走进数字化——

中国历史地理信息系统（CHGIS）1820 年数据简介

复旦大学历史地理研究中心　　满志敏

一，中国历史地理信息系统包括的内容

历史地理信息包罗万象，有自然的，如气候、地貌、灾害动植物等；也有人文的，如政区疆域、聚落、经济、人口、文化社会等。但其中有一些是最基础的信息。从有关地理内容的表达来看，无非有两个基础的因素，即空间位置和要素内容（当然要素中又可以延伸出许多相关的类型）。但表达地理信息在地球表面位置的科学方法经纬度系统是近代科学的产物，在中国历史上的大部分时间中还没有这个系统和完善的概念，实际上历史文献记载中的地理信息是通过另一个系统来表达空间位置的，即众所周知的地名点和区域（行政的或自然的），如某地发生水灾、某地有多少人口、某地有什么样的社会文化习俗，或某现象在某山某河之阳等等。这个大部分以人文要素标记的地点和地区实际上起着与经纬度相似的作用，用以标记地理要素所属的空间位置。因此这种用以标记其他地理信息的空间位置可以称为基础历史地理信息。

当然这个基础信息的局限和缺陷是显而易见的，如相对性、名称随时间变化等等。但这是一个事实，历史地理研究所依赖的历史文献信息就是这样表述的，我们必须面对这个问题。谭其骧先生主编的《中国历史地图集》出版历史地理意义重大，如果从信息角度来看，是把历史上传统的空间信息描述方法和内容，转移到以现代地理坐标为基础的现代地图上，大大方便了阅读和研究历史地理信息，提高了历史地理信息空间位置的准确性。以计算机技术为基础的中国历史地理信息系统从基本目的上来看，是传承了《中国历史地图集》的主要目的，也是把基础历史地理信息标定到现代空间位置基础上。但 CHGIS 也提供了纸面历史地图所不具有的功能：

1)，历史地理信息的连续变化，CHGIS 数据不是描述一个或多个时间截面的空间信息，而是描述这些信息在时间上的连续变化。

2)，空间信息分布与文字属性信息的有机结合。

从具体内容来看，CHGIS 的基本功能是用地理信息系统技术编制基础历史地理信息，但同时 CHGIS 数据也应当承担普通历史地图的功能。也就是它在为历史地理或其他方面研究中提供基础数据外，其本身的内容需要有相当的可读性，无论是图面内容的表达还是相关数据的阅读。

