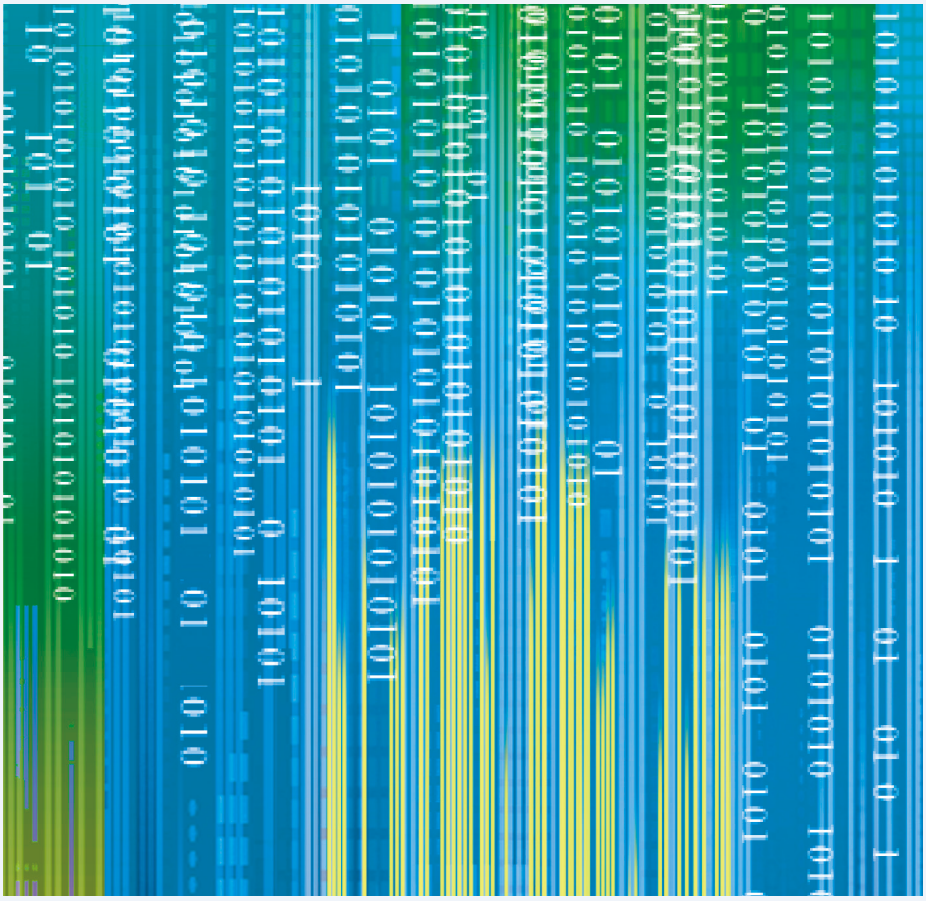
大数据与中国历史研究

第❷辑



Big Data and the Study of Chinese History

付海晏　主编

大数据与一流历史学科建设 …… 马　敏

1736～1739年中国宗教人口大普查 …… 高万桑

1933年中国纺织业产量与产值的再估计
——1933年中国区域制造业的量化研究之一 …… 徐　毅　葛　冰

近年我国经济区划及动态变化 …… 葛　非

留学前后：留日士官生群体特征研究 …… 王志敏

寻求对过去更好的理解：江南近代早期经济研究的新途径 …… 李伯重

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

大数据与中国历史研究

第❷辑

付海晏 主编

Big Data and the Study of Chinese History

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

本刊编委会

目　录

·中国历史研究中的数据库建设·

大数据与一流历史学科建设 …… 马　敏 / 3
清代商税数据库建设的初步构想…… 倪玉平 / 10
中国近代寺庙登记表数据库刍论…… 付海晏 / 21

·专题论文·

1736—1739 年中国宗教人口大普查 …… 高万桑 / 47
1933 年中国纺织业产量与产值的再估计
——1933 年中国区域制造业的量化研究之一 …… 徐　毅　葛　冰 / 88
近年我国经济区划及动态变化 …… 葛　非 / 108

