# 浅谈社会科学研究领域中数据资源的有效利用
## —以中国边疆史地研究中心为例

白　妍

社科院 中国边疆史地研究中心

**【摘要】**为了开发具有边疆史地研究特色的、适合边疆史地研究工作者需求的公共数字资源平台，发挥数据库的强大优势，我们从语音研究工作中引入了「标注」概念。本文将探讨如何通过「标注」的方法提高学术资料的可用性和针对性，在扩大数据检索范围的基础上提高数据检索的命中率。从而达到辅助科研的目的。

**【关键词】**数据资源、全文检索、存储性能、检索效率

### 一、时代背景

信息科学与技术的飞速发展及广泛应用，对社会生活、经济发展、科学研究等领域都产生了深刻的影响。信息技术在为这些领域创造新的可能性的同时，也产生了网络时代所独有的新的行为模式。信息技术的产物，数字资源是文献信息的表现形式之一，是将计算机技术、通信技术及多媒体技术相互融合而形成的以数字形式发布、存取、利用的信息资源总和。而数字资源整合是一种数字资源优化组合的存在状态，是依据一定的需要，对各个相对独立的资源系统中的数据对象、功能结构及其互动关系融合、聚类和重组，重新结合为一个新的有机整体。

我们所处的这个时代，呈现出社会科学研究与科学技术密不可分的特征。社会科学研究是精神的、知识的生产，是利用旧文献产生新文献的过程。而现今的文献存在方式已经发生了根本性的变化，随着社会历史的不断发展，载体从历史早期的

甲骨、竹简等材料已转化为今天的光盘、磁盘等各种电子信息存储设备。

当网络、电脑、数字电视和家庭宽带这些原本遥不可及的数字化产品渗透到当代人生活中以后，数字神话就不断上演。今天的数字化技术已让天涯变咫尺，天堑成通途，让许多梦幻成真。如果说今天还有什么神话的话，那就是数字化。

当今科学研究面临的新机遇、新问题是信息技术正在改变着学术研究的手段和方法。从 World Wide Web 的兴起，以及随之而来的社会信息数字化、网络化的浪潮，不仅给社会科学带来了一个新的研究对象，而且其中气象万千、随历史长河奔流的海量数据，以及各种新型的工具，例如搜索引擎等，也给社会科学的传统研究方法和研究手段带来了新的机遇和挑战。然而，目前社会科学领域利用信息技术开展工作的现状与信息技术以及社会数字化信息不断丰富积累带来的潜力之间，还存在很大的空间。这个空间既表现在大量的数据有待人们应用新的社会科学方法来挖掘和开发，也表现在信息技术为适应社会科学的需要还应积极参与并发展出相应的新工具。社会科学研究在一定意义上和自然科学一样，都是信息（数据）的加工和处理。

## 二、边疆中心信息化的现状与设想

成立于 1983 年的社科院中国边疆史地研究中心（以下简称「边疆中心」）是目前中国唯一以「边疆」为研究对象、成建制的学术机构，系中国社会科学院直属的开放性研究机构，主要任务是继承和弘扬中国边疆史地研究的优秀遗产，组织和协调本单位及全国边疆史地领域的学术研究。就边疆的界定与边疆地域范围而言，国内从东到西，依次分为东北边疆（辽宁、吉林、黑龙江）、北部边疆（内蒙古）、西北边疆（甘肃、新疆）、西南边疆（西藏、云南、广西）与海疆（黄海、东海与南海，以及台湾、海南两省），这些均为边疆中心的研究领域。与此同时，沿着这些边疆地域的外缘，还有邻国的毗邻地区，这些也是边疆中心的研究领域。另外，边疆学理论也是中国边疆史地研究的重要方面。

自 2002 年以来，边疆中心先后主持了《东北历史与现状系列研究工程》、《新疆历史与现状系列研究项目》，协助中国社会科学院实施《西南边疆历史与现状系列研究项目》。从本年初开始主持《北部边疆历史与现状研究》项目，同时申报「海疆研究」项目。此诸项目的实施，在使得边疆区域研究范围愈益扩大的同时，也对资料