考虑到上述的要求，CHGIS 数据从空间属性和文字属性来看，至少需要包括以下内容：

1)，聚落　　地名有许多类型，但其中无疑最重要的是聚落。历史上的许多信息都与聚落有关，同时它也是其他历史地理信息的基本空间信息点。

2)，行政区域和疆域　不同等级的行政区域也是基础地理信息，并且表达了不同行政区域在空间上的位置关系，以及统辖关系。

3)，自然地理要素　　包括海岸线、河流、湖泊、山脉、山峰、地形等。

图 1 是将来 CHGIS 数据应该达到的图面表达样式。

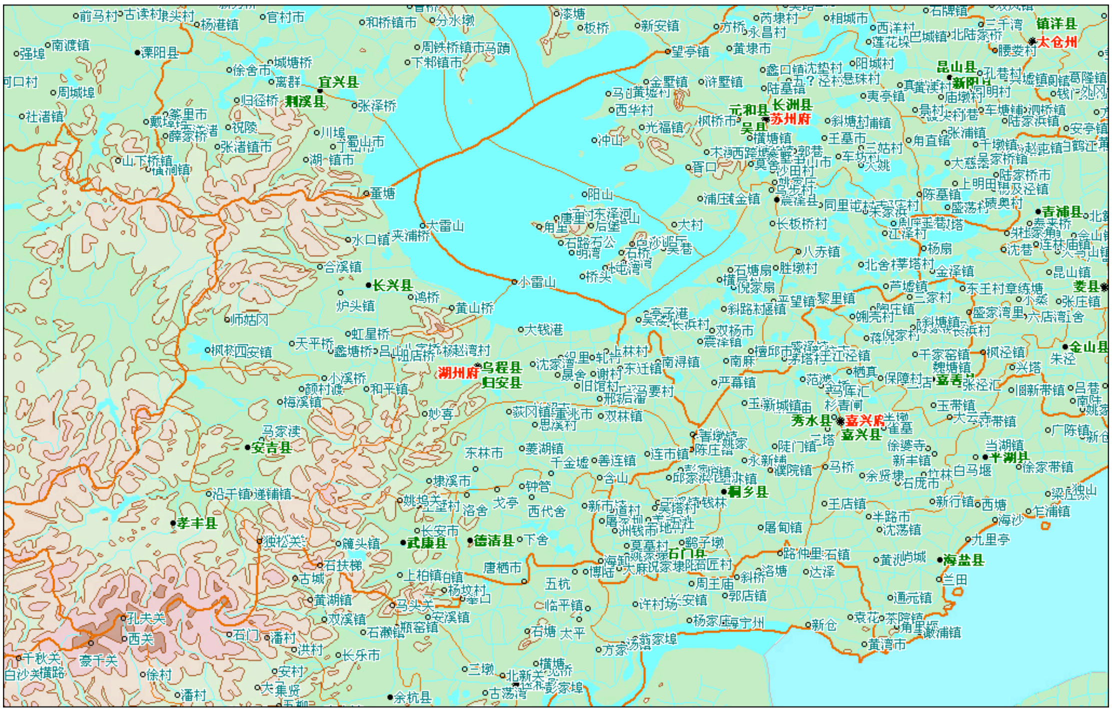

图 1，CHGIS 数据的图面样式

二，目前已经完成的内容

按照 CHGIS 项目最初的计划，从 2001 年 1-8 月为编辑演示数据阶段，在这个阶段主要集中在 1820 年这个时间截面上。主要完成的内容包括数据库结构的研究和改善、数据编写的格式和要求、工作流程和概念，以及用于体现上述要求的工作数据（以 1820 年时间截面为主，以下简称 1820 年数据）。

		聚落点	点释文	政区区域	区域释文
太湖地区	村镇级	956	有		
	县级	70	有	71	有
	府级	10	有	10	
	合计	1036		81	
松江府地区	村镇级	413	有		
	县以上级	122	有	122	有
	合计	535		122	
其他地区	村镇级	7790			
	县级	1725			
	府级	296		304	
	省级	21		26	
	合计	9832		330	
总计		11403		533	

表 1，中国历史地理信息系统 1820 年演示数据内容统计

1820 年数据以谭其骧先生主编的《中国历史地理集》（以下简称谭图）疆域为界，根据研究人员的数量和编制数据需要的时间，我们把 1820 年数据分为三个区域（图 2），不同区域的数据实现不同的演示目标。目前已经完成的内容参见表 1 的统计。