·学位论文·

留学前后：留日士官生群体特征研究 …… 王志敏 / 143

·讲座实录·

寻求对过去更好的理解：江南近代早期经济研究的新途径 … 李伯重 / 189
从学生学籍卡到量化历史研究 …… 梁　晨 / 208

·史料选编·

民国二十年（1931）一月立大成裕记（五个月）转本老账账目
一览表…… / 219

稿　约…… 231

中国历史研究中的数据库建设

大数据与一流历史学科建设*

马　敏

今天的主题是“大数据与一流人文社会科学创新战略研讨会”。我觉得刚刚邢老师在更广的层面谈一流学科，谈得非常好。我今天围绕大数据历史学科谈几点思考。主要从以下三个方面谈：一、根据历史谈谈目前发展的趋势；二、我校大数据历史研究的基础和设想；三、简单谈一谈需要突破的困难和需要的支持。

大数据在历史学界正在兴起，特别是最近十年，发展得很快。尤其是在国外，形成了几个大的人文社会科学数据库。一是美国整合公共微观数据库（Integrated Public Use Microdata Series，IPUMS），这是一个跨学科的数据库，有很多数据在里面。二是加拿大巴尔扎克人口数据库（BALSA Population Database，BALSAC），是进行人口研究的。三是荷兰历史人口样本数据库（Historical Sample of the Netherlands，HSN），这个数据库我在欧洲的时候，他们专门给我演示过，很厉害。我请他们给我演示中国的人口历史，他们的数据库包含从秦开始一直到现在的数据。好多朝代人口的数据都在里面，基本上是连贯下来了，所以可以做很多的分析。这个数据库做了近二十年。四是瑞典斯堪尼亚经济人口数据库（Scanina Economic Demographic Database，SEDD）。五是美国犹他人口数据库（Utah Population Database，UPDB），主要内容是犹太人和他们的家属。

---

* 本文根据马敏在“大数据与一流人文社会科学创新战略研讨会”（2017 年 6 月 9 日）上有关“大数据历史”学科的发言整理而来。

2006—2010年的五年间，引用过以上五个数据库的学术成果已达2360余篇。而就纯粹跟历史相关的三个数据库，从2006年到2010年，历史学术成果达到117篇，这证明大家都在用。哈佛大学数字人文研究走在前面。包弼德（Peter Bol）是哈佛大学的副教务长，也是很多学术机构中心的主任。他的学术研究领域实际上是中国古代史，主要偏重于唐宋思想史，但他现在成为哈佛大学数字人文科学的领军人物，包括我们熟知的哈佛的网络课程都是他主导建设的。他首先建立了一个地理分析中心，然后主持了哈佛的在线课程，这个杨宗凯校长非常清楚。另外他还担任了哈佛中国历史地理信息系统管理委员会主任，以及国际历史人物传记资料库主任。下面我想详细介绍一下中国历代人物传记资料库（China Biographical Database，CBDB）。它是北京大学中国古代史研究中心与哈佛大学费正清研究中心、台湾中研院历史语言研究所联合开发的。他们输入了中国历代37万多个人物的传记资料，大概是从秦汉至19世纪，现在正在输入明清时期的人物传记资料。我去看过这个资料库，它非常好，好在什么地方呢？比如，我想让它找几个人，很生僻的人，我就输入了两个人名，一个搜索出来了，一个没有搜索出来，这是怎么回事？没有搜索出来的那个人是晚清商会的刘先加，一个进士，而晚清的人物他们还在做。但是古代的一输进去，相关的资料、研究都出来了，所以说他们这个做得还是很漂亮的。中国历史地理信息系统（CHGIS）是由哈佛大学和复旦大学合作完成的，这是一个从秦始皇到辛亥革命的历史地理数据库。利用这些数据库可以进行人口的研究、人事官员的研究、宗教的研究等等。哈佛大学在线公开课程（HarvardX），开始有约20万个注册用户，到2014年即超过100万个，到目前为止大概已有200万个遍布全球的注册用户。