的依赖程度越来越深。

另一方面，自成立至今，边疆中心通过多年探索，逐渐形成研究趋向与重点，主要围绕三大研究方向展开，即（1）以中国边疆史地研究为基础，开展中国古代疆域史、中国近代边界沿革史、中国边疆研究史三个方向的综合性研究；（2）开展当代边疆地区的稳定和发展方面的研究；（3）构建中国边疆学的理论框架。此诸研究，离不开古今中外各种文字记载。

从上述的中国边疆史地的研究范围来看，边疆研究所涉及的资料面广且杂，古今中外，各种涉及边疆历史与现状的文字资料，都在边疆研究者视线之内。但也因此使得学术资料占有量，越来越成为制约边疆中心科研发展的重要因素，学术资料保有量的多寡，在某种程度上决定着学术研究的深入程度。

信息技术的发展不但改变了人们的阅读习惯，而且正在改变着科研的手段。通过网络，科研人员可以利用搜索引擎快速地获得大量信息，并希冀从中提炼出真正有用的资料信息。但令科研人员苦恼的是，通过搜索引擎检索到的成百上千条信息中的大量条目，绝大多数与科研主题无关。也就是说，搜索引擎能够提供给我们的是「某一篇网页里包含的信息」，而科研人员所需要的则是在一些网页集合中所蕴含的某一方面的信息。可能性与现实性之间，存在着较大的距离。实际上，一些已经发展与应用多年的技术，例如信息过滤、信息提取、文本挖掘、文本综述等，在某些特定领域的场合（例如：电子商务）已经被成功地应用，但在学术领域，要解决上述问题，在效果和效率上尚有很大的改进空间。

可以说，海量的数字化、网络化信息固然带给我们方便，其有效利用的潜力也愈益显现，但这并不意味着实际利用没有困难。当信息技术与社会科学的结合在技术与条件的准备上，还没有达到水到渠成的境界时，面对动辄成百上千条的数据量，如何在新的社会科学研究方法有待诞生的现有条件下，使科研人员能够相对准确、比较专业地从海量信息中提炼出其所需要的真正有用的资料信息，从而达到辅助科研的目的，是值得从事网络信息工作与研究的人们积极思考并着力去解决的。

从信息资源的源头着手，即从产生数字化信息的学术史料这一信息源出发，在浩如烟海的史料和学术资料中，找到一种能够更准确地提炼出科研所需要的专业信息的方法，或许是解决上述问题的重要途径。这个途径就是我们将引入的「标注」

方法。

## 三、「标注」的提出和应用

### 1、「标注」的提出

所谓「标注」是我们从「语音」研究领域引入的一个概念。在语音研究中，语音库的标注是按照一定的原则，对语音或相关物理信号中的某个语音或信号片段，从语音学、语言学及副语言学的层面进行分析和描写。「标注」即指这种描写所生成的符号标示。语音标注一般标注那些有语音学、语言学或者副语言学意义的信息。语音学层面的标注包括音段和韵律标注；语言学层面的标注可以有词性、句法、语用等信息标注；副语言学层面的标注包括一些非语言现象和情感状态等。没有标注的语音库是生语料库，研究用途非常有限，而具有标注信息的语音库则可以用来进行语音学基本研究、语音识别建模研究以及构建语音合成系统等。

目前，在边疆研究中，对经数字化处理的史料检索方式主要是通过搜索引擎对「关键词」进行检索。「关键词」源于英文「keywords」，特指人们在制作使用索引时，所用到的词汇，来源于图书馆学。关键词搜索是数据检索的主要方法之一，是希望访问者了解的资料、信息等具体内容的用语。用户输入一个词或句子，搜索引擎据此检索出来的网页即为搜索结果，其中用户输入的内容就是「关键词」，关键词是用于表达文献主题内容的词句或短语。

那么在对史料进行数据化的过程中，所适用的「关键词」即是直接从题目、小标题、正文或摘要里抽取的部分词汇，标引的都是单一概念。这样处理的结果有利也有弊。所谓「利」就是方便查找，一看便知。而学术资料的应用并非如此单纯，不具备这种「单一」性。正是这种「单一」性限制了学术资料的广泛利用，也就是「弊」之所在。以一篇学术资料为例，虽主题突出，但内容涵盖则非常广泛、丰富，若从不同学术角度进行分析，则有不同的学术价值。比如一篇探讨边疆地区少数民族婚姻状况的文章，会涉及当地的社会状况的各个方面，其信息量包括历史、政治、经济、文化、教育、宗教、民俗、语言、服饰等等一系列信息，而这些信息对于相应研究领域的学者来说都是非常宝贵的学术资料。如果仅用几个「关键词」来描述这篇文章的话，则很难将所有的信息涵盖其中。也就是说，「关键词」所表述的只是