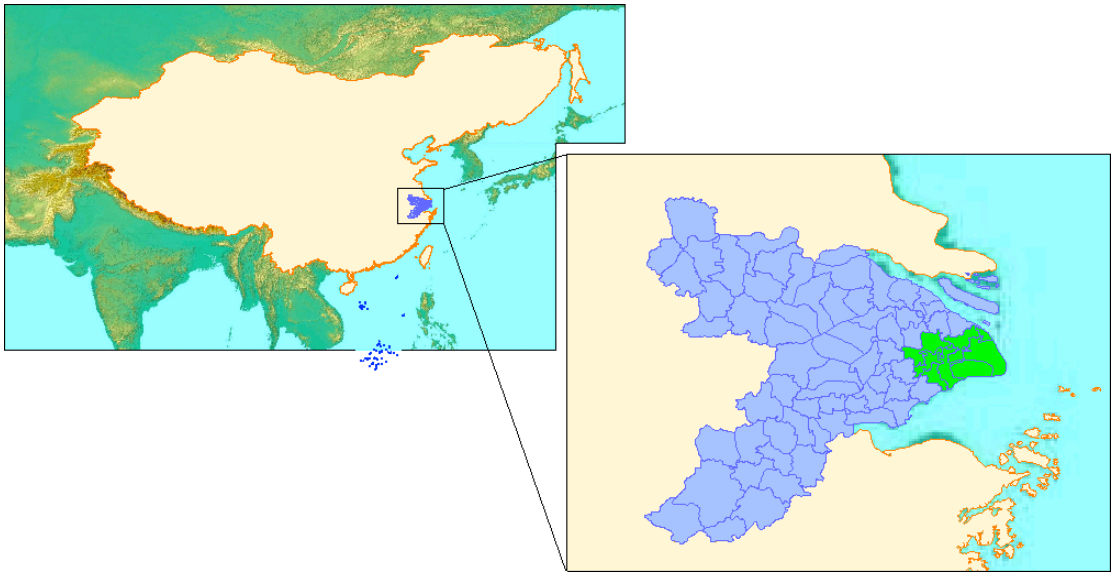

图 2，1820 年三个不同区域数据的空间关系

三，太湖地区数据

江宁、镇江、常州、苏州、太仓州、松江府、湖州、嘉兴、杭州、严州等十府州共有 1036 个聚落点和 71 个县界（另有 5 个府已经完成数据编写，但未入库）。全部数据是原始文献记载中重新编辑，每一个聚落点和县界都有相应的释文，下面其中是其中的两个例子：

龙窟镇　（村镇聚落）

乾隆《武进县志》卷二驿站，北塘九里铺，南接府总铺，北接龙窟铺，计程俱十里，又北塘龙窟铺，在龙窟镇，北接江阴火庄铺，计程十里，可见龙窟镇在县北二十里。光绪《武进阳湖合志》卷三水道龙窟荡，俗名龙虎塘，今有龙虎塘，在武进县北近二十里，接江阴县界，符合上述记载，当即清之龙窟镇。

武进县　（县界）

东界南段，据光绪《武进阳湖合志》（参照康熙《常州府志》卷五）卷二武进怀德南乡，南接阳湖延政乡界，地名有陈渡桥、慈墅等，阳湖延政乡地名有游塘（牛塘）、塘口、韦庄，又卷五丫河铺，地属阳湖，铺属武进，上述地名今犹见在参照光绪《江苏全省舆图》武进县图，陈渡桥、慈墅以南，大致武宜运河至滆湖段稍西，包括今韦庄，此西属武进，东属阳湖。北段武进德泽乡，东南至阳湖县永丰西乡为界，地名有三井头，低头田，阳湖永西乡地名有洋头、青龙桥，上述地名今犹存在，参照《江苏全省舆图》，大致以澡港运河为界。

南界见宜兴县界条北界。

西界《嘉庆一统志·常州府·关隘》云，武进县奔牛镇二十里至镇江府丹阳县吕城镇。光绪《武进阳湖合志》卷一武进县西抵九里铺镇江府丹徒（阳）县界，此段界线与今相同。南段与金坛县界、北段与丹阳县界，据地名及参照光绪《江苏全省舆图》与今界相同。但寺城村图在武进县界，今在丹阳县界，合志鸣凤乡地名有寺城上，原当属武进。

北界长江，道光二十年《江阴县志》卷首孟河口以西有武进县陆沙，江中大沙分属武进、丹阳，光绪《武进阳湖合志》卷六永生洲，峙立江心，分属常州、镇江、扬州三府，无所专统，今丹阳与武进交界处有永胜洲地名（属丹阳），此段界线大致以今孟城北至扬中县一线（清末设太平厅，民国初改扬中县）。与江阴县界据上引县志以桃在港为界与今同。

