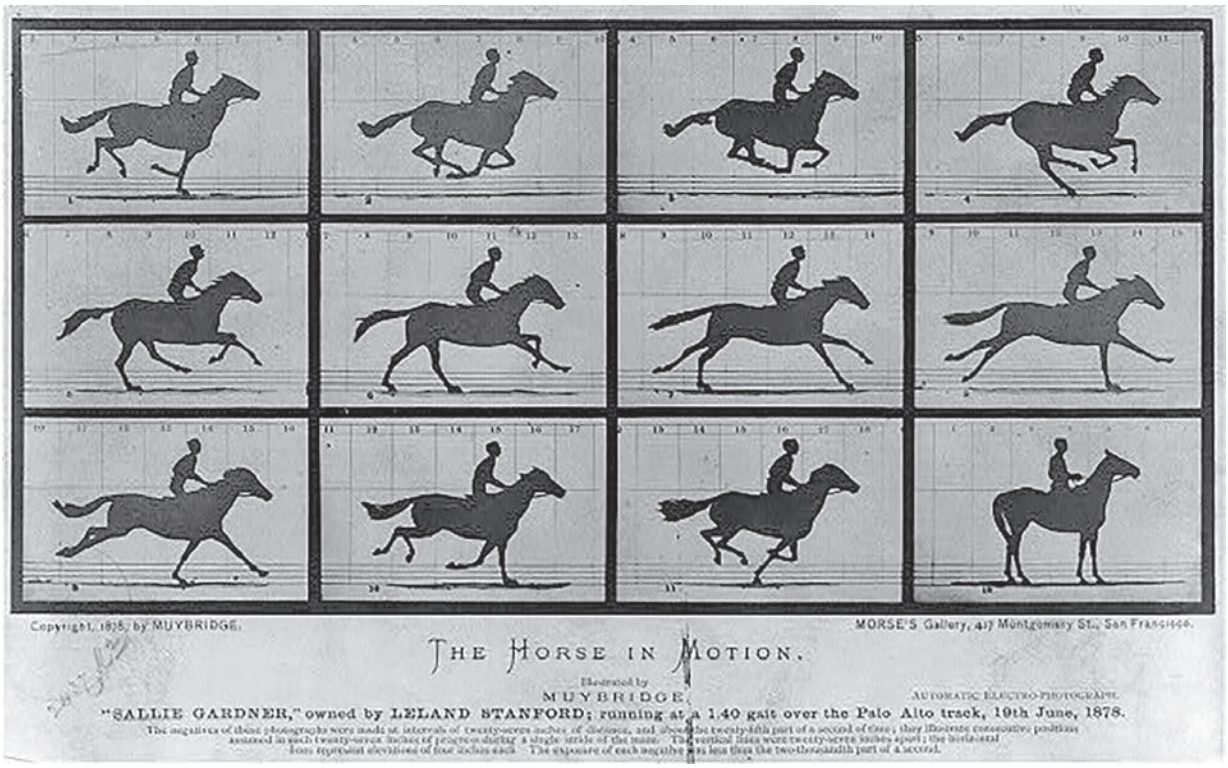
> 这是一项意志控制的内部斗争，这一意志给予机器客观性很高的道德论调……照片是一种类型的机器图片，成为了不干涉主义客观性各个方面的象征……这并非因为照片显然比手工绘制的图片更忠于自然——许多绘画如果使用了色彩，比早期的照片更接近它们的记录对象，而是因为照相机显然消除了人类的介入。（187）

科学家越来越担心自己的顽固意志可能会玷污他们研究发现的真实性，而机器方法的图片制作则提供了一个解决方案。

最终，照相机也终于能够看到仅用人眼无法观察到的自然。例如，随着拍照所需曝光时间的缩短，它重新解读了运动。埃德沃德·迈布里奇（Eadweard Muybridge）在 1878 年至 1884 年拍摄的著名赛马照片，证明了马匹在飞奔过程中的某一个时刻的确四蹄同时离地。在这一系列照片和随后的照片中，摄影技术令科学家对自然的概念发生了翻天覆地的改变，科学家们能够冻结住运动，单独分析运动的每一个组成部分。换言之，这些照片将流动的动作或一个过程转换成了数据：可观察、测量或分析的离散单元。

科学史上的这一时刻让在人文学科领域从事计算文本分析的人们回想起一篇试金石般的文章。在《文本：可批量寻址的对象》（Text: A Massively Addressable Object）一文中，Michael Witmore 认为“将 [ 数字 ] 文本对象与其他对象区别开来的”是：

> 它在不同规模级别上都是可以批量寻址的。可寻址（Addressable）在这里意味着人们在一定




*埃德沃德·迈布里奇的“运动中的马匹”，1878 年末拍摄。感谢美国国会图书馆提供。*

> 的抽象级别上可以查询文本中的一个位置……那么，书或物理实体就是众多地址级别中的一个。退回到更大的范围内，我们可能将一种体裁的作品当做相关的地址级别。或者我们可以探讨印刷书籍中的单个行，每一行中的所有名词，每隔两行的每第三个字母。所有这些变量都意味着地址级别的巨大灵活性。而且更具吸引力的是，当我们创造数字化文本集群时，我们的地址模式变得越来越抽象：例如，集合中所有项目中的具体名词，或被赫明斯（Heminges）和康德尔（Condell）在《第一对开本》（First Folio）中认为是“历史”的所有项目。每一个级别都是一个临时的统一体：为了地址而稳定，也因为是地址的对象而稳定。书籍是这样的临时统一体。电话簿里所有的专有名词亦是如此。

正如 Whitmore 承认的，通过一些学术手段，例如目录索引、词语索引或旁注，“文本实体在被数字化之前就已经是可批量寻址的”。地址的灵活性则是全新的概念。突然之间，书本甚至是图书馆，不需要转变为新的形态就成为了数据：书本或图书馆本身就可以是数据。

显然，其他形式的人文数据在计算机出现之前也已经存在：例如图书馆目录卡片或主题书目就是数据库的一种形式。但就像迈布里奇的照片，计算机重新塑造了我们与文本之间的关系，让它可彻底分离并可操作：一个统一体可以按其组成部分在各个层级上重新分解，就如 Whitmore 所描述的。很多人文问题的主要研究对象——文本仅是其中之一，因为我最为了解它，所以在此我一直围绕着它——可以被认为是可观察、可测量或可分析的离散单元。

## 走向数据人文主义

本文提出并尝试回答两个相关的问题：首先，作为人文课程一个教学领域的“数据分析”到底意味着什么？其次，我们如何区分人文数据分析与自然和社会科学中的类似研究方法？正如 Datson 和 Galison 所述，科学家转向使用机器是为了避免人为判断的失误，在数字人文的一些节点上，计算的使用是为了解决数十年来困扰人文学者的一些意志问题。我们知道，英国文学的经典地位是由其守门人的身份以及这些书中所蕴含的美德所共同塑造的。女性主义、后殖民主义及其他理论流派让我们认识到我们的偏见限制了我们的批判视野，导致了经典的萎缩，不足以代表所有的人类身份或文学。

诸如遥读、宏观分析或文化解析这样的方法，有时作为能够绕开不可靠的人类意志的方法而被提出，正是人类意志通过一种机械客观性建立了现有的经典。人类因为各种原因而选择性地仅看到某些方面，其中就包括我们一直尝试尽可能减轻的、有意或无意的偏见。相比之下，计算机则不受这些先入之见的干扰，专注于寻找文本规律。从这个视角看计算研究，机器更不容易受到社会、政治或身份操控塑造经典的影响。

当然，如果我们更近距离地去看，我们发现这一论点只是将问题转嫁到了书写代码的人身上，而不是塑造经典的人身上。正如 Safiya Umoja Noble [1]、Cathy O'Neil [2] 和 Virginia Eubanks [3] 等多位学者充分展示的，计算机程序常常满载了它们创造者的偏见和疏漏。无论是故意为之还是无心之举，设计用以提高商业或政府效率的软件通常复刻了结构性的不平等，例如白人至上。Ben Fagan（2016）等学者们已经证实，许多大型数字化项目也存在同样的现象，文献选择的过程复刻了主流文化的主导地位。在这样一种现实下，我建议我们不能指望以自动化的方式实现更公平的经典塑造。相反，我们应当倡导人文数据分析（HDA），主要目的是让我们注意到数字化与批判性分析之间的空缺。对我而言，这一可能性最吸引人的一点在于 Lauren Klein 对主题建模的描述（2015）“搅动档案资料的一项技术”，

反之，在一种双向关系中，它也会被档案资料搅动（需旁注的是，Klein 与 Katherine D'Ignazio 正在写作的新书《数据女性主义》（Data Feminism）[4]，完成后将成为这一领域的重要检验标准）。

在 2015 年 Klein 的阐述中，人文数据分析的方法让研究者和学生们能够发现数据集合中的关联或路径，而这在现行阅读中并不易被发现。同时，至关重要的一点是，我们所使用的数据集的性质决定了我们能从中发现什么样的关联和路径，因此“领域专家们……必须能够探究模型呈现出的语义关联”。为了理解主题建模，研究者必须理解工具训练所使用的数据集，才能明确模型存在的各种可能性，并注意到缺失了哪些可能性。

Klein 描述的基本计算分析结果并非统计模型——尽管他们在分析过程中可能使用了这种模型，而是研究者与数据之间的推测性对话。我们可以在 Stephen Ramsay 的《阅读机器》（Reading Machines）一书中找到类似的人文数据分析概念（2011a）。Ramsay 因一篇争议性的三分钟意见书而为人所知，“谁入局、谁出局”（Who's In and Who's Out）这篇论文发表在 2011 年 MLA 大会的一次圆周会议上。这次会议令人难以忘怀，所有的与会者都被要求起草一篇争议性的意见书，以发起讨论——而他就在意见书中提出数字人文学者应当了解如何编程。然而，在《阅读机器》一书中，Ramsay 对文学研究中构成“算法考证”的要素可能有哪些，提出了深入细致且坦白地说更具特色的观点。Ramsay 尤其排斥计算工作的科学框架，而青睐于对话框架：

> 如果文本分析以事实核查之外的方式参与文学评论研究，那么它必须帮助评论人揭示各种可能的解读……（10）无论科学辩论的范围可能有多广，无论多少种诠释被提出，我们仍假设有待我们解决的问题仅有一个答案（或唯一一组答案）。文学评论不存在这样的假设。在人文学科中，对某一问题的研究成果多寡常常严格取决于它对我们待解决问题所提供的不同解决方法的分歧程度。我们并非尝试解答伍尔夫。我们在努力确保对《海浪》（The Waves）的讨论能够延续下去。（15）

在这些段落中，Ramsay 强调各自不同的诠释方式支撑起人文研究。在本文中，我会依循 Klein 和 Ramsay 的框架，探讨探索性的、迭代和对话式的人文数据分析，目的是将我们的关注点引向数字化资料集内外出乎我们意料的地方，同时对这些资料集以及它们在研究中的缺席提出新的问题。

这种“数据人文主义”并非完全不同于“数据科学”一词所囊括的研究方法。至少在美国，全新的数据科学硕士研究生课程似乎每天都在涌现，随之而来的是对于这类文凭能够获得高回报工作机会的承诺。仅从修辞学的角度看，数据科学似乎类似于、但并非完全等同于计算机科学：需要编程，可能用到更适合统计学的编程语言，例如 R 语言，但焦点并非搭建系统或编写软件。相反，数据科学家擅长“适时”编程，探讨经济、政治或民众数据集，找出政府、企业、非盈利组织、新闻业或其他行动方立即能够使用的趋势。数据科学家的重要特质是灵活性，因为他们的工作是在各个领域和各种数据集中寻找有意义的规律。尽管这些特质在人文数据分析中也很有帮助，但我建议数据人文主义必须从想象和实践上都有别于数据分析，无论是内容还是方法学。