哈佛大学有一个著名的社会科学量化研究和大数据中心——量化社会科学研究所。量化社会科学研究所注重的是人文科学与大数据的结合，和北京大学有合作项目，还独立开辟了很多站点。他们在人文科学方面进行了大量的数字化、量化研究。我2016年11月访问哈佛大学的时候，专门去跟他们进行了交谈，他们愿意和我们进行合作，这个事情我跟杨校长也讲过。我们和他们正在积极地接触。陈志武也在做这方面的工作。陈志武是耶鲁大学的终身教授，非常有名的经济学家。他现在在北京大学成立了数字人文研究所，并担任所长。他现在也偏向数字人文研究，还出版了这方

面的著作，召开这方面的会议。北京大学在 2016 年召开了第一届“数字人文论坛”，该论坛以“跨界与融合：全球视野下的数字人文”为主题，对数字人文在历史学、语言文学和艺术学中的应用进行了深入探讨。在历史学方面，主要围绕着“史学与大规模史料的深度挖掘”展开，包括人物的维度与史料挖掘——中国历代人物传记资料库（哈佛大学）、虚拟仿真技术与考古实验教学（北京大学）、古籍数字化与史料的深度挖掘（北京大学）、清代人口册数据库项目（中国社会科学院）。第二届“数字人文论坛”的主旨是“数字人文”与“史学研究”的互动与共生。在此主旨下又有 4 个分主题：“数字人文与历史地理研究”“数字人文与史学研究”“数字人文与历史文献整理”“数字人文基础设施建设”。

香港科技大学的人文数字研究，在李中清（James Z. Lee）的领导下做学生学籍研究。他们收集了 1952—2002 年北京大学与苏州大学 15 万名本科生的学籍卡，并通过统计分析这些学生来自哪里，考察新中国建立以后社会流动的情况，还对高等教育的改革提供了不少建言。《无声的革命：北京大学与苏州大学学生社会来源研究（1952—2002）》这本书引起了很大的反响，我在政协开会的时候，俞正声主席专门提到过，它通过大量数据找出了证据证明社会流动是令人信服的。刘延东也鼓励要多做这方面的研究。李中清之前来过我们这里，现在也在和我们合作。目前他的想法，就是把他的研究扩展到更多的大学，也包括我们华中师范大学。因为华中师范大学前身是教会大学，保留了许多可以数据化的资料，他想把华中师范大学的数据收入他的数据库。李中清团队有一位重要成员——康文林（Cameron Campbell），已经被我们学校聘为长江学者讲座教授。他本科学的是理科，后来转学文科，他的特长就是研究大数据。目前他们正在做三个大的数据库：一是中国多代人口数据库，地点在辽宁双城，做得很细；二是精英教育数据库，解放以来的北京大学、苏州大学，现在扩展到了全国，包括华中师范大学；三是清代官员数据库，缙绅录以及相关资料数据库，通过这个数据库可以了解官员的出身和分布等信息情况。

民国大学生量化数据库，包括了很多大学学生的个人信息，目前已有超过 11.3 万名学生的个人信息。已完成的有上海交通大学、暨南大学、上海商学院，校对中的有金陵大学，筹备中的有齐鲁大学和华中大学等。原始档案种类有学生入学履历表、学生卡片、投考人履历书（报名表）、学生

履历书（学籍卡）。更重要的是该数据库可以和很多其他数据库匹配。基于清代缙绅录数据库的实证分析，可知获得功名越高的人，在京师做官的机会越大。通过这些数据，可以做一些相关性的分析和研究，比如旗人和其他人在其官员中的分布等。

接下来我就分析一下量化研究的意义，讲一下为什么要进行量化历史研究。该研究方法重视对长时段、大规模记录中的各种人口和社会行为进行统计描述及彼此间相互关联的分析，有助于避免传统研究方法（短时间和个案）极易犯的“以偏概全”的错误，是“选精”或“集萃”的史学研究方式；有助于揭示隐藏在“大人口”中的历史过程与规律，它能够丰富、完善我们对微观人类历史和行为的认识，帮助构建更为可靠的宏大叙事，促进我们对人类社会发展规律的进一步认识。从科研立项而言，2015年度国家社科基金重大攻关立项项目中涉及大数据的项目仅中国近现代史学科就有三项：抗战“大后方”资料数据库建设（西南大学）、蒋介石资料数据库建设（浙江大学）、清末民国社会调查数据库建设（中国人民大学）。