这篇文章直接指向的内容，而其它层面的信息则被忽略了。由于「关键词」在信息方面的严重缺失，使得从事相关研究的学者，无法通过搜索引擎利用「关键词」检索出所需要的资料信息，从而导致资料信息的不完整性，其学术价值也自然地被「人为」地降低。

此外，还有很多因素也会直接影响或制约检索的结果。譬如从事海疆研究的学者要查找一篇有关钓鱼岛的资料时，就会在搜索引擎中输入关键词—「钓鱼岛」，结果，涵盖的只是含有「钓鱼岛」的资料。事实上，「钓鱼岛」在汉语语境中又被称为钓鱼台、钓鱼台群岛、钓鱼台列岛、钓鱼列岛等，在日本则称为「尖阁列岛」等，那么，在检索的结果中自然无法体现出「钓鱼岛」之外的其他文献。这样的搜索结果，既无法满足研究者的需求，也无法体现出文献的完整性。类似的情况实际上并非个案，而是普遍存在的现象。

文献对社会科学研究的重要性不言而喻，在对边疆问题进行学术研究的实践中，我们愈发体会到，一方面是海量数据的现实存在，另一方面则是利用上的局限性。准确而全面地体现文献的信息，最大限度地囊括、涵盖文献的信息含量，是我们引入「标注」概念的基本出发点与动力。

**2、「标注」的应用**

首先，我们应该发挥数据库的强大优势，在创建边疆研究数据库之初，即对数据库进行严格、细致的分类。第一级数据库依据地域原则，区分为东北边疆、北部边疆、西北边疆、西南边疆、海疆等类别；第二级数据库按照专题原则，区分为政治、经济、文化、地方史、宗教、民族、边防等类别；第三级数据库根据断代原则，区分为古代、近代、现代等类别；第四级数据库基于疆域理论要求，区分为国家、民族国家、国民国家、边界、边疆、边境、国民、领土、领海、国际法、海洋法、十二海里领海、二百海里专属经济区等类别；第五级数据库则依照毗邻原则，对于毗邻中国边界线的邻国沿边地区的历史、文化及边界演变等予以分门别类的标示。基于上述五个类别，形成树状结构、分类明晰的数据库。

其次，数据库中的每篇文献均由专业研究人员根据数据库的分类形式，对文献做出详细的学科分类，从不同的学科角度进行解读，并将这种解读标示出来，明确标出并不作为该篇文献主题所传达的其它学科领域的隐含内容，即为每篇文献撰写

题注或评述。提炼出其中隐含层面的非主题性的其它学术领域的信息，并做出相应的学术评述，这一评述就是我们所提出「标注」的核心内容。经过「标注」的文献具有明确的指征特点，使文献的「隐性」特征得以凸显，强调同一篇文献的多学科属性。

将经过「标注」的文献资料放入相应的数据库中，这个数据库即是一个功能强大的「标注数据库」。当不同研究领域的科研人员在「标注数据库」中，通过搜索引擎，输入「关键字」查找某一方面的资料时，不仅可以找到该关键字作为主要指向的文章内容，还可以同时检索出隐含在其它学科领域以非主题形式出现的相关学术文章。从而实现在更大范围内进行检索的功效，达到提高文献检索的数据匹配能力的目的。

**3、「标注」的资料价值与学术意义**

（1）经过「标注」处理的文献资料具有较高的学术价值。因文献本身不再拘泥于原有的主题范围，「标注」作为前期的学术加工形式，使得文献资料的潜在学术价值得以淋漓尽致的体现，具有了非常强的可用性和针对性，从而在扩大检索范围的基础上极大地提高了数据检索的命中率。