数字人文学者常常被指试图将文学、历史及相关领域转变为计算机科学的分支。这一指责在很大程度上并非如此。特别是，数字人文研究工作的诠释目标与计算机科学截然不同，许多将这两个领域相提并论的言论来源于对计算机科学家工作的错误假设。而且，我们传授计算能力的方式、我们使用的特定计算技术，以及我们描述计算工作结果的方式，都完完全全来自于语料库语言学、数据科学和相关社会科学领域。我们很大程度上未能建立起专门的计算数据分析的人文方法，这类方法的严重缺

失已经成为了我们这个领域发展的拦路虎。将数据科学技能简单强加于人文专业学生的企图丢失了我们这一领域的独特属性，而我们的学生则可能被带入了关于数据和文化的更宽泛对话中。设想出一种完全不同的数据人文主义的最重要原因是，这样做又会进一步设想出一些学生，他们将会进入学术界之外的世界并对其作出改变，而不是简单挤进现有的框架中。我并不主张数据人文主义与数据科学相对立，而是一种补充和（偶尔的）修正。

## 实践

那么在实践中数据人文主义看起来是什么样的？我们如何培养学生的相关技能？本文的剩余部分将阐述我在美国东北大学英语系开展本科生和研究生数据分析教学的经验，主要是通过“数字时代的阅读和写作”[5]、“文本技术”[6]、“人文数据分析”[7]和“阅读机器：技术与书本”[8]等课程。在我为加拿大维多利亚大学 2019 年数字人文暑期讲习会的人文数据分析课[9]备课时，这些问题也常常萦绕我的心头。我从这些课堂教学中总结出的原则体现出了我自己想法的变化，如何创建一个数据丰富的人文教学方法以及它可能需要实现哪些目标，也就是说我会同时反思哪些失败了的方法和取得了成功的方法。我列出的“理想”并非是我认为我的教学已经达到的，但我尝试在每个学期都努力更靠近它们。在列出这些理想时，我也并没有暗指计算机科学、数据科学和数字人文或文化分析领域内没有人遵循这些实践原则。很多人已经在这样做，我的建议几乎全部都基于他们的榜样（我会尝试合理的引用）。

### 1. 从创造力开始

很多年来，我都从文本分析开始我的课堂编程教学单元：文本计数、关键词上下文、N 元语法（n-grams）、主题建模（topic modelling）。在花费数周时间（较大课堂的一个教学单元）或几乎整个学期（人文数据分析专业课堂）在这些教学活动上后，我们可能用最后一周的时间从事更具创造性的编程活动，例如创造一个推特诗歌机器人。然后，在近几年中，我已经认识到这恰恰是一种错误的方式为人文学生介绍编程或数据。去年我调换了教学活动的顺序，我认为我课堂上的所有学生都极大受益于此。

在我的“数字时代的阅读和写作”课中，我从 Giorgia Lupi 和 Stefanie Posavec 的项目“亲爱的数据”[10]开始我们的数据教学单元，这个项目始于一个网站，后来成为了一本很漂亮的书。Lupi 和 Posavec 用一年的时间记录每周他们日常生活的不同方面，创造出一种的新的方式将这些数据以图形呈现在明信片上，然后互相邮寄这些明信片。这本书汇编了他们的 104 次实验，外观设计非常漂亮。我们从这里开始是因为我希望我的学生们首先理解数据是比数字数据更大的分类，搭建在现有计算机程序中的系列可视化工具仅是数据可视化无限可能性中的一部分。然后学生的作业[11]就是花一周的时间记录然后可视化他们自己数据的某一方面，思考他们从数据中了解到什么、他们为呈现数据而做出的选择，以及他们所选择的呈现方式能够（不能够）让他们从数据中了解到什么。这样的练习帮助学生们从有效批判的视角做好准备，迎接计算数据的研究工作。

今天我也在两项课程——本科生 “文本技术”课和研究生 “阅读机器”课的编程教学单元中将之前总结环节的推特机器人作业用作课程的开始。总体来说，这一练习相当简单：学生选择一首诗或其他短文，然后编写一个程序，用恰当词性的词语代替诗中的相应部分：类似于儿童游戏 Mad Libs。得到的结果有时毫无意义，有时甚是滑稽，有时又有莫名其妙的深奥感。实际上，学生通过这一课程学会了几种重要的技能，包括如果操作 R 语言中的文本串，如果使用应用程序接口（API）查询网络服务，

以及如何将 R 程序的结果输出回网络服务，例如推特。但最为重要的是，他们通过一个生产性的和创造性的过程学会这些技能，同时也用到了他们现有的技能，例如理解和分析一首诗歌的能力：至少足以判断哪些替代词语可能会起作用，以及可能会是有趣的。尽管简单，这一练习符合学生们所熟知的剪辑诗歌和实验主义诗歌的悠久传统，帮助他们从一开始就将编程理解为比他们想象中更有创造性和表现力的事情。

我们介绍数据分析的方式很重要。如果你在 Courser 或 Codeacademy 这样的网站上调研现有的“学习编程”公开课，你会发现，无论何种语言的讲座，他们都几乎以同样的方式开始：教学生如何做数学。我并不是说数学不会用在人文数据分析中——尽管我会在下文进一步做探讨，而是我认为，考虑到在 2019 年的今天编程的应用范围如此巨大，我们不一定非要以这样的方式开始，它可能会立即成为横亘在人文学生面前的障碍，他们认为自己在数学方面没有足够的技能，或对此并不感兴趣。

目前大部分教授编程或数据分析的课程都假定这项工作必须扎根于计算机科学，但我赞同 Annette Vee 的尝试，将编程作为一种实践方法从计算机科学中剥离出来，成为一个学科，可以令二者均从中受益。在 2017 年她的《编程素养》（Coding Literacy）一书中，Vee 认为在当今世界的大部分地方，“计算机代码属于基础设施”，一种被重视和善加利用的素养：“位于技术写作的上层和下层，计算机代码如今构成了我们当代交流沟通的大部分”（3）。

> 当我们将编程当做一种书面沟通模式时，它就不再与计算机科学领域捆绑在一起。它不再仅仅扎根于数学、工程和科学，也同样包含写作沟通。从计算机科学中剥离编程不仅仅帮助我们理解作为一种沟通方式的编程，也将计算机科学解放出来，不再仅被看到其中的一种方法……无论是编程还是计算机科学，二者完全相互对应的想法都没有起到积极作用。（41）

Vee 在她的书中提出了“编程从不是技术专家专属的领域”，并倡导我们将编程视作一种可以以多种方式应用于许多领域的素养。这也是我的目标，是我向学生们展示处理数据的人文方法可以从一开始就有迥然不同的目的的主要原因。扩展开来一点就是：我们需要新的方法学资源向人文学生开展数据分析教学，从问题、数据和编程等能够引起学生共鸣的任务开始。

## 2. 使用特定领域的数据教学

我的下一个建议非常直截了当：当教授学生人文数据分析时，我们应当使用人文数据。这一点可能看似显而易见，但实际中并非总是这样，部分原因是它的实践难度可能比想象中更大。坚持用人文数据开展教学不可避免地引出了另一个补充的想法：我们不应当回避复杂且杂乱的数据。计算机科学和数据科学的教学通常使用专门设计的规整数据，目标明确地帮助学生熟练掌握特定的方法。我想到了 R 语言编程环境中内置的“mtcars”演示数据集。它是一个规模较小、结构有序的信息表格，汇总了各种车型的数据，包括每加仑油行驶英里数等细节。这个数据集来自于 1974 年的《Motor Trend》杂志，被设计为一项用以学习如何使用 R 语言中数学或绘图函数的资源，在教程演示过这些函数的使用方法后才在教程的其他部分被调用。其他常用的 R 语言内置数据集包括“iris”——包含鸢尾科 50 种花的各种测量尺寸，“ToothGrowth”——包含研究维生素 C 如何影响牙齿发育的实验结果，以及“USArrests”——包含美国暴力犯罪的统计数据。

在人文学生的教学中默认使用这些数据集存在一些问题。首先，我们的学生常常发现很难从车辆或牙齿发育的数据外推到他们感兴趣的领域。对数据分析心存犹豫或畏惧的学生尤为如此，因为很多

现有的股票练习数据集从技术上更响应企业或政府的需求。说的更直白一些，这些数据集在我们的学生看来和感觉上更像是数据应该的样子，也就是说，它们看起来和感觉上是距离人文问题非常遥远的一些领域内的信息，它们让学生们更加疏离，而不是在欢迎他们的加入。

其次，这类数据集本质上区别于人文数据，也就是从文学、历史和相关领域范围内获得的数据，这类数据的特点（至少在一定程度上）是其复杂性和杂乱性。事实上，很多计算科学教授都将我们的数据用于他们自己的研究中：它有着令人着迷的难度。几年前，我在美国东北大学与其他教师共同开设了一门“绘制波士顿”（Bostonography）的计算机科学本科生课程。参加课程的学生们不断表示他们在处理这些所谓“真实数据”（即与波士顿市区相关的历史数据）时如何地感到挫败但又很兴奋，与他们在计算机科学课堂上曾使用的“罐装”数据集完全不同。我听到学生们说，我们课堂上的数据极具挑战性，但比他们曾一次又一次使用的枯燥数据更有成就感。人文学生需要从一开始就处理这些极具挑战性的人文数据。

不巧的是，如何使用人文数据教学并非不言自明的，因此在开设这类课程的教师间，最常见的怨言就是在文学研究、历史或其他相关领域内缺少规模较小、可用于教学的数据集。我们在这些领域的文献资料中见到的很多数据集往往太大，无法让一组学生同时在课堂环境中使用，事实上在课堂外也常常令人受挫，学生每次都需要几分钟甚至更长的时间才能知道一行代码是否按预期运行。学习编程所必须的迭代过程——运行代码、查看结果是否正常、修改、再次运行，无法在大数据中进行，这也就是为什么计算机科学和数据科学领域开发了我之前所描述的教学数据集。随着越来越多的人士尝试在学生中培养数据人文主义，我们也需要构建易管理的但仍适度凌乱的数据集，用于课堂教学。

我个人的研究很大程度上围绕历史报纸，我已经基于美国国会图书馆的“美国报纸名录”（U.S. Newspaper Directory）开发出了几种极为有效的教学练习。[12] 这个数据集并非他们的数字化报纸典藏，这些典藏过于庞大，无法用于课堂作业，而是他们对 1690 年至今所有已知在美国创办的报纸的元数据标引。同 R 语言中的“汽车”数据集一样，可以在其中开展多种自然实验，学习如何比较不同类别或绘图：19 世纪 70 年代创办了多少周刊和日报？在世纪之交哪个州的报纸最多？我们是否可以分析或可视化过去数十年间报纸标题中最常见的词语，关于编辑和读者对报纸概念的理解如何随时间变化，这种分析能告诉我们些什么？并非我所有的学生都从事报纸研究，但因为这一数据集更贴近他们的兴趣，他们知道我们可能会在其中探讨什么样的问题，因此这样的研究过程更容易转移到他们自己的领域中。

### 3. 语料库优先于研究方法

关于我们在课堂中所使用数据的问题直接引出了我的下一个建议，也就是在人文数据分析中，语料库或数据集的教学应当优先于研究方法的教学。在她 2018 年的《虚构的世界》（A World of Fiction）一书中，Katerine Bode 将她在图书史方面的专长用于批评许多计算文学学者处理他们的核心分析数据的方式。Bode 特别提出：

> 在他们的文学史著作中，[Franco] Moretti 和 [Matthew L.] Jockers 都将文学数据和数字化典藏呈现为预批判性的、固定的且不言自明的。在将数据和计算设想为访问文学 - 历史记录的直接且综合渠道时，他们否定了构成数据和数字记录并让它们能够用于分析的批判性和解释性活动。（20）

Bode 并不是说遥读学者未能揭示他们从中得出结论的语料库，而是说他们常常没有充分的描述这

些语料库：它们包含什么，不包含什么。因此，Bode 认为遥读看起来很像新的批判性精读，如果我们手中现有的文本资料——无论是济慈的一首诗还是 HathiTrust 书库，被低估或忽视了“文本”或“档案”的理想化概念价值。