记录表的内容如图 3 所示，需要说明的是，该表是 CHGIS 工作数据表，不是最后数据发布样

式。

key_id	charact	ch_uni	beg	en	known	X_coord	Y_coord	place_present	dyn_c	rej	prov_ch	level2_	le	le	le	cnty_c	uncnty_	note_i	spatial_	compile	checker
16722	鑫塘桥	村镇			1820	119.8327484	30.8948765	今浙江长兴县南里塘	清朝		浙江省	湖州府				长兴县	鑫塘桥	16722	FROM_FD	邹逸麟	傅林祥
16723	天平桥	村镇			1820	119.7627487	30.9078751	今浙江长兴县西南天平桥	清朝		浙江省	湖州府				长兴县	天平桥	16723	FROM_FD	邹逸麟	傅林祥
16724	夹浦桥	村镇			1820	119.9358673	31.1031132	今浙江长兴县北夹浦镇	清朝		浙江省	湖州府				长兴县	夹浦桥	16724	FROM_AC	邹逸麟	傅林祥
16725	吕山	村镇			1820	119.9127502	30.8938751	今浙江长兴县东南吕山	清朝		浙江省	湖州府				长兴县	吕山	16725	FROM_FD	邹逸麟	傅林祥
16726	小溪桥	村镇			1820	119.8211899	30.8453274	今浙江长兴县南小溪口	清朝		浙江省	湖州府				长兴县	小溪桥	16726	FROM_AC	邹逸麟	傅林祥
16727	师姑冈	村镇			1820	119.6467514	30.9498768	今浙江长兴县西南师姑冈	清朝		浙江省	湖州府				长兴县	师姑冈	16727	FROM_FD	邹逸麟	傅林祥
16728	枫树冈	村镇			1820	119.6127472	30.8958759	今浙江长兴县西南仙山（枫和	清朝		浙江省	湖州府				长兴县	枫树冈	16728	FROM_FD	邹逸麟	傅林祥
16729	鸿桥	村镇			1820	119.9767532	30.9908752	今浙江长兴县东鸿桥镇	清朝		浙江省	湖州府				长兴县	鸿桥	16729	FROM_FD	邹逸麟	傅林祥
16730	德清县	县			1820	120.0848007	30.5531330	今浙江德清县东城关镇	清朝		浙江省	湖州府				德清县		16730	FROM_AC	邹逸麟	傅林祥
16731	新市镇	村镇			1820	120.2903137	30.6166763	今浙江德清县城关镇东北新市	清朝		浙江省	湖州府				德清县	新市镇	16731	FROM_AC	邹逸麟	傅林祥
16732	唐栖市	村镇			1820	120.1811218	30.4800148	今浙江德清县东南唐棲镇	清朝		浙江省	湖州府				德清县	唐棲市	16732	FROM_AC	邹逸麟	傅林祥
16733	钟管	村镇			1820	120.1787491	30.6508751	今浙江德清县城关镇北钟管	清朝		浙江省	湖州府				德清县	钟管	16733	FROM_FD	邹逸麟	傅林祥
16734	西代舍	村镇			1820	120.1957474	30.6308765	今浙江德清县城关镇东西代舍	清朝		浙江省	湖州府				德清县	西代舍	16734	FROM_FD	邹逸麟	傅林祥
16735	下舍	村镇			1820	120.1737518	30.5558758	今浙江德清县城关镇东下舍	清朝		浙江省	湖州府				德清县	下舍	16735	FROM_FD	邹逸麟	傅林祥
16736	洛舍	村镇			1820	120.0797501	30.6358757	今浙江德清县城关镇北洛舍镇	清朝		浙江省	湖州府				德清县	洛舍	16736	FROM_FD	邹逸麟	傅林祥

图 3，工作数据记录表内容的样式

太湖地区数据中，每个研究人员分别承担分府的聚落点与县界的草图绘制，各县聚落点是研究人员用手工的方法绘制在 1：50 万地形图底图上，同时绘出县界，没有 1：50 万地形图的地区，利用 1:100 万 ArcChina 数据的打印地图作为工作底图。县界和聚落点数据必须有释文，写明资料的来源和定点定线的文献依据和判断意见。各府草图扫描后，在 MAPINFO 中注册，注册的扫描底图与 ArcChina 的 RESPT 图层叠加比较，如果 RESPT 图层上有相应的点，拷贝该点的空间位置，如果研究人员绘制的聚落点位置和地名在 RESPT 图层没有相应的点，则用 MAPINFO 的点工具，依据草图的位置直接在聚落点图层上绘制。县界的画法与此类似。这样可以保证我们 CHGIS 数据可以统一到 1:100 万 ArcChina 这个工作底图上。

太湖地区数据主要演示将来 CHGIS 数据全部完成后可以达到的精度和样式（参见图 1），从图形数据的空间精度来看，与《国家普通地图集》的精度相当。同时可以在图像显示的条件下，查询相应的表格数据和释文（图 4）。