因为时间关系，我就简单谈一下我们学校的大数据建设。章开沅先生在《关于改进研究中国资产阶级方法的若干意见》（《历史研究》1983年第5期）一文中就提倡要重视中国近现代史研究中的人口等统计资料的重要价值，我也在1997年意识到要重视中国近代社会发展指标体系，经济研究不能简单地从举例来研究，而要从统计数据，从长时段进行研究。2001年，我和陆汉文发表了《建构民国时期（1912—1949）社会发展指标体系的几点思考》，2005年发表了《民国时期政府统计工作与统计资料述论》，另外我们还编了《民国时期社会发展统计资料汇编》，目前还没有出版。我们很早就开始从事数据化方面的研究工作。郑成林教授做了很多很好的工作，他和国家图书馆合作编了很多统计资料，如《民国时期经济调查资料汇编》（全30册）、《民国时期经济调查资料续编》（全30册）、《民国时期经济统计资料汇编》（全50册）、《民国时期社会统计资料汇编》（全20册）、《民国时期国情统计资料汇编》（全45册）、《民国时期国情统计资料续编》（全36册）等，资料很多，而且都是影印的，也都可以分析。另外现在由我主编，之前是刘望龄先生主编的《苏州商会档案丛编》（共12卷），还有50年代的两卷，正在出版。这是个很大的工程，世界上很多著名大学的图书馆都保存有这份资料。还有章先生主编的《辛亥革命史资料新编》（共8

卷)，这都是基础性工作。在人才培养方面，我校高研院开设了一个大数据历史研究生基地班，研究方向主要就是大数据历史，现在一共招了 35 个人。目的，一个是参与、引领历史研究的方法革命，另一个是实现人才培养的国际化。我们已经邀请了香港科技大学李中清教授来讲课，另外还有我们的长江学者客座教授康文林，再就是李伯重教授。李伯重是清华大学教授，现在也是我们学校的特聘教授。他以经济史研究著名，基本以明清经济史研究为主，在全世界都有很大的影响，也非常注重大数据历史的研究。康文林，他目前正在给我们的研究生上课，他的目的是在华中师大开设以“量化历史研究”为核心方向，以“中国多代人口数据库”“民国大学生信息数据库”等为训练平台的人文研究创新基地班。再就是与海内外相关专家合作撰写以“数字人文与历史研究”为主题的导引类教材，为华中师大乃至国内各大学历史专业开展“数字人文”方向的教学与研究工作提供基础性指导。另外，付海晏现在主编了一本《大数据与中国历史研究》，收集了有关大数据历史研究的文章，由社会科学文献出版社出版，这在国内是开辟性的。

那我们有什么设想呢？结合我们学校一流学科中国史建设情况来说，我们就是想要传承过去章开沅先生开辟的一些学术领域。同时在信息化时代，我们还要创新，把实证和理论相结合，做“有思想的学术，有学术的思想”。其中一个方向，就是结合时代的发展，进行可靠的、长时段的大数据与中国历史研究。我们的设想主要是建一个中国现代化历史进程大数据库，这个大数据库着重从数据的意义观测中国近代化的发展过程以及量化的说明。这个库现在有四个大的子库，当然这是目前初步的设想，以后有可能增加。分别是中国商会数据库、中国近代教会大学数据库、中国博览会数据库、中国近代宗教统计数据库。下面我会仔细介绍这些数据库。中国商会数据库，实际上我们已经搞了十年，但做的是基础建设，目前我们已经将很多有关商会研究的信息和资料放进去了，主要是完成学术文献总库系统开发（含后台和前台系统），还累计完成了 10 万余条数据的上传工作。此外，还有近 1T 数字化资料，正在进行数据描述，这是我们一期的工作，现在正在开展中。二期建设将在一期成果的基础上进行数据的深度挖掘和利用，这个是比较难的，还要和杨校长数字工程中心配合，做深度挖掘。怎么挖掘？这个方面，我们比较缺乏经验，需要提供一些技术方法来