Bode 的观点是对 Johanna Drucker 等多位学者见解的扩展。Johann Drucker 认为“数据”一词恰恰意味着它所标识的信息是“一种能够被记录和观察的‘所在’”或仅仅是“已存在事实的一种自然表征”（Drucker 2011）。尽管我不太支持 Drucker 的建议，用“capta”一词来代替“data”（数据）一词，因为“capta 含有主动获取的意思”，但我的确同意数据人文主义必须从探讨数据的结构开始，而不是从数据分析开始。为了对美国国会图书馆的“记录美国”（Chronicling America）报纸档案做出分析——正如我们在“疯狂传播的文字”（Viral Texts）项目中所做的，[13] 我们需要理解收录其中的报纸是如何通过美国国家数字报纸项目（National Digital Newspaper Program）编撰的，该项目为各个参与项目的州设置了特定的选择标准。这些标准包括具有代表性的思想，发行量和影响力最终决定了整个报纸档案中包含哪些类型的报纸、排除哪些类型的报纸。

回到对美国国会图书馆“报纸名录”（Newspaper Directory）的课堂讨论，我会注意到，关于“我们从对它的分析中学到了什么”的讨论很自然地引发了关于数据集本身结构的丰富讨论。谁编撰了这些清单？他们如何获得这些报纸的相关信息？哪些信息没有包含在内？例如，数据包含每种报纸目标读者的信息吗？如果我们分析标题中最常用的词语，我们需要将特定年代内的哪些报纸版面排除在外？人文方法的数据分析必须将这些讨论置于首位，并指导学生如何深入探究他们所遇到的每一个数据集或语料库蕴含的思想、社会和政治基础。

## 4. 将思维放在方法的前面

我在本文中引用的课程的大部分教学内容是通过利用 R 语言编程分析人文数据。我不会说编程是有效开展人文数据分析教学和实践的唯一道路，很大程度上是因为这大大提高了入门的门槛。这是一个需要更多而非更少声音的研究领域。现在已经有很多出色的工具可以用来完成最常见的分析任务，研究者可以在图形用户界面（GUI）中思考所有我以上列出的问题，与使用编程语言一样便利。坦白说，我自己在这一领域的工作就是从使用 GUI 和应用程序开始的，这帮助我建立了回答问题的方法。虽然这些问题最终需要我具备编程能力，但这些早期工作的意义不逊于我现在从事的工作。

然而，我将会阐述我在课堂中越来越倾向于采用编程而非常见数字人文应用程序来开展人文数据分析教学的原因。事实上，我应当立即修正这一说法为，我的人文数据分析教学使用的是练习簿，是应用程序和纯编程方法之间的一种过渡形式。在 R 语言中，这样的编程练习被编辑成 RMD 文件，[14] 它们与包含可执行代码块的描述性文字交织在一起。这样我就可以在课堂上使用较为复杂的大代码块，甚至在我们第一天开始讲解编程的时候。当我们立即转向应用程序时，学生们会更加了解前因后果，比仅仅打印出“你好，世界”的效果更好。在面对这样一个 RMD 文件时，学生可能并不理解每一行文字的含义。但与图形用户界面相比，我们可以检视每一行代码，一起从事大量的工作，最初可能包括运行代码块、检查结果，然后重新修改代码再次运行，循序渐进地理解代码到底能做些什么。

我遵循这一流程的原因有几个。首先，我非常同意我的同事、历史学家 Ben Schmidt（2016）所说“数字人文学者根本不需要理解算法”——如果理解意味着准确掌握算法等式中每一个希腊字母代表的含义，或能够准确重现其中的数学方法。Schmidt 还说，人文学者需要的是“理解算法尝试带来的转变”。如果学生要认真负责地使用计算技术，他们需要理解，比如将文本或表格数据从一组小说移动到一个

矢量空间模型中时会发生什么。这种方法的假设条件是什么？哪些变量可控，当它们发生变化时会怎样？在图形用户界面中能够做同样的操作吗？“Gale 的数字学术实验室”就相当的强大，对它所收录的分析方法以及研究者能够在其中控制的变量都提供了详细的解释。

我并不想要假装使用像 R 这样的编程语言让其他所有一切都苍白无力，因为实际上某些特定算法的许多组件都打包在功能模块中，不需要有太多的理解也能够运行。但对于我而言，讲解代码能够帮助学生更全面地理解转变发生的过程，看到某一过程中的各个主要步骤，开始询问某些功能如何操作的问题。我的练习簿通常包括多个预先写好的代码块、对学生能够操作的代码行的讲解，以及空白的代码块，他们可以用来复制、粘贴和修改我提供给他们的代码。这种修修补补的方式让学生感受到代码是一种媒介——它实际上是可以复制、粘贴和编辑的文本，并且为后续的练习簿做好准备。在更高阶的练习簿中，他们将更直接地参与编程，在独立的研究项目中对他们自己的数据提出问题。

对于我所描述的所有教学方法，有一个需指出的重点：我无法在四周、甚至是一整个学期的课程中将学生转变为熟练、独立的程序员，而且这也并非我的目标。我的目标也不是让他们变得专长于某种方法：例如分类或主题建模。我尝试培养的是一种思维，如何处理数据、探索数据，发现计算技术可能帮助回答哪些关于数据的问题，然后决定什么方法可以用于帮助回答这些问题。最终我希望他们开始理解如何制定一系列的步骤——某种意义上的一种算法，但并不以数学的形式表达，这些步骤将帮助他们将数据从原始的形态转变为与他们想要探讨的问题相关的一种形式。为了让这一过程切实地发生，我们可能需要更多的研究，甚至可能需要展开协作。但我认为人文学科的学生最难掌握的技能是程序式的思考，而非编程。

以人文为本、探究式的数据分析方法不仅仅令学生们能够对人文数据使用专为计算机科学或社会科学而开发的方法，也能够激发对能够分析人文数据的新方法的需求。我认为这是数字人文或相关领域内仍未被大规模探索的领域。我们已经看到一些相当复杂的项目，将诸如矢量空间分析或主题建模这样的方法应用于小说和历史报纸，但我们还没有看到从历史、文学及相关领域数据中衍生而来且专用于这些领域的新方法。但这是我想要开展教学的方向。我们需要人文数据分析的一个原因是我们不能将数字文化遗产的维护和算法架构彻底让权于工程师，或完完全全交给公司企业。像“Gale 数字学术实验室”（DSL）[15] 这样的产品无疑是一个良好的起点，在他们的历史典藏中开始研究，帮助学生理解他们整合的分析方法所能够实现的效能。Gale 的实验室平台非常强大且灵活，我对它的各项功能印象深刻。

无论这种说法是否是真的：没有一套标准的分析工具——尽管经过了深思熟虑的开发，能够将所有我们想要从历史、文学或社会科学数据集中了解到的东西一一捕获；人文研究的最大挑战之一仍然在于待探讨的问题很难普遍化或提前预测，因此很难开发一种通用的工具来涵盖它们。仅仅教授学生使用现成的工具会限制他们的各种可能性。此外，我的教学经验告诉我大部分的学生未来往往不会在资源良好的大学中工作，更遑论这些大学又恰好订阅了能让 DSL 这样的产品大显身手的数据库。他们中的大部分人未来会在美国大学和学院的更广阔领域内工作，或者进入企业工作。如果这些学生将会继续从事处理大量数据的人文研究工作，或者会训练他们自己的学生从事同样的工作，他们将主要利用开源工具（包括图形用户界面和编程语言）和开放获取数据。我的学生需要理解这类工具的基本内在原理，能够灵活地适应不同院校的环境，并且在必要的情况下，了解如何开始组建自己的一套工具，开展他们所需的研究和教学。

## 注释：

1. https://safiyaunoble.com/.
2. https://mathbabe.org/.
3. https://virginia-eubanks.com/.
4. https://bookbook.pubpub.org/data-feminism.
5. https://f18rwda.ryancordell.org/.
6. https://s19tot.ryancordell.org/.
7. http://s17hda.ryancordell.org/.
8. https://s19rm.ryancordell.org/.
9. http://www.dhsi.org/courses.php#DataAnalysis.
10. http://www.dear-data.com/theproject.
11. https://f18rwda.ryancordell.org/assignments/dear-mydata.html.
12. https://chroniclingamerica.loc.gov/search/titles/.
13. https://viraltexts.org/.
14. Examples of RMD files from my recent DHSI class are here: https://github.com/rccordell/DHSI-HDA/tree/master/exercises.
15. https://www.gale.com/intl/primary-sources/digitalscholar-lab.

## 参考文献：


Bode, Katerine. 2018. A World of Fiction: Digital Collections and the Future of Literary History. Ann Arbor: University of Michigan Press.

Daston, Lorraine, and Peter Galison. 2017. Objectivity. New York: Zone Books.

Drucker, Johanna. 2011. "Humanities Approaches to Graphical Display." Digital Humanities Quarterly 5, no. 1. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html.

Fagan, Benjamin. 2016. "Chronicling White America." American Periodicals: A Journal of History & Criticism 26, no. 1: 10–13. https://muse.jhu.edu/article/613375/summary.

Klein, Lauren. 2015. "The Carework and Codework of the Digital Humanities." http://lklein.com/2015/06/the-carework-and-codework-of-the-digital-humanities/.

Library of Congress. Chronicling America (newspaper archive). https://chroniclingamerica.loc.gov/.

Library of Congress. U.S. Newspaper Directory. https://chroniclingamerica.loc.gov/search/titles/.

Lupi, Giorgia, and Stefanie Posavec. 2016. Dear Data: A Friendship in 52 Weeks of Postcards. New York: Princeton Architectural Press.

Ramsay, Stephen. 2011a. Reading Machines: Toward an Algorithmic Criticism. Urbana: University of Illinois Press.

Ramsay, Stephen. 2011b. "Who's In and Who's Out." Position paper delivered at the 2011 Modern Language Association (MLA) Conference. https://web.archive.org/web/20170426170232/http://stephenramsay.us/text/2011/01/08/whos-in-and-whos-out/.

Schmidt, Benjamin M. 2016. "Do Digital Humanists Need to Understand Algorithms?" In Debates in the Digital Humanities, edited by Lauren F. Klein and Matthew K. Gold. Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/99.

Vee, Annette. 2017. Coding Literacy: How Computer Programming Is Changing Writing. Cambridge, MA: The MIT Press.

Witmore, Michael. 2012. "Text: A Massively Addressable Object" [2010]. In Debates in the Digital Humanities, edited by Matthew K. Gold. Minneapolis: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/28.