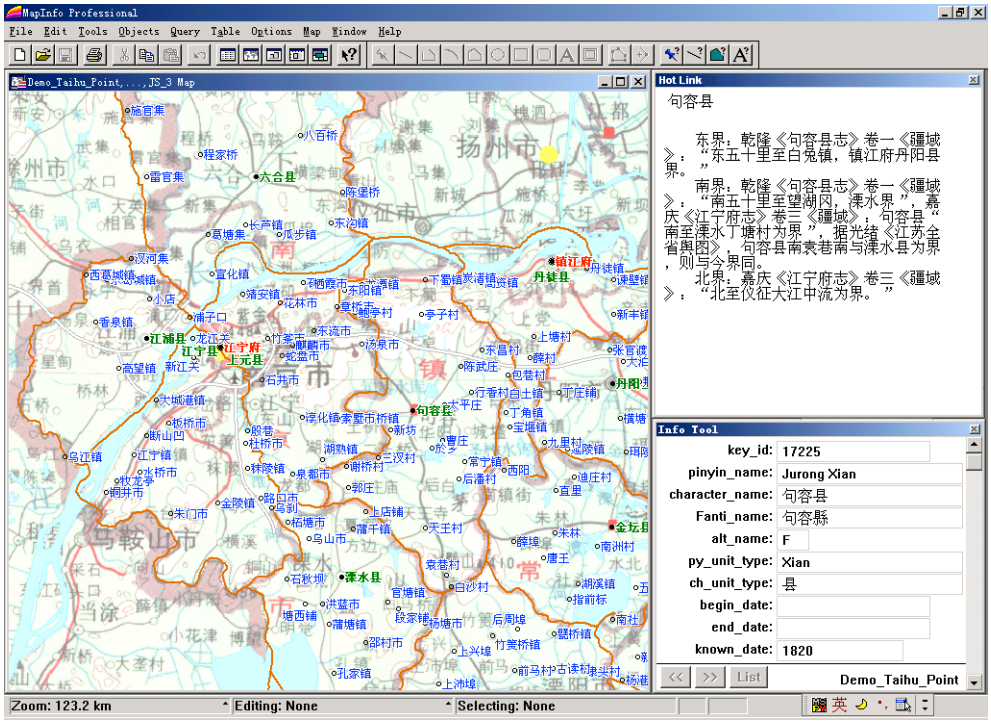

图 4，太湖地区数据与《国家普通地图集》图幅比较以及表格数据和释文查询

四，松江府地区数据

松江府地区数据包括 751 年华亭县建县以来一直到 1990 年的聚落和县界，其中 1958 年前仅包括松江府的地域，1958 年后包括上海市领县后的地区。共有聚落点 535 个和 122 个县级和县级以上的区域界线。该区域主要演示具有时间变化的历史地理信息数据编写方式、相应概念和数据查询方式。

在 GIS 数据库中，处理具有时间概念的数据时，有两种基本的方法，1，时间截面描述法，即用特定时间截面数据描述单一时间面上的空间信息。2，生存期描述法，即用每个聚落的生存期（起讫时间）描述聚落存在的时间。

对于时间截面不多的数据可以把不同时间数据放在各自的数据库中，各个独立的数据库就代表一个特定时间的地图。当时间截面很多的时候，这个方法明显受到数据量的限制，例如，一个有 200 年历史的地名数据库系统，以 10 年为单位，只要 20 个数据库，当时间间隔缩小到年的时候，数据库将扩展到 200 个，数据量将大大地增加。从基础历史地理数据的时间变化特性来看，大部分内容都不是持续变化的。一个地名（或区域）从开始出现，到更名成另一个地名（或区域），通常有一定的时间（当然设有个别变化很频繁的），在这个时间里地名的名称、行政隶属关系、空间属性稳定不变，CHGIS 数据库设计中，把这样一个地名（或区域）定义为一个记录。如果这个地名的上述属性发生任何变化，则用一个新的记录描述。我们把具有这样一个属性特征的地名（或区域）从开始到结束的时间段称为**生存期，**相应的记录称为生存期纪录。生存期是 CHGIS 数据中重要的概念，生存期记录是 CHGIS 数据库中的基本记录。历史地理信息信息的时间变化描述就是通过许多具有不同生存期的记录数据来实现的。在大量数据的情况下，生存期描述显然要比截面描述法有效的节约数据空间。图 5 是时间截面描述法与生存期描述法的比较示意图，图中的每一个矩形代表一个记录。

