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

中国社会科学院中国边疆史地研究中心

目　　录

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。

研究背景

信息技术正在改变着学术研究的手段和方法。由于网络的兴起，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，搜索引擎等各种新型工具，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。

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。

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。

中国社会科学院中国边疆史地研究中心成立于1983年。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：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。

边疆中心信息化的现状与设想

边疆中心研究领域：

- 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）
- 海疆（黄海、东海与南海，以及台湾、海南两省）
- 邻国的毗邻地区等。

边疆中心先后主持、参与的国家级项目：

- 《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》
- 即将开展《北部边疆历史与现状研究》项目

边疆中心的研究方向：

- 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；
- 开展当代边疆地区的稳定和发展方面的研究；
- 构建中国边疆学的理论框架。

从上述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。

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。

海量的数字化、网络化信息带给我们方便，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、快捷地提炼出科学研究所需要的专业信息的方法，或许是解决上述问题的有效途径。这个途径就是我们将引入的“标注”方法。

“标注”的提出

“标注”是我们从“语音”研究领域引入的一个概念。

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。关键词是用于表达文献主题内容的词句或短语，标引的是单一概念。

“标注”的提出和应用

这样处理的结果有利也有弊。“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。

例如：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。

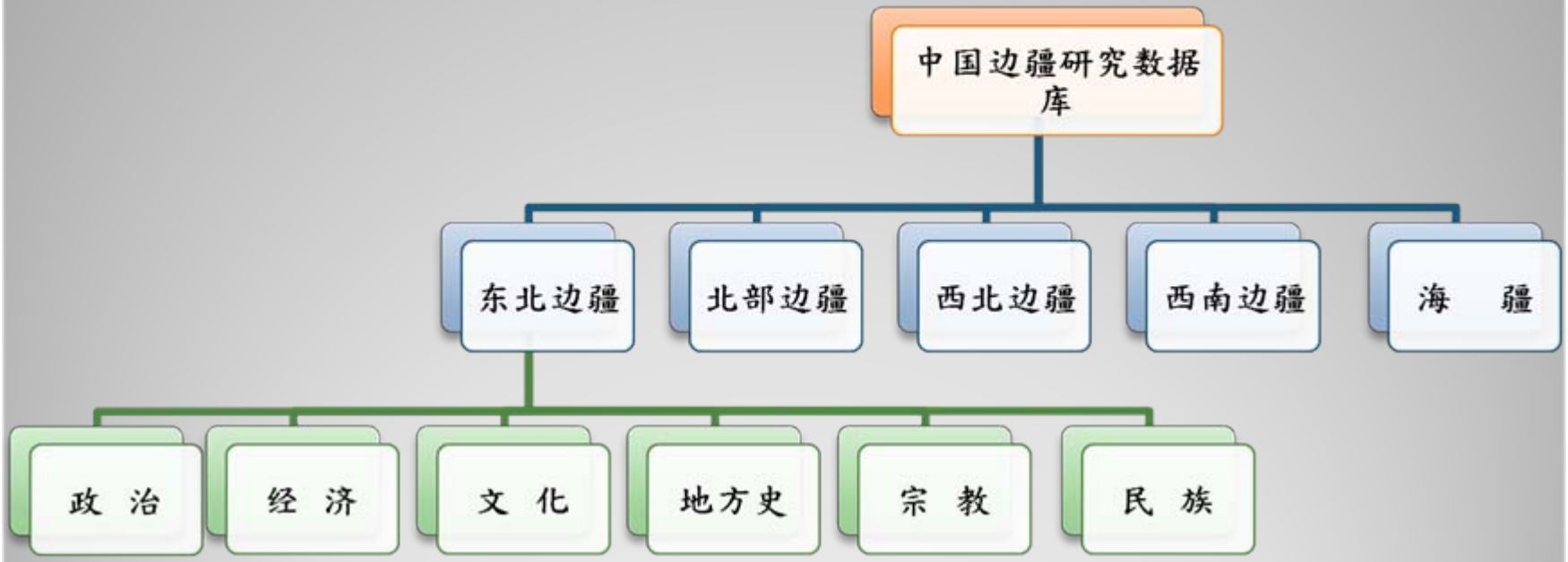
而这些信息对于相应研究领域的学者来说都是非常宝贵的学术资料。如果仅用几个“关键词”来描述这篇文章的话，则很难将所有的信息涵盖其中。也就是说，“关键词”所表述的只是这篇文章直接指向的内容，而其它层面的信息则被忽略了。

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（即：检索命中率偏低），其学术价值也自然地被“人为”地降低。

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。

“标注”的应用

对数据库进行严格细致的分类



“标注”的提出和应用

“标注”的核心内容:

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的“标注”的核心内容。

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。

“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。

“标注”的提出和应用

2. 文献资料本身是以“源数据”状态存在的，通常的数据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。

我们提出的“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。

结　语

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。

感谢各位 敬请指正