# 将数字人文引入本科生课堂：策略、方案及利用 Gale 数字学术实验室的教学实践

**Sarah L. Ketchley**，美国华盛顿大学数字人文及埃及学讲师和教授
Ketchley@uw.edu

**摘要：**本文探讨了以人文和社会科学教育背景为主的本科生数字人文教学中的一些实践问题。结果呈现在一项名为“2015-2018 年华盛顿大学的数字人文概论课”的案例研究中。尽管选修这门课程的很多学生之前在这一领域很少或几乎没有经验，但大部分人选择这门课程的目的是想要熟练掌握技术并提升自己毕业后的职场竞争力。这一案例研究探讨了在这门课前三次开课时课堂教学中的经验教训，描述了在第四次开课时课堂教学中采用的策略，如何解决教学和课堂管理中的很多挑战。这些策略包括使用基于云存储的数字工具——“Gale 数字学术实验室”，创建和管理与课堂研究主题相关的一次文献资料的语料库。学生们学习在实验室平台上清理 OCR，然后使用整合其中的数字工具进行定量和定性文本分析。这些分析包括主题建模、N 元语法、命名实体识别和情感分析等。学生们将他们的分析结果导出为原始数据（CSV/JSON），进一步使用 Google Fusion Tables 和 Voyant 等外部工具进行探讨；或导出为图像文件，用在他们最终的 Omeka 展示中。

**关键词：**数字教学法、本科生、数字工具、文本挖掘、数字人文、基于云的数字人文工具

数字人文领域的蓬勃发展为学生们提供了在人文内容的环境下培养技术技能的机会。这一领域因其核心价值观而逐渐被人们所知，也就是开放、协作、多元化、实验、联合及互联，这些精神延伸到了课堂环境和教学实践中（Spiro 2012）。数字人文并非人文学科的分支学科，它也不归属于任何院系。它的范围和潜力仅由参与其中的研究者的想象力而决定。数字人文的本质是协作——很多工作是基于研究项目的，最终的目标是让内容或研究以数字化的（通常是在线的）形式呈现。这些内容可以是一套一次和 / 或二次文献、一个数据库，或可能是地图位置。计划和开发一个成功的数字项目需要有效的项目管理和团队合作，考虑项目保存和可持续性的最佳方式，以及建立和执行多种工作任务，例如文本建模，处理图像、地图和统计数据，同时还要考虑在线演示和设计的美观性。这是试验性教育和技术训练与传统人文学科的结合。数字人文训练和实际操作会吸引很多的学生，因为学习到的这些技

能能够随时应用到各个行业中。

如何将这些我们作为研究者使用的数字工具和实践技能以最佳的方式带入本科生课堂？本文讨论了将数字人文引入课堂的策略，而课堂中的学生们则是鱼龙混杂，来自大学的各个院系，各自具备的技能有所不同。这一经历是从最早利用数字工具探讨历史研究数据的人文学者的教学视角描述的，而并非以技术为核心的学科的教学视角。

## 案例研究：数字人文概论课

“数字人文概论”是华盛顿大学 2015 年和 2018 年面向四组不同学生开设的一门课程。这门 3-5 个学分的课程没有选课条件要求，每周两次课，每节课 2 小时。

前三次开课时课程内容本质上非常近似，不同院系的本科生和研究生共处一室，按院系分成几组，研究生在各自所在的小组内担任项目小组长。教学大纲（见图 1）范围宽泛，引导学生在某一特定研究问题的基础上完成一项数字展览的计划与搭建过程。在前两次开课时，讲师根据她熟悉的数据选择了研究问题：中东游记，但她对数字人文教学还不太熟悉。到第三次开课时，学生们可以从几个选项中自行选择数据集。课堂的教学进程重现了一个成功的数字人文项目的进程，从建立团队工作规范和计划开始，找到收集和管理数据的最佳方法并付诸行动，分析和可视化收集到的资料，最后搭建数字展览以展示每个学生小组的工作成果。

课堂的学习目标包括：

1. 培养人文和社会科学专业学生的核心计算能力；
2. 推动跨学科学习和协作
3. 利用数字工具探讨人文数据
4. 呈现数字研究项目的完整过程，包括计划、执行，然后是总结，强调可持续性的构建；
5. 让学生们接触到数字人文范畴内的多种研究项目、数字工具和研究方法。

无疑，前三次开课时出现了很多的教学和实践难题（不仅仅因为课程在一定程度上是实验性质的），每次当然也在进步和改善。到 2018 年秋季第四次开课时，以下所述困难的各项解决方案都应用于课堂，讲师在课堂上的控制力更强，而学生在课堂上的体验也更加顺畅。

### 困难 1：教室

- 课程最初的推广较为艰辛，它被列在近东语言和文明（NELC）课程目录中，但学生们通常不会到 NELC 院系寻找一门数字人文课。为了吸引更多的学生选课，讲师在校园中做了大量的推广工作，包括制作海报、向各个学生团体和班级推荐这一课程，在多个社交媒体渠道中发布消息。

- 每个人能够实际操作的时间有限，华盛顿大学采用的是学季制，每学季仅有 10 周。数字人文课程每周上两次课，每次上课时间约 2 小时。

- 学生们的技能和知识背景有很大差异，因此想要让这门课成为对所有人而言都是有意义的学习经历就会存在一些问题。

| 2018 秋季学季教学大纲 -INFO 498C：<br>数字人文概论 | |
|---|---|
| **题目** | **主题** |
| **计划** | 欢迎！教学大纲综述 |
| | 什么是数字研究项目？评价数字研究项目 |
| | 版权和开放资源文献资料 |
| | 期末展示计划和 Omeka 概述 |
| | 数字人文实验室：简介与演示 |
| | 实验：项目计划和协同工作 |
| | 风险评估和数据管理方案 |
| **收集和管理** | 元数据 |
| | 创建数字档案，什么是 OCR？ |
| | 语料库创建和管理，利用 OpenRefine 清理数据 |
| | 利用 RegEx、Lexos 和数字学术实验室清理文本 |
| | 图书馆和数字学术，在 Hathi Trust 中清理文本 |
| **数据处理** | 实验：清理内容集 |
| | 利用 Voyant 和数字学术实验室进行文本和语料库分析 |
| | XML 和 TEI |
| | 利用 Clavin 进行地理剖析，绘制地图 |
| | 实验环节 |
| | 命名实体识别、情感分析和 N 元语法 |
| **数字展示** | 实验环节 |
| | 期末小组展示和数字展览 |

*图 1：数字人文概论课教学大纲*

- 课堂本身的空间布置是一项重要的考虑因素。传统的大学教室通常是按讲座式的 / 被动学习的方式布置的。而数字人文教学的性质则令教学工作无法在这样的环境中获得主动学习的效果。
- 讲师首先是一位学科研究专家，其次才是数字人文学者。常常出现的情况是，人文学科的教师很不情愿向他们的学生讲解实践性环节，特别是因为相比于较为传统的人文主题，向一组学生传授技术知识可能让学科研究专家面临挑战。在这种情况下，本课程的讲师花费了八年的时间自学技术技能，开拓利用数字人文方法探讨自己研究项目数据的道路。在这个过程中，她与持同样想法的同事的协同合作，最终建立了一个数字人文本科生实习项目（位于 Newbook Digital Texts 出版社）以及后续的补助资金，以支持前三次数字人文概论课程的发展与教学工作。这样漫长的学习期对于大量时间需用于教学和研究的教师而言是不可行或不实际的。

## 课堂解决方案

- 信息学系被证明是数字人文概论课程最合乎逻辑的开设地点。这门课可供主修和辅修，吸引了不同院系的学生，在没有推广或宣传的情况下，课程在开放注册的第一天就满员了。

- 为了尽可能增加学生与教师和助教一同动手操作的时间，本课程讲师选择从花费时间做综述性的讲解转变为翻转课堂。（Stommel 2013）学生们需在上课前进行预习，包括观看为每一堂课创建的概述视频，并辅以课程读物以及参加在线讨论。基于已经完成的课外作业，课堂教学从问答环节开始，学生们来上课时已经准备好将他们学到的一些理论付诸实践。

- 学生完成课前调查，其中的问题涉及之前的经验、兴趣、技能和目标等，以便从第一天开始就将他们放在相匹配的小组中。他们能够立即相互间交流和学习，同时培养起宝贵的社交技能。课前调查也为讲师提供指引，每一项课堂主题的讨论需要达到哪种详细程度，有哪些机会推动学生间的学习和参与。学生们在课程进行到一半时接受第二次调查，发现任何悬而未决的问题或关注点，并征集需讨论的话题。在每堂课上与学生一对一互动的机会异常宝贵，帮助讲师为学生创造适合自身的学习体验。

- 华盛顿大学有很多的主动学习教室（ALC），是这类以小组实操为特征的数字人文教学的理想之所，在课堂活动中强调讨论、一对一互动、及时反馈和解决问题（Hornby 2017）。一些实用的特点包括家具易于移动，围成多个小组工作台，同时也易于将学生合并成较大的组。每一个工作台的桌子上都有一台较大的显示器，便于协同工作，还有可书写的墙面，便于计划和讨论活动；最为重要的是，多个位置合适的电源插座。课堂设置能够促进教师、助教和学生们之间的常规互动，反之又能快速反馈和解决技术问题并回答学生提出的问题。学生们发现这种非常规的设置既有趣又实用，因为它不仅有助于他们处理手边的资料，也有助于他们更大程度地参与学习过程，他们有方法能够掌控自己的学习体验。小组式的课堂设置也能够促进协作，进一步培养社交技能。

- 认识到她自己并非课堂上论及的每一个话题的专家之后，讲师将这个概论课程打造成了与图书馆和大学其他院系同事协作的机会，目的是为学生提供一个综合全面、内容丰富的教学大纲。例如，图书馆和信息科学项目的一位讲师讲解版权以及如何利用开源资料这样的主题，而元数据图书馆员则探讨“都柏林核心标准”以及清晰和连续的研究记录对项目可持续性的重要意义。在学生准备开始搭建他们的数字展览时，由来自人本设计与工程专业的教师谈论有效的项目计划和设计。

在 2018 年秋季的课程中，讲师与商业出版商 Gale 合作，首次在课堂上使用了最新发布的“Gale 数字学术实验室”（Gale Digital Scholar Lab）。

## 困难 2：技术

前三次开课时，教学内容对许多数字工具和平台进行了试验。回想起来，这是一项有些雄心勃勃的目标，但却往往制造出一些问题，而非解决问题。其中的一些困难描述如下：

- 学生上课时带来了各种型号的平板电脑和笔记本电脑，出现了多种操作系统，甚至是系统语言。由于课堂活动主要依赖软件，学生们或者下载到本地，或者在线使用，至少有一整堂课用于下载和安装，而在这一季度的课程中，排查故障的工作始终在继续。在这样的环境下，很难实现

生动有趣的教学体验。

- 学生们在前三节课中开展的多种分析工作决定了他们需要在这一个学季的课程中使用多个平台。每种平台都有自己独特的模式，每一种都有一个学习曲线（有时曲线走势相当陡峭）。通常出现的情况是，工作流程既不清晰也不直观，需要耗费大量的时间学习如何访问和使用这些平台。
- 在更大的范围内，技术在课堂教学中的使用因机构设置而受阻，特别是 IT 和软件支持方面，以及缺乏专为数字学术提供的支持。

## 技术解决方案

与 Gale 合作，或使用“Gale 数字学术实验室”，让学生能够在一个平台中进行大部分的工作，构建语料库、数据管理、分析和可视化。因为“Gale 数字学术实验室”是基于云存储的，学生带到课堂上的电脑类型（甚至不带电脑）不再是一个问题。类似的，通过限制学生们安装和使用的辅助工具的数量，课堂管理方面的问题大大减轻，教学的焦点转为建立研究问题并最终以数字展览呈现研究结果。

## 2018 年秋季的数字人文概论课

本文的以下部分将围绕 2018 年秋季课程的细节，探讨教学大纲以及学生们在课堂上使用的数字工具。2018 年课程面向的学生群比前几次开课时都更宽广，因为学生们通常都很热衷于信息学院的课程（见图 2）。几个小时之内课程就注册满员了，共有 35 位学生注册。学生们来自大学的 21 个院系，大部分是大三和大四学生，90% 的学生之前没有使用数字工具探讨人文数据的经验——大部分人都感到困惑为什么数字人文会被描述为一个学科。他们选择这门课程或者是为了培养数字素养，或者为了将现有的技术或某些学科的专用工具应用于人文数据集。

笼统的课程学习目标上文已经列出，为了实现这些目标，讲师努力营造出一种环境和课堂文化，鼓励自学和学生间的学习以及知识共享、参与和协作（参见 Savonic and Tagliaferri 2017）。讲师通过邀请客座演讲人分享他们的专业知识来塑造这一学习理念的典范，强调跨学科协作的益处。翻转课堂不再强调传统课堂模式中教授的权威地位，而是将课堂的焦点转向与每位学生建立关系，同时留意小组的整体动向。最初的几节课着重于讨论和确立小组内专业角色划分的策略，包括建立和保持小组合作方式下的工作预期值，按时完成各自的任务，报告流程和最终的结果。因为每位学生在他们各自的小组中都被分配了一个角色（例如，项目经理、内容专家、数据管理人、可视化专家等），重要的一点是要营造一个环境，其中每个角色的都有着相同的重要性，正如实际中的教师和技术专家的角色。