进行挖掘数据以及利用数据。中国近代教会大学数据库，经过近三十年的建设，我们拥有一个庞大的数据库。这个我很清楚，因为美国亚洲基督教高等教育联合董事会联合鲁斯基金会支持了我们好几回。在他们的支持下，我们现在已经有了13所教会大学全部档案的缩微胶卷，美国耶鲁大学把这个送给我们一套，全中国就这一套。今后要查教会大学的一些档案，不必去美国了，到华中师范大学来查就可以了，我们这里是开放的。另外我们陆续购进2000余盒在华基督教差会档案缩微胶卷，那时候好多来华传教的差会的档案全都在我们这里，来查阅也是非常方便的。我们还有5000册有关基督教史的研究图书和百余种期刊，所以说我们这里的资料是很丰富的。文献中心已成为国内教会大学与中国基督教史研究资料收藏最丰富的单位(外界评价的)，并已将主要文献编成目录，录入数据库。下一步要进行的就是数据的整理、利用。还有很多其他的步骤，包括建一个相关的成果库。再就是中国教会大学建筑数据库，教会大学有很多中西结合的建筑，建筑方面有大量的数据资料，我们想把这些东西收集起来，章先生的女婿就在专门做这个研究，做得非常好。我们还想建设中国教会大学师资与学生数据库，把很多教会大学学生的学籍卡信息输进去，包括师资的情况，这个研究可以有很多分析和借鉴。以上就是我们教会大学数据库准备做的一些事情。中国博览会数据库，博览会我们也做了二十年，收集了大量的资料，包括我在美国、欧洲、我国台湾等这些地方收集了很多资料。现在我们要把这些资料做成数据、成果库、史料库，及相关的统计、动态库。中国参与了很多的博览会，有很多数据是可以统计的，要进行统计整理和运用。中国近代宗教统计数据库，是付海晏教授在做的。他从佛教这个角度，从事近代寺庙财产研究、人口研究、信仰与精神研究、寺庙的宗教地理研究。他们也找了大量的数据，尤其是我们国家从民国以来的寺庙统计，他们基本上都搜集到了，非常齐全。我们已经谈到，用四个大的子数据库合成一个中国现代化历史进程大数据库。这在以后可能是要发展的，包括彭校长(彭南生)在做的近代工业化、产业史和工业史，也可以搞数据库，下一步还在设想中。

目前我们正处于发展的关键时期，前面我说过，现在大数据历史发展方兴未艾，大家都在朝这个方向努力，哪个抢先一步占领了这个制高点，就抢占了先机，占领过后，很多人就要来你这里学，所以北大、复旦好多

学校在这方面发展得很快。但是大数据历史这方面，特别是我们构想的中国现代化历史进程大数据库，还是非常有特色的，我们要抢先把它做出来，这才是最关键的，我们要占领这个制高点，所以我们需要学校的支持。

首先，需要一个平台，我们的地方小，没有设备，需要大批存储的地方。从技术上说，历史数据的采集、提炼与整理的工作是非常重要的。因为数据库可不可靠，关键在于数据的采集是否可靠，而历史的数据，虚的成分很多，所以中间还有一个去伪存真、辨析的过程，这都是很难的。还有资料库、数据库的设计、匹配、管理、检索、分析，这都不是我们的强项，希望相关的技术专家来帮助我们。

其次，就是人才，大数据研究的人才，这类人才是跨学科、跨领域的。所以我们要思考，怎么进行建设，怎么进行考核，怎么加大资助的力度，特别是人才引进的力度。我们引进人才一定要根据学科的需要，尊重专家的意见。尤其是看重的一些人才，要不惜一切办法把他引进来。

再就是经费方面，希望进一步加大对数据库持续性的资助。目前我们遇到的困难，关键还是经费，如果没有经费，再要进步是很难的，这个工作做起来是很花钱的。我们现在落实了一流学科建设的经费，下一步，研究的经费、大数据库的经费，也要进一步落实。

最后就是要进行跨学科的合作，这不是我们哪一个学科能做的，也不是我们历史学家能做的，包括历史学、社会学还有其他的政治学、信息科学等都要配合好。资料的采集方面、方法方面、学科建设方面，我觉得都需要跨学科的合作。