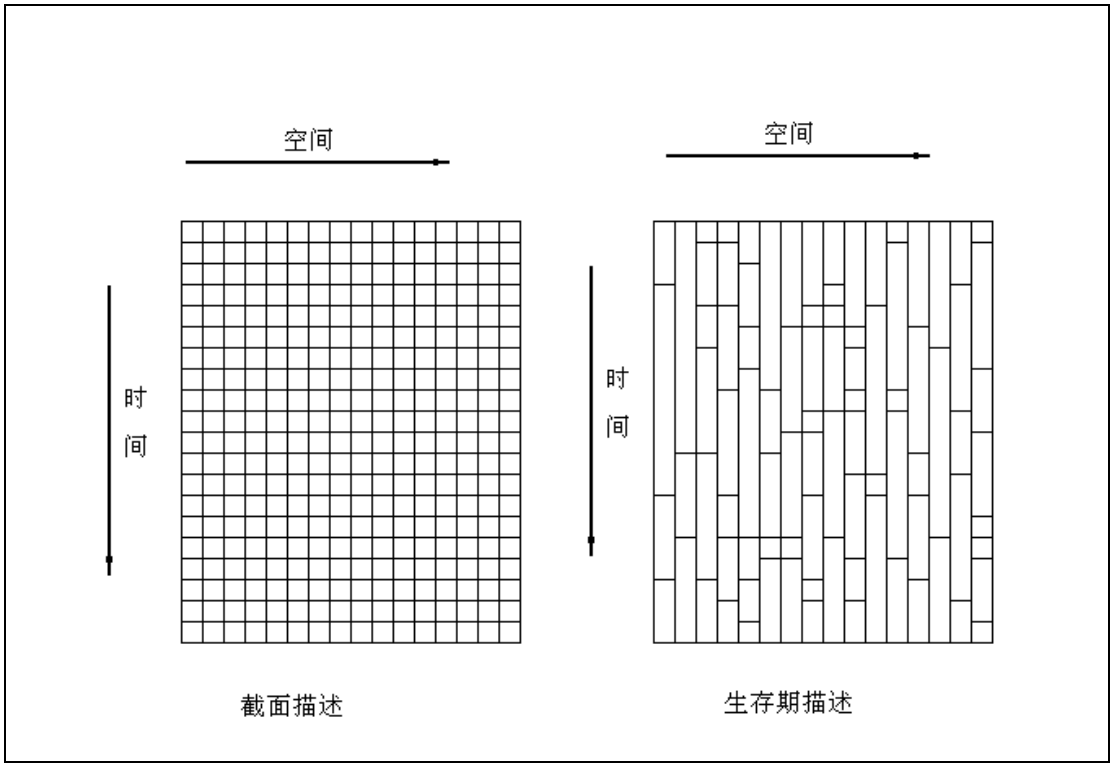

图 5，历史地理信息数据库的生存期描述法与截面描述法比较

生存期记录是数据库的基本单位，但在编制数据的时候，可操作性较差。例如一个省级单位的名称发生变化的时候，按照数据库中记录的定义，作为这个省范围的所有记录必须重新有相应的新纪录，因为表征这些地名的行政隶属关系名称发生了变化。因此在编制数据的时候，研究人员必须记住所有与上级单位变化有关的信息，实践中我们发现这是非常麻烦的，很容易遗漏。同时，隶属关系的变化常常与这个地名本身变化关系不大，如果每个生存期记录仍然有一个单独释文的话，释文的编写也很麻烦。为此我们定义了**地名存在期**的概念。地名存在期是指一个地名在其本身名称、空间位置不变的情况下，它所存在的时间（起讫时间）。相应的地名称为一个存在期地名。研究人员只要在释文中描述这个存在期地名，并按存在期地名绘出相应的柱表，只要在柱表中继承存在期地名的上级政区的变化，就可以方便地把存在期地名分割成生存期记录，而一个地名存在期内的不同生存期纪录共用一个释文。图 6 是关于存在期地名与生存期记录关系的示意图。图中蓝色短横线表示存在期地名的起讫时间，红色横短线表示继承的上级隶属关系变化的时间，横短线之间的时间就是记录生存期，对应一个生存期纪录。Life-phase Record and Existing-phase Place Name