为了鼓励探寻和求知的精神，而且因为学生们探讨的数据集涉及的主题对他们而言是全新的，所以教学目标之一就是鼓励学生们大胆试验，不怕失败。因此，一项基本的任务就是每周的工作日志，记录课堂上的笔记、阅读材料和讨论，以及所使用工具和所运行分析的完整操作步骤。假如学生们保存了他们所有这些研究工作的完整记录，当他们运行的分析返回无意义的结果，或试验的工具最终没有达到预期时，他们不会遭受严重的损失。

课程的研究主题是历史菜单，选择这一主题的原因是它较为宽泛且能够让课堂上的所有学生参与其中，范围足够每个学生小组建立自己独有的研究问题。这个主题的灵感来自于纽约公共图书馆的“菜单上有什么”（What's on the Menu）项目，学生们主要使用来自 Gale 原始档案（Gale Primary

2018 秋季学季学生情况

| 百分比 | 大学院系 |
|---|---|
| 24% | 地理 |
| 15% | 经济学 |
| 12% | 情报学 |
| 6% | 图书馆和信息科学 |
| 6% | 英语 |
| 6% | 数学 |
| 3% | 哲学 |
| 3% | 环境科学和资源管理 |
| 3% | 商业管理 |
| 3% | 传播学 |
| 3% | 历史 |
| 3% | 艺术 |
| 3% | 法语 |
| 3% | 商业管理（金融） |
| 3% | 地球和空间科学 |
| 3% | 近东和中东研究 |
| 3% | 政治学 |
| 3% | 建筑学预科 |
| 3% | 材料科学与工程 |
| 3% | 心理学 |
| 3% | 社会学 |

*图 2：2018 秋季学季学生情况*

Sources）中的内容，以及来自纽约公共图书馆项目的一些额外文献资料和其他开放资源。学生和教师通常情况下可能都不太了解他们的图书馆网站能够提供多少及哪些数字典藏。使用“Gale 数字学术实验室”的一个重要好处是它提高了对院校现有数据库馆藏的了解和使用，其中就包括随时能够支持研究、教学和数字学术的 Gale 原始档案。

在一个季度的课程中，学生们学会了以最佳的方式计划他们的研究项目、创建和管理数据集，以及使用数字工具分析他们收集的资料，然后搭建一个数字展览展示他们的工作成果。我们在展示阶段使用了 Omeka，而在数据集创建、管理和分析的协作过程中使用了“Gale 数字学术实验室”。

我们在课程中遵循的工作流程总结在“Gale 数字学术实验室”的主页上（见图 3）。学生们从创建文献语料库（在“Gale 数字学术实验室”中描述为“内容集”）开始，经过数据管理步骤，然后清理光学字符识别（OCR）数据，最后对他们的内容集进行各种文本挖掘分析，生成分析结果的可视化图表。讲师录制的教学视频强调了工作流程和步骤的重要性，并指引学生使用“Gale 数字学术实验室”每个

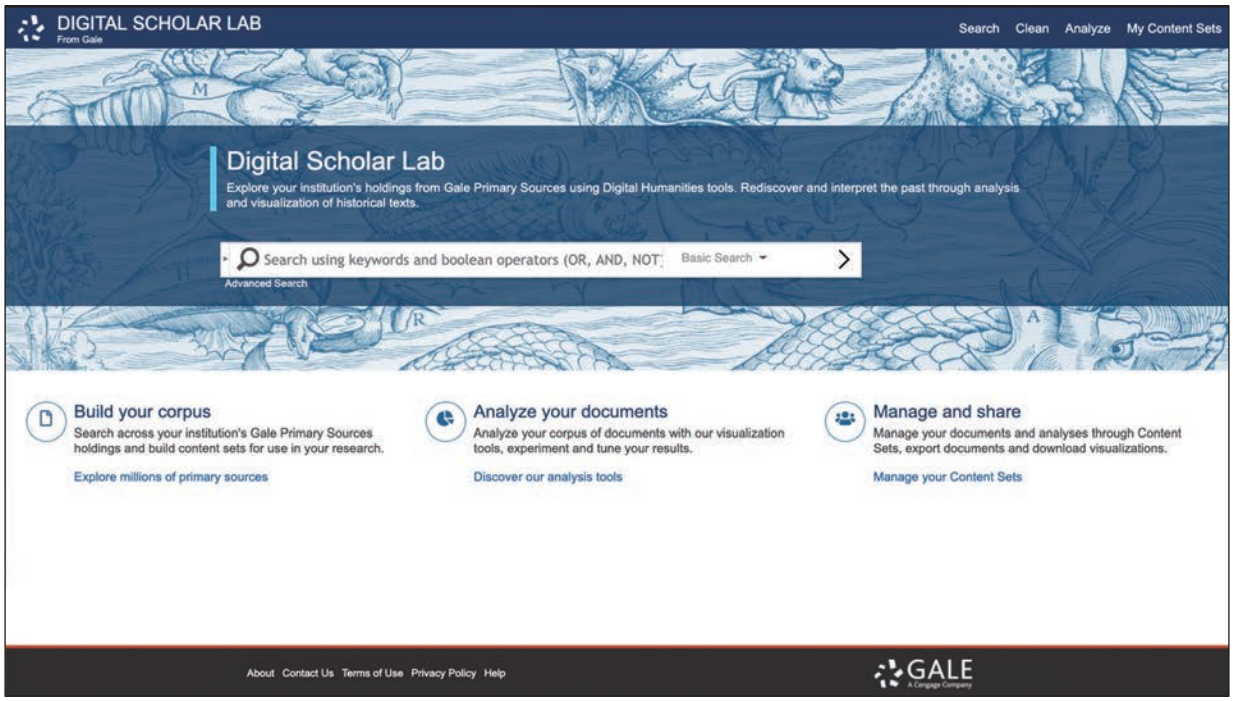



*图 3. “Gale 数字学术实验室”主页截图*

页面上的背景知识帮助文档，其中包含的数字人文术语表尤为重要，因为词汇往往是理解一个新领域的最大障碍。

每个学生小组都有一个 Gmail 登录账户，便于对相同的文献资料开展协作，而不会干扰课堂中其他小组的工作。当学生们登录后，他们会访问到一个个性化的工作平台，存储了他们创建的内容集以及他们运行过的分析项目，以及每个内容集中文档的描述性元数据（见图 4）。学生们也能够导出各种信息，包括每篇一次文献的扫描图像、OCR 文本、可视化图表和原始表格数据，他们可以将这些数据用于外部工具，在此次概论课中则是一种数字展览工具。

## 1. 创建和管理内容集

学生们研究工作的起点是创建与小组课题相关的内容集。用“menus”（菜单）一词做宽泛的检索会返回数千条结果，因此学生们使用了高级检索筛选选项，将检索结果加入到易于管理的内容集中。各个小组讨论各自感兴趣的、与菜单相关的主题，然后反复多次重复检索和数据管理步骤，琢磨并建立他们的研究问题，并找到足够的资料用于分析。

当学生们确定哪些文档与他们的研究最为相关时，他们使用元数据作为指引，并利用了检索结果页面上每篇文档的 OCR 文本片段。内容集可以包含最多 10,000 篇文档，对于课堂教学而言完全足够。

即便处理得当，数据管理也需要较长的时间，如果要得到有意义的分析结果，这是至关重要的组成部分。学生们能够进入文档浏览页面，仔细地并排查看文献原始扫描图像和 OCR 文本，决定这篇文

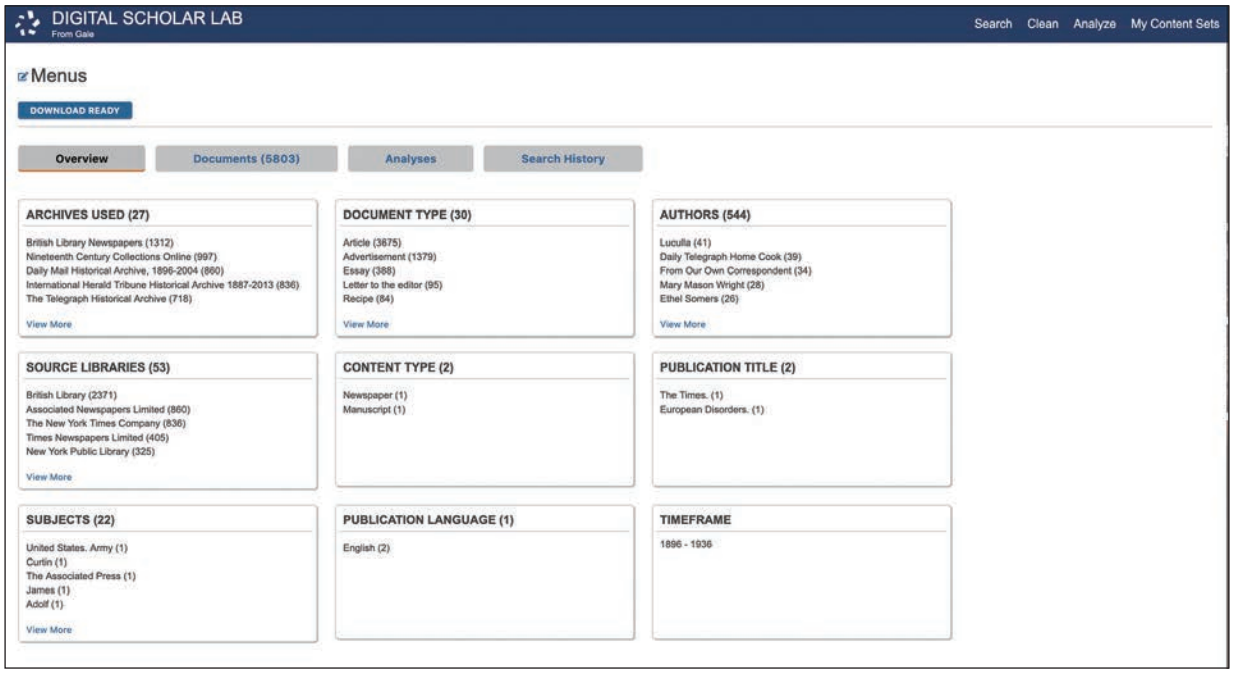



*图 4. “Gale 数字学术实验室”中“我的内容集”页面截图*

献是否适合包括在内（见图 5）。

在进行这一步骤时，他们利用了每篇文档的 OCR 置信度水平作为一项数据管理指标，低置信度水平的文献不一定会被排除在外，因为它可能有大篇幅的图片或从质量很差的原稿或微缩胶片扫描而来，或者 OCR 扫描软件的版本较低。

这个数据管理的迭代过程最终让七个学生小组都建立了自己的研究问题并在“Gale 数字学术实验室”中创建了一组数字研究对象。他们选择的主题千差万别，从“第一次世界大战前的素食主义意味着什么？”、“美国人对中国食物的态度变化”和“繁荣与萧条时期的饮食规律”，到“19 世纪以来感恩节菜单的演变”、“著名的纽约客酒店露台餐厅的历史”、“火腿蛋松饼作为早餐的历史”以及“杯盏之间”——一个将菜单上的葡萄酒与它们的历史背景相关联的项目。

## 2. 为分析做准备

很重要但常常被忽视的一个文本挖掘和分析步骤是 OCR 文本的清理。在课程刚刚开始的时候，“Gale 数字学术实验室”平台还不具备文本清理的功能，在 2018 年 11 月才新增了这项功能。因此，至少在最初的时候，学生们必须导出他们的 OCR 文本，在“Gale 数字学术实验室”以外利用 Lexos 和 OpenRefine 清理文本。一些学生经历了我们所熟知的下载和安装难题，特别是对 OpenRefine，学习曲线异常的陡峭。清理功能加入到“Gale 数字学术实验室”中之后，学生们能够更流畅、更直观地将他们的内容集随时转入清理步骤。