（马敏，华中师范大学中国近代史研究所教授）

清代商税数据库建设的初步构想*

倪玉平

20 世纪 30 年代中国经济史学形成伊始，学者们即开始关注清代的商税，产生了一批有影响力的学术成果，特别是罗玉东对厘金的研究堪称经典。新中国成立之后，尤其是改革开放以后，愈来愈多的学者投入这一领域，如吴承明、郭松义等人对商品流通的研究，汤象龙、陈诗启、戴一峰、任智勇等人对海关的研究，经君健、许檀、祁美琴、邓亦兵、廖声丰等人对关税的研究，何烈、周育民、徐毅等人对厘金的研究，都取得了积极的进展。

不过，总体上看，到目前为止国内外学者对清代商税的研究尚可以进一步深入。商税的种类既多且杂，难以方便为学者所用。考虑到清代商税研究内容的复杂性和史料整理的滞后性，很有必要在对清代商税的各个方面（常关、海关、厘金和杂税）具体研究的基础上，汇总出清代商税与社会经济的互动关系，并在各分项史料长编的基础上，建立清代商税数据库。这也是国家社科基金重大项目“清代商税研究及其数据库建设（1644—1911）”（项目号：16ZDA129）立项的初衷之一。

本文将就项目组关于清代商税数据库的构想做一简单分析，不当之处，敬请方家指正。

* 本文系国家社科基金重大项目“清代商税研究及其数据库建设（1644—1911）”（项目号：16ZDA129）的阶段性成果。

一 目标设计

史学研究的基础是史料，历史数据库的建设更是离不开海量的数据。清代商税研究存在着涵盖内容的复杂多样性和研究成果的失衡性特点。为学术界提供系统的常税和杂税长时段系列数据，对海关和厘金的既有数据进行适当的修正，从而为学术界提供一套完整而可靠的清代商税收支数据系列，并在此基础上分析商税与财政和社会的互动关系，是本项研究的重要目标。而为达到这一目标，就非常有必要花大力对大量的第一手史料进行发掘和整理，下最笨的功夫，一步一个脚印，踏踏实实，勤勤恳恳。

考虑到这一点，项目组关于子课题的设计逻辑框架如图 1 所示。

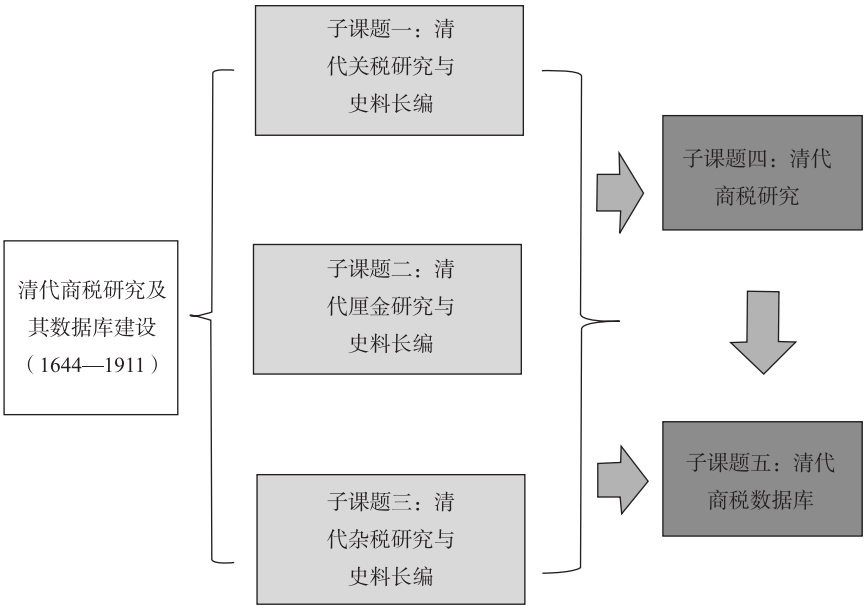


图 1 “清代商税研究及其数据库建设（1644—1911）”子课题设计框架

也就是说，要通过子项目的积累和研究，完成有清一代关税史料长编、厘金史料长编、杂税史料长编等子项目史料长编，尽可能完备地搜集相关史料，然后在此基础上，建设清代商税数据库。

清代商税数据库的建设，以全面、丰富、准确、便捷为最高宗旨。

目的：分享商税各门类的历史数据，方便同行交流与查阅。