（2）文献资料本身是以「源数据」状态存在的，通常的处理过程包括几个方面，即数据挖掘、数据处理、数据整合、数据分析等，此谓「数据处理链条」。而「标注」作为一种数据加工方式，是通过把握数据处理链条中的第一个环节，即「数据挖掘」，对文献数据进行深度加工、充分发掘其潜质，从而达到极大地提高数据的利用率的目的。

（3）无论是对数据库进行分类还是将文献资源数据化，无不是利用了文献资料这一数据资源的自然形态，所不同的是「标注」更关注的是文献资料的内容及内涵，即在「挖掘」二字上下功夫。通过「标注」形成「标注数据库」，使传统意义上的资料数据库的应用水平得到极大提升。形成具有边疆史地研究特色的、适合边疆史地研究工作者需求的公共数字资源平台，发挥数据库的强大优势。

（4）对文献资料的「标注」过程还是一个学术积累的过程。凡使用过该文献的人都可以从某一学科角度出发提出新的评述，形成新的「标注」，以对「标注数据库」进行补充。在良性互动循环不断持续的状态下，「标注」的积累日益增多，文献的可

用性得以大大增强。充分实现了对数据的深度挖掘，进一步丰富了学术资料的来源。

（5）「标注」过程本身更是学术研究的过程，其特点是要由专业研究人员与网络信息人员协同工作，共同完成。「标注」的过程是利用现代信息技术对学术研究予以支持的过程，它是通过一系列可能的技术和服务来促进和改善科研环境和过程，从而达到促进科研生产力，使科研辅助由信息服务转变为知识服务的目的。

总之，从支持社会科学学术研究的角度来看，计算机技术作为一个支撑工具，其应用范围与手段，远不像自然科学和工程科学领域中那么广泛与突出，但随着时代的发展，特别是社会科学对计算机依赖程度的日益加深，改变数据库的传统模式，以使之符合社会科学研究发展的要求，已是大势所趋，且前景广阔。

## 四、小结

我们所提出的「标注」模式，是一种新的尝试，而这一尝试旨在试图实现信息技术与社会科学研究更加紧密的结合。相信随着实践的不断深入，「标注」的实际效用将得到最大限度的展现。

我们期待信息技术能更加广泛地应用于社会科学研究，这也是国内外众多学者的共识。在英国，e-science 计划之下有专门的 e-social science 分支，在美国的 cyber infrastructure 计划中也专门有对社会科学研究的安排。在中国，自然科学基金委已开始支持相关的项目，正在构思一个「实证社会科学网格」。

由信息技术带动的历史潮流为社会科学家们更细微、全面、实时地理解和认知历史和社会本身带来了「四两拨千斤」的支点和杠杆，这也是社会科学发展所面临的新的机遇。对这种机遇的及时把握，不仅有望给社会科学带来突飞猛进的发展，也预示着信息技术的应用与进步有了一个激动人心的新方向。

在可预见的未来，信息技术和社会科学的相互促进会是一种日益重要的发展方向。插上信息技术的翅膀，社会科学研究将展现出一种新的面貌；应用于社会科学研究，信息技术将迎来一个新的、生机勃勃的春天。

【作者简介】

白妍　女

职　　称：社科院 中国边疆史地研究中心 高级工程师

职　　务：社科院 中国边疆史地研究中心「网络资料室」主任

研究领域：网络信息

个人简介：**学习经历：**

1980～1984 年，就读于北京航空学院机械设计及制造专业

1984～1985 年，在航空工业部 625 所计算机软件中心学习

1987.8.～1987.10.，在美国 GE／calma 公司培训中心接受 CAD／CAE／CAM 系统培训

**工作经历：**

1984～2004 年，在航空工业部工作

2004 年～至今，在社科院中国边疆史地研究中心工作

**研究成果：**

1. 在航空工业部工作期间，所参加课题《CAMAUX—F 加工中心自动编程系统》，获 1989 年机电工业部科技进步三等奖。本人在其中主要负责该系统「输入翻译」部分的编程及调试工作。
2. 在航空工业部工作期间，曾承担国家科委「星火计划」研究项目—《CAD／CAM 技术在大型复杂模具中的开发和应用》，获 1992 年航空工业部科技进步三等奖。本人在其中承担模具的 CAD 设计部分。

联络电话：86-13520419394；86-65134986

联络邮箱：baiy@cass.org.cn