这类工作的一个重要起点是对数据集的熟悉程度，这将有助于研究者思考并选择如何清理和清理

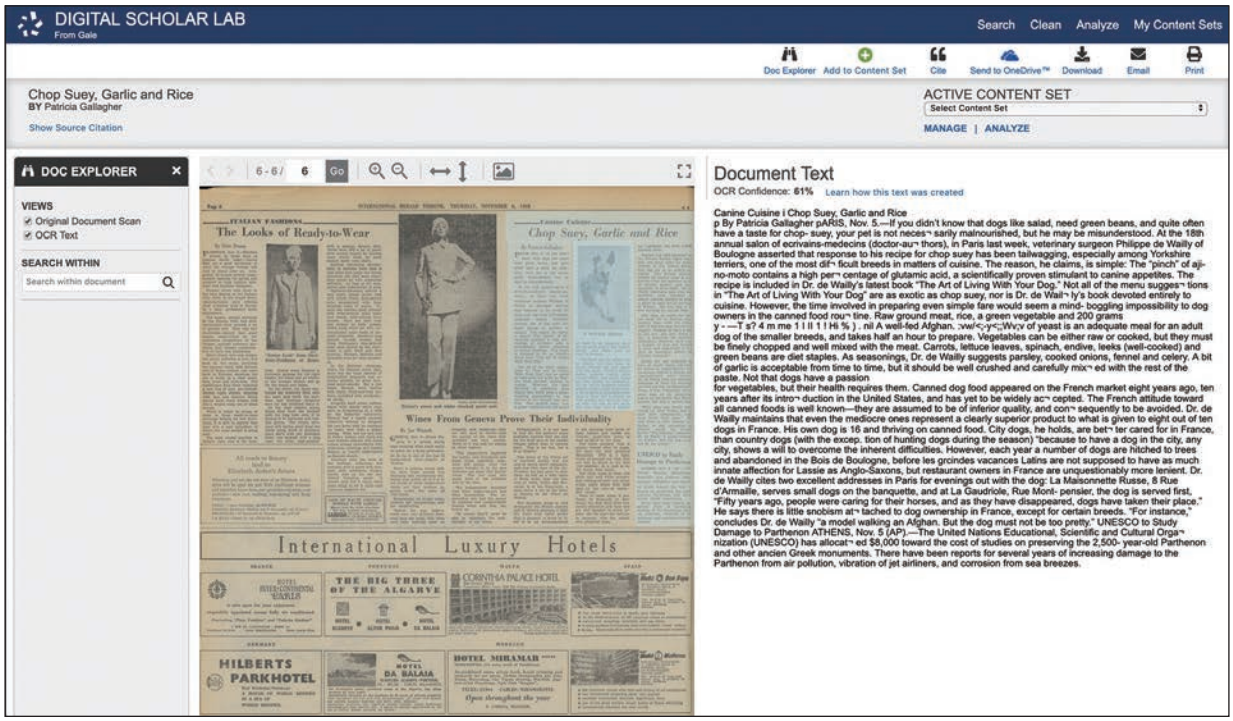



*图 5. “Gale 数字学术实验室”文档浏览页面截图*

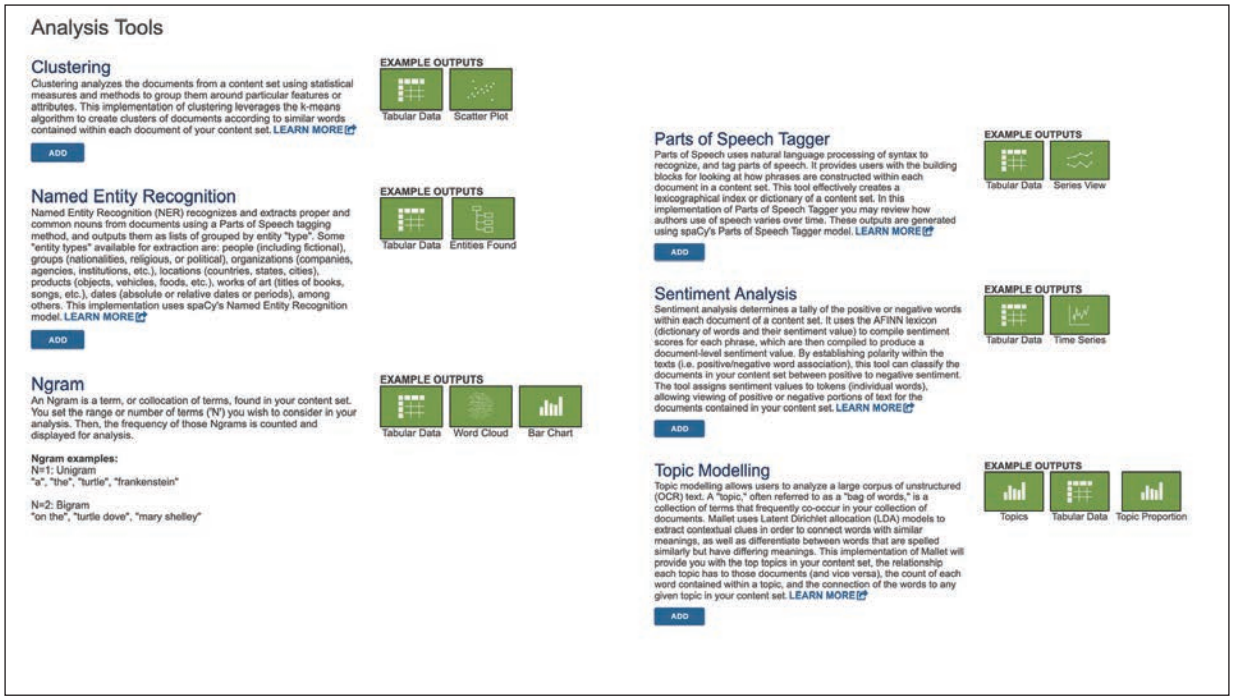



*图 6. “Gale 数字学术实验室”分析工具页面截图*

哪些部分。学生们在细致的内容集管理步骤中形成了一定的看法。文本清理又是一个缓慢且需反复多次的步骤，而在“Gale 数字学术实验室”中则包括下载 10 个清理后的文档样本和 10 个未清理的文档，并排比较，评价对数据集设置的文本清理选项是否有效。学生们不可避免地会看到他们疏忽了的错误，然后返回清理配置，做相应的调整。通过比较未清理数据集的分析结果和学生们花费时间规范和清理的数据集的分析结果，这一步骤的价值不言自明。后者的结果总是更加明显且有意义。

## 3. 分析文本

“Gale 数字学术实验室”最有用处的功能之一是一次文献内容与文本分析工具整合在一个直观地工作环境中。平台共整合了六种文本挖掘和分析工具，四种是开源工具（主题建模、命名实体识别、词性标注和聚类分析），两种为 Gale 定制工具（情感分析和 N 元语法）。分析工具页面提供了每种工具的详细定义，以及开发者说明文档和算法的链接，这对于研究结果的重现和引用非常重要（见图 6）。平台的帮助文档也详述了每种工具的用途和局限性。

学生们主要使用了这些工具中的四种，主要是因为时间的限制。他们一开始运行了主题建模分析，以判断他们的内容集中是否存在无法直接看到的主题或话题但可能值得进一步的研究。这一工具基于 MALLET，一种开源的、数字人文学者广泛使用的工具。直到目前，它都必须通过命令行运行，超出了这门概论课的教学范围。在使用“Gale 数字学术实验室”内置的主题建模工具时，学生们能够选择他们想要让算法返回的主题数量，以及每个主题下的单词数量。在大部分的情况下，我们使用默认值（10 个主题，每个主题 10 个单词），为每个小组的数据集返回最为相关的主题，其中很多都出乎学生们的预料，帮助他们确定了进一步研究的方向。

学生们运行 N 元语法分析找到他们的数据集中最相关的词语，并学习通过文本清理的步骤去掉常见停止词，得到更有意义的分析结果和可视化图表。当学生们开始探讨他们收集的资料时，这个工具也是一个良好的开端，在不同长度的 N 元语法（一元语法、二元语法、三元语法等）之间切换，帮助他们在更宽广的范围内看到特定的词语。

其中一名学生注意到了他们对数据集运行情感分析工具的重要意义，他评价说“情感分析到目前为止是最有用的，因为它真正捕捉到了 1934 年发生的事情。然而，我们了解到当人们逐渐习惯于大萧条时期的生活状况时，他们开始重新变得积极起来，即便仍处在大萧条的中期”。学生们能够从他们的分析中得出的一些结论，包括注意到现代菜单中与感恩节相关的资料比十九世纪更加世俗化，记录了所谓“中国餐馆综合征”的增多是反中国情绪的一种延伸，以及追溯了炒杂碎的发明以及中国食物的美国化，绘制了火腿蛋松饼作为一种早餐品种的起源及其随时间变得越来越流行。

学生们使用的最后一种工具是命名实体识别。这一工具从特定的内容集中提取并列出实体，例如地名、人名、组织名等。学生们使用这一工具作为他们绘制数据集的基础，例如从 1972 年尼克松访华相关的中国菜单数据集中提取地名，下载数据的 CSV 或 JSON 文件，然后利用 Google Fusion Tables 可视化数据。

## 4. 导出数据和外部分析

“Gale 数字学术实验室”的导出功能让学生们能够将在平台上创建的数据集与从其他来源收集的数据集结合在一起，然后利用外部工具分析所有资料。主题建模分析工具的 CSV 下载文件中包含所有 MALLE 老手希望看到的数据，额外的一列标出了每篇文献的 Gale 文档识别号。

一些小组也使用了来自其他开放资源库中的资料，并包括在了他们最终的数字展览中。学生们使用的其他外部平台也都是云存储的，包括用于创建项目陈述的 Knightlab 的 StorymapJS，以及用于文本挖掘的 Voyant。

## 5. 研究结果和数字展览

课堂教学的最后一个阶段是每个小组搭建一个 Omeka 数字展览，包含以下项目：

*探索如何在你的 Omeka 展览中创建页面和子页面来呈现相关信息。综合考虑设计、美学和实用性。（注：你选择的主题将会固定某些版式，例如菜单是边栏式的或位于页面上方。）*

***1. 主页 / 引导页*** *呈现出小组项目的细节，具备适当的导航功能。它应当链接到课程主页面上。为你的研究工作想出一个创意的题目！（例如，不要采用“第 1 组的项目”，而是“世纪之交的美食”，或类似的题目）。*

***2. 关于页面***

- *小组成员——包含照片和个人简介。*
- *每位小组成员应当将他们本学季的工作日志各自编辑成一个文档，存储在 Omeka 中。在每位小组成员个人简介的最后添加这个文件的链接。*
- *完成项目任务书*
- *个性化项目页面*

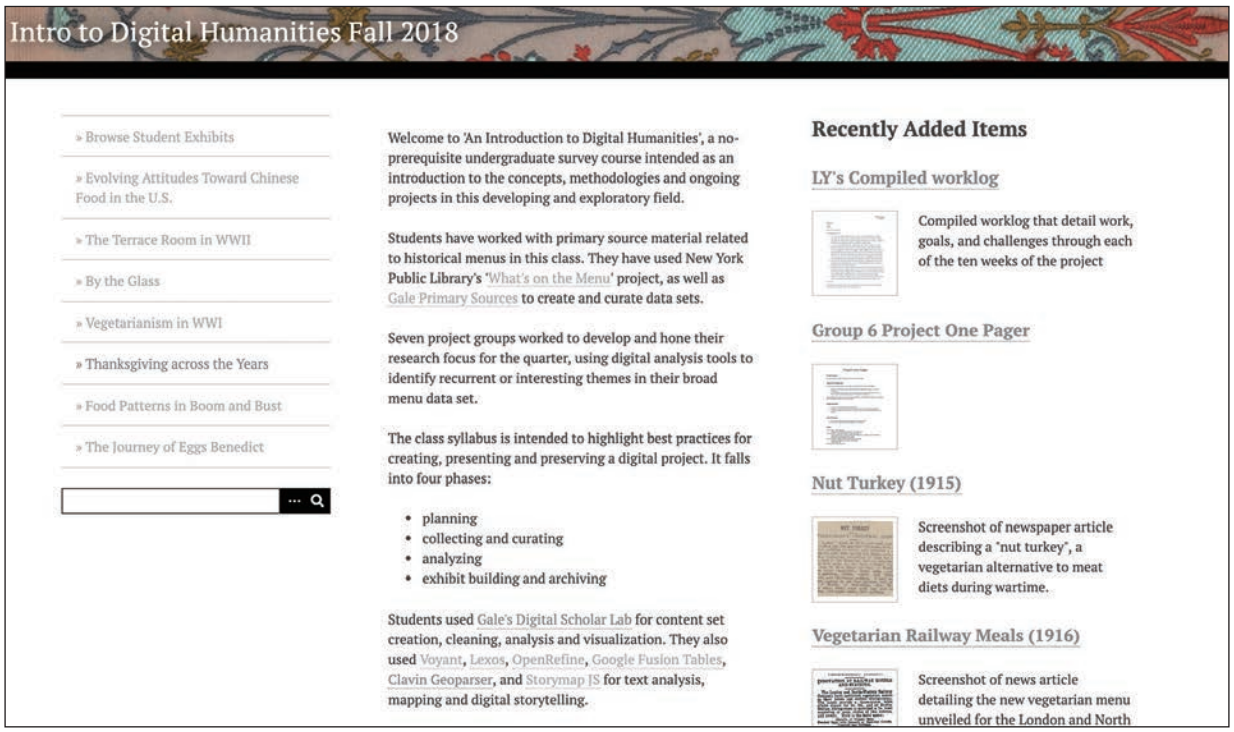



*图 7a. 2018 年数字人文概论课研究项目主页*

- *起草数据管理计划（你不必担心经费问题，但给出对备份和存储的一些考虑）。*

### *3. Omeka 中的资料集*

*每个人应当收集最少五份、最多十份资料，根据小组的规模创建一个约二十到五十份资料的合集。每个人应当为各自的资料正确完成元数据部分，并且整个组的资料集需保持一致且完整（没有空白区域）。*

### *4. 数据可视化*

*包括一个页面链接到至少一个详细的数据可视化图表，包含陈述、分析和来自你所在小组成员共同完成的 Omeka 资料集的图片。（Storymap 是满足这些标准的一个平台样例。）从与我们课堂教学相一致的类别中选择：*

- *文本分析（例如词语频率分析、情感分析、主题建模、主题分析等）*
- *地图绘制*
- *数据可视化（例如网状图）*

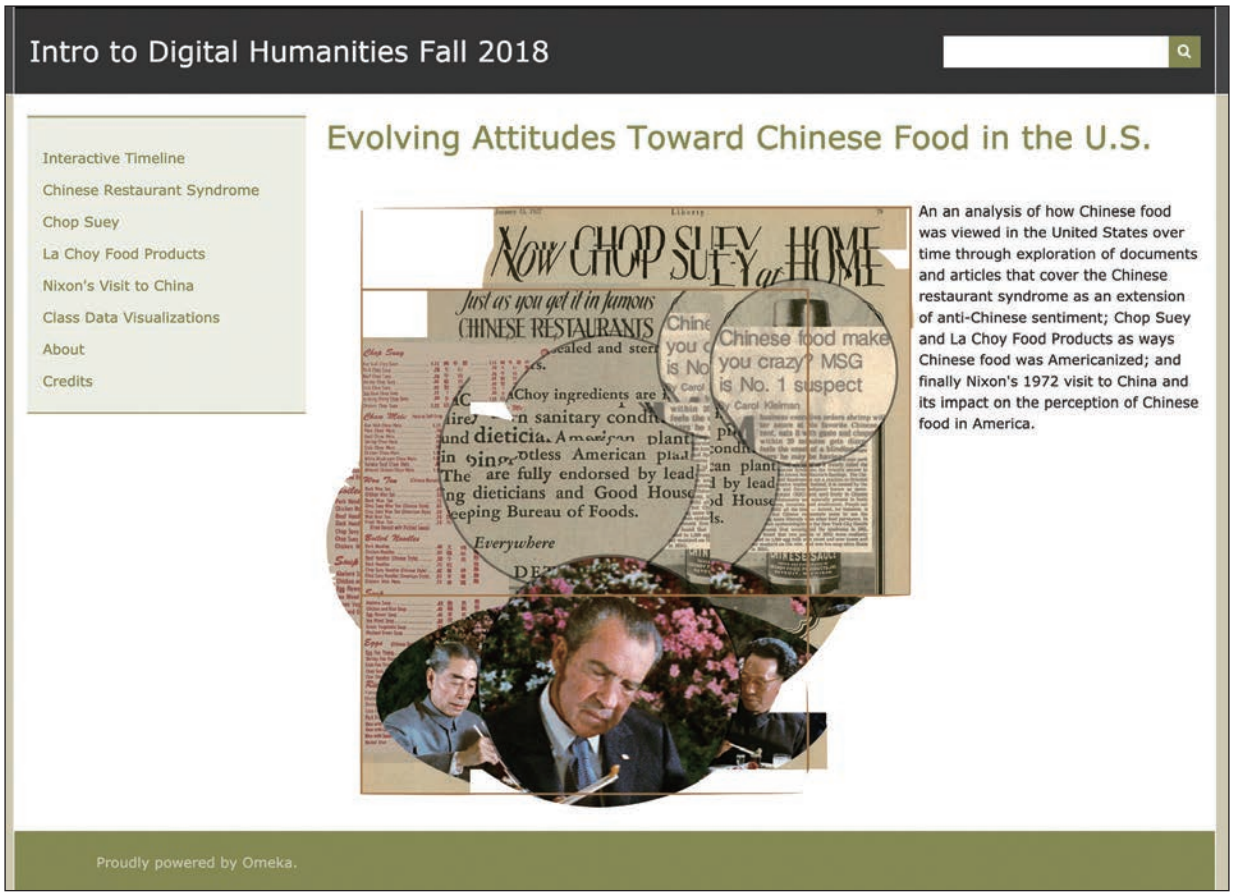



*图 7b. 学生展览：美国人对中国食物的态度演变*

*并且包括每个剩余类别中的示例（链接、图片或截屏的形式），这些类别基于你在课堂上实际操作环节的工作，可能包括：*

- *词云，基于 Voyant、N 元语法、主题建模或文本分析的课堂教学*
- *一幅或多幅地图，来自地图绘制课*
- *词语频率图表或主题建模旭日图*

最终的展览和学生们的演示令人印象深刻，展现出学生们对研究工作的热情和在各自团队中投入的努力。图 7a 到 7c 分别是该课程的项目主页、一项学生展览的首页和学生展览主页上一个链接的示例。

## 学习成果：将技能应用于工作

我们专门用了一节课的时间讨论如何找到和记录学生们正在学习和使用的可转移技能，不仅仅是在此数字人文课上，而且也包括他们的整个本科教育过程。这一对话发生在第七周，学生们已经和他们各自的团队工作了一段时间，他们的数字研究项目也有了相当大的进展。这一讨论主要是由讲师在和学生们做每周小组工作检查的过程中提出的，很显然，同学们都对毕业后的就业前景感到担忧，尤

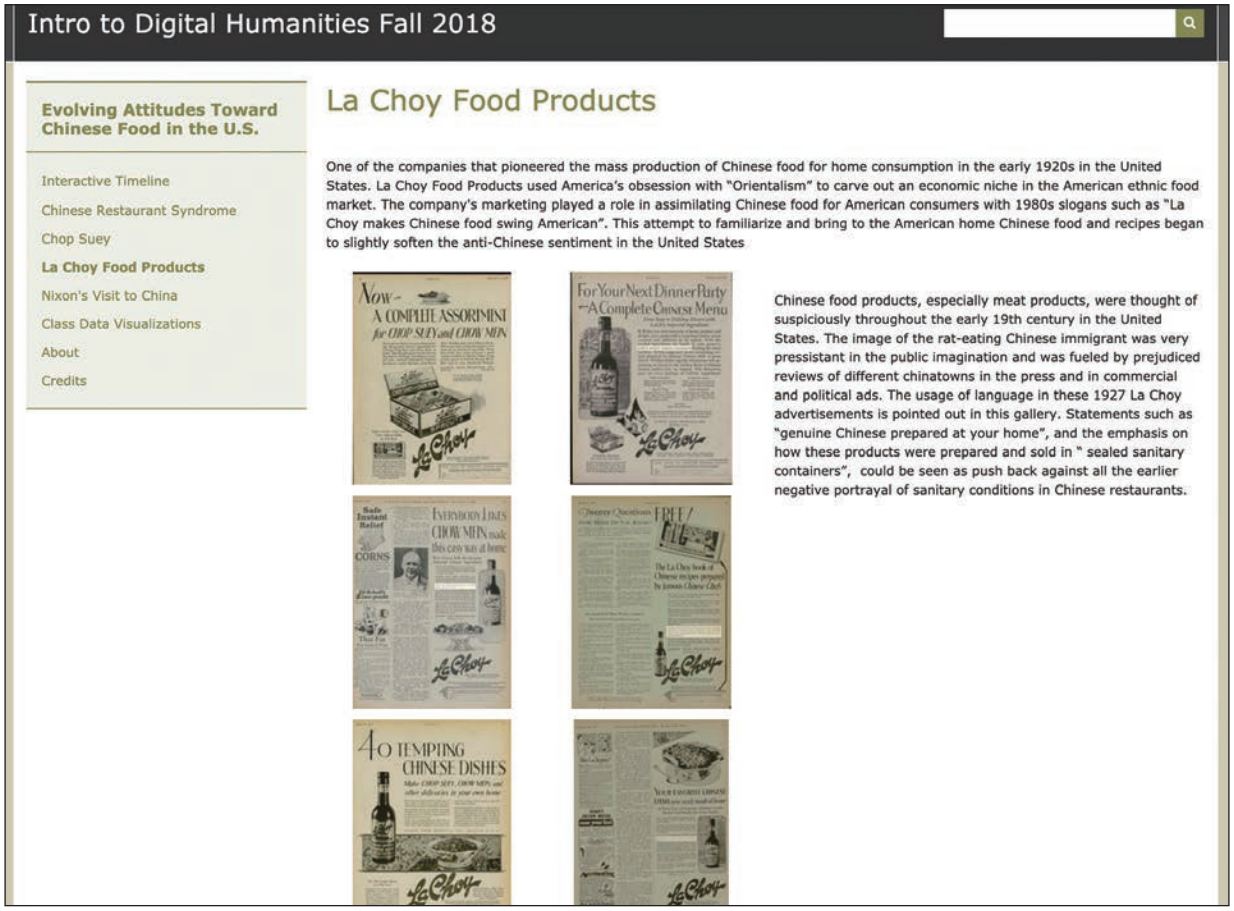



*图 7c. 美国人对中国食物的态度演变：La Choy 食品公司的链接*

其是人文专业的学生。除了提出本科生应当学会如何批判性思考、分析性推论和有效沟通外，学生们还明确提出了一系列他们在课堂上正在培养和使用的技能，都将会在职场上大有用处且对他们有利。从数据的获得和管理，到项目计划和设计，这些技能也包括各种形式的写作（发表在网上，记录技术决策等等），诠释和解读数据分析和可视化结果，作为团队的一份子开展研究工作，协同合作并承担责任。简言之，学生们从他们在课堂上从事的这类工作中看到了巨大的价值，将他们与即将面对的大学以外的生活连接在一起。

## 参考文献：


A´lvarez Sa´nchez, Adriana, and Miriam Pen˜a Pimentel. 2017. "DH for History Students: A Case Study at the Facultad de Filosofı´a y Letras (National Autonomous University of Mexico)." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Bretz, Andrew. 2017. "The New Itinerancy: Digital Pedagogy and the Adjunct Instructor in the Modern Academy." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Brown, Susan. 2016. "Tensions and Tenets of Socialized Scholarship." Digital Scholarship in the Humanities 31, no. 2 (June 1, 2016): 283–300. https://doi.org/10.1093/llc/fqu063.

Christian-Lamb, Caitlin, and Anelise Hanson Shrout. 2017. "Starting from Scratch? Workshopping New Directions in Undergraduate Digital Humanities." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Christie, Alex. 2017. "Building a Toolkit for Digital Pedagogy." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Clement, Tanya. 2012. "Multiliteracies in the Undergraduate Digital Humanities Curriculum: Skills, Principles and Habits of Mind." In Digital Humanities Pedagogy: Practices, Principles and Politics, edited by Brett D. Hirsch, 365–88. Cambridge, UK: Open Book Publishers.

Dunn, Stewart. 2016. "Quantitative, Qualitative, Digital. Research Methods and DH." Stuart Dunn (blog), September 21, 2016. https://stuartdunn.blog/2016/09/21/quantitative-qualitative-digital-researchmethods-and-dh/.

Estill, Laura. 2017. "Collaborative Knowledge Creation and Student-Led Assignment Design: Wikipedia in the University Literature Class." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Faull, Katherine M., and Diane K. Jakacki. 2015. "Digital Learning in an Undergraduate Context: Promoting Long-Term Student–Faculty Place-Based Collaboration." Digital Scholarship in the Humanities 30, no. suppl_1 (December 1, 2015): i76–82. https://doi.org/10.1093/llc/fqv050.

Gale. n.d. "Gale Digital Scholar Lab: Open New Research Pathways." Accessed March 5, 2019. https://www.gale.com/intl/primary-sources/digitalscholar-lab.

Gold, Matthew K., ed. 2012. Debates in the Digital Humanities. Minneapolis: University of Minnesota Press.

Google. n.d. "Google Fusion Tables." Accessed March 30, 2019. https://fusiontables.google.com/DataSource?dsrcid=implicit.

Hirsch, Brett D., ed. 2012. Digital Humanities Pedagogy: Practices, Principles and Politics. Cambridge, UK: Open Book Publishers. https://doi.org/10.11647/OBP.0024.

Hswe, Patricia, Tara LaLonde, Kate Miffitt, James O'Sullivan, Sarah Pickle, Nathan Piekielek, Heather Ross, and Albert Rozo. 2017. "A Tale of Two Internships: Developing Digital Skills through Engaged Scholarship." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

HumTech. 2015. "A Student Collaborators' Bill of Rights." University of California, Los Angeles, June 8, 2015. https://humtech.ucla.edu/news/a-studentcollaborators-bill-of-rights/.

Jenstad, Janelle, Kim McLean-Fiander, and Kathryn R. McPherson. 2017. "The MoEML Pedagogical Partnership Program." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Kelley, Shannon. "Getting on the Map: A Case Study in Digital Pedagogy and Undergraduate Crowdsourcing." 2017. Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Kennedy, Kara. 2017. "A Long-Belated Welcome: Accepting Digital Humanities Methods into Non-DH Classrooms." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Koh, Adeline. 2014. "Introducing Digital Humanities Work to Undergraduates: An Overview." Hybrid Pedagogy, August 14, 2014. http://hybridpedagogy.org/introducing-digital-humanities-work-undergraduates-overview/.

Ledolter, Johannes, and Lea VenderVelde. 2019. "A Case Study in Text Mining: Textual Analysis of the Territorial Papers." Digital Scholarship in the Humanities.

Lexos (website). Wheaton College (Norton, MA). Accessed March 30, 2019. http://lexos.wheatoncollege.edu/

upload.

Locke, Brandon T. 2017. "Digital Humanities Pedagogy as Essential Liberal Education: A Framework for Curriculum Development." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Mahony, Simon, and Elena Pierazzo. 2012. "Teaching Skills or Teaching Methodology?" In Digital Humanities Pedagogy: Practices, Principles and Politics, edited by Brett D. Hirsch. Cambridge, UK: Open Book Publishers, 2012.

Martin, Allan. 2008. "Digital Literacy and the 'Digital Society.'" In Digital Literacies: Concepts, Policies and Practices, edited by Colin Lankshear and Michele Knobel, 151–76. New York: Peter Lang.

Mauro, Aaron, Daniel Powell, Sarah Potvin, Jacob Heil, Eric Dye, Bridget Jenkins, and Dene Grigar. 2017. "Towards a Seamful Design of Networked Knowledge: Practical Pedagogies in Collaborative Teams." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." University of Massachusetts, Amherst. http://mallet.cs.umass.edu.

MLA Commons: Digital Pedagogy in the Humanities (website). Accessed February 26, 2019. https://digitalpedagogy.mla.hcommons.org/.

Modern Language Association. n.d. "Text Analysis." MLA Commons: Digital Pedagogy in the Humanities. Accessed March 26, 2019. https://digitalpedagogy.mla.hcommons.org/keywords/text-analysis/.

Morris, Sean Michael. 2013. "Decoding Digital Pedagogy, Pt. 1: Beyond the LMS." Hybrid

Pedagogy, March 5, 2013. http://hybridpedagogy.org/decoding-digital-pedagogy-pt-1-beyond-thelms/.

Murphy, Emily Christina, and Shannon R. Smith. 2017. "Introduction." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Murphy, Emily Christina, and Shannon R. Smith. 2017. "Undergraduate Students and Digital Humanities Belonging: Metaphors and Methods for Including Undergraduate Research in DH Communities." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

New York Public Library. n.d. "What's on the Menu?" Accessed March 12, 2019. https://www.nypl.org/node/163851.

Newbook Digital Texts (website). Accessed March 30, 2019. http://www.newbookdigitaltexts.org/.

Omeka (website). Accessed March 30, 2019. https://omeka.org/.

OpenRefine (website). Accessed March 30, 2019. http://openrefine.org/.

Saum-Pascual, Alex. 2017. "Teaching Electronic Literature as Digital Humanities: A Proposal." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Savonick, Danica, and Lisa Tagliaferri. 2017. "Building a Student-Centered (Digital) Learning Community with Undergraduates." Digital Humanities Quarterly 11, no. 3 (October 2, 2017).

Spiro, Lisa. 2012. "'This Is Why We Fight': Defining the Values of the Digital Humanities." In Debates in the Digital Humanities, edited by Matthew K. Gold. Minneapolis: University of Minnesota Press, 2012.

Stommel, Jesse. 2013. "Decoding Digital Pedagogy, Pt. 2: (Un)Mapping the Terrain." Hybrid Pedagogy, March 5, 2013. http://hybridpedagogy.org/decodingdigital-pedagogy-pt-2-unmapping-the-terrain/.

StoryMap (website). Accessed March 30, 2019. https://storymap.knightlab.com.

University of Washington Libraries. 2017. "Active Learning Classroom Resources: ALC Research & Pedagogy." Updated June 26, 2017. http://guides.lib.uw.edu/c.php?g=342298&p=2838500.

Uprichard, Emma. 2014. "Big-Data Doubts." The Chronicle of Higher Education, October 13, 2014. https://www.chronicle.com/article/Big-Doubts-About-Big-Data-/149267.

Voyant Tools (website). Accessed March 30, 2019. https://voyant-tools.org/.

Wiersma, Ashley. 2013. "Talking about Digital Pedagogy." Cultural Heritage Informatics Initiative (blog), March 12, 2013. http://chi.anthropology.msu.edu/2013/03/talking-about-digital-pedagogy/.

# 作者介绍

**Ryan Cordell** 是美国西北大学英语系副教授，西北大学文本、地图和网络实验室（https://northeastern.edu/nulab/）核心教职创始人。他的研究方向是揭示生产、接收和修复技术如何影响了社区中文本的含义。他主要探讨十九世纪美国报纸的发行和再版，但他的研究范围也拓展到了计算和数字化对当代阅读、写作和研究的影响。Cordell 参与了多个数字人文项目，包括 NEH 和 ACLS 资助的“疯狂传播的文字”项目（http://viraltexts.org），并且是挖掘数据项目“海洋交流”（(http://oceanicexchanges.org)）的主要研究人员。在西北大学，他担任英语系研究生部主任，并领导 Huskiana 凸版印刷实验工作室（http://oceanicexchanges.org）。他也是善本研究中心 Andrew W. Melon 书目评析学会（http://rarebookschool.org/admissions-awards/fellowships/sofcb/）的高级成员，以及 MLA 书目学和学术编辑论坛执行委员会的成员。

**Joris van Eijnatten** 是荷兰乌特勒支大学文化史教授。作为一名文化历史学家，他涉足多个相互关联的领域，包括思想、宗教、媒体和传播学的历史。他尤其擅长数字历史研究（数字历史、数字人文）。他目前的研究项目方向是十九世纪和二十世纪文化模因（词语和概念不断重复出现的规律）的数字分析，这项研究基于报纸、期刊、议会记录和其他数字化文本。尽管他对 Python 充满热情，但他仍在撰写一本有关十八世纪荷兰共和国历史的传统著作。他的教学涵盖多个方面，从“大历史”和文化史到现代的“主义”（isms）概念史。Joris van Eijnatten 是开放获取期刊《历史、文化和现代性国际期刊》（International Journal for History, Culture and Modernity）的共同创刊人和主编。

**Tessa Hawswedell** 是一位文化历史学家，参与了两项大型数字人文研究项目对数字化历史报纸的大量研究。作为“海洋交流”项目（由挖掘数据挑战项目和 EHRC 资助）的一员，她开展了对“《泰晤士报》数字典藏”（TDA）的文本挖掘工作，以探讨十九世纪对英国女性移民的争议。她也参与了 HERA 资助的“不对称的相遇”项目，研究多种语言欧洲报纸典藏中对欧洲城市的争议。最近，她刚刚成为《重新绘制中心和边缘：欧洲和全球背景下的不对称相遇》（Re-mapping Centre and Periphery: Asymmetrical Encounters in European and Global Context）一书的共同编者。

**Sarah Ketchley** 是美国华盛顿大学近东语言和文明系的讲师、马里兰大学学院兼职副教授。她目前的研究和教学内容是十九世纪末二十世纪初埃及学黄金时期的历史。她开发并多次为本科生和研究生开设数字人文入门课程。2012 年，她参与创办了一家创新性的出版社——Newbook Digital Texts，旨在重新定义和重新构建传统学术研究、出版和教育。出版社欣欣向荣的实习生项目为学生们提供了独一无二的机会，从事一次文献资料研究，转录文本和编码，以及创建数字地图和可视化图表。Sarah 于 2018 年作为数字人文专家加入 Gale，负责通过“Gale 数字学术实验室”推动和支持数字学术研究。

## Authors

**Julianne Nyhan** 是英国伦敦大学学院数字信息研究副教授，并且主管文 / 理科硕士课程。她也是伦敦大学学院数字人文中心的副主任。Julianne 发表了很多有关数字人文历史的论文，最新（与 Andrew Flinn）发表的一篇是《计算和人文：数字人文的口述历史发展方向》（Computation and the Humanities: Towards an Oral History of Digital Humanities，2016 年春）。她也是 Leverhulme 基金会资助的斯隆爵士（Sir Hans Sloane）手稿目录大英博物馆联合研究项目（https://tinyurl.com/y7zvrthm）的研究员之一，挖掘数据挑战“海洋交流：追溯历史报纸中的全球信息网络”项目（http://oceanicexchanges.org/）的英国首席研究员，以及“重要传承研究和欧洲的未来”居里夫人行动项目（http://cheurope-project.eu/）的研究员之一。

**Ulrich Tiedau** 是一位历史学家和数字人文学家。他在大部分的职业生涯中都在人文和传播技术的边缘附近开展研究工作，并发表了很多有关比利时、荷兰、德国历史以及远程教育和数字学术方面的论文。他是伦敦大学学院数字人文中心的副主任，也担任《荷兰十字路口：低地国家研究期刊》（Dutch Crossing: Journal of Low Countries Studies）的主编。这本期刊在 2009 年 12 月获得了学术期刊编辑委员会颁发的“杰出编辑成就凤凰奖”（Phoenix Prize for Significant Editorial Achievement）的荣誉奖。他是历史研究学院（IHR）低地国家历史研究研讨会的联合召集人。


**圣智学习（北京）教育科技有限公司**

北京市海淀区科学院南路 2 号融科资讯中心
C 座南楼 707 室
邮编：100190

电话：+86 10 8343 5000
传真：+86 10 8286 2089
邮箱：GaleChina@cengage.com
**www.gale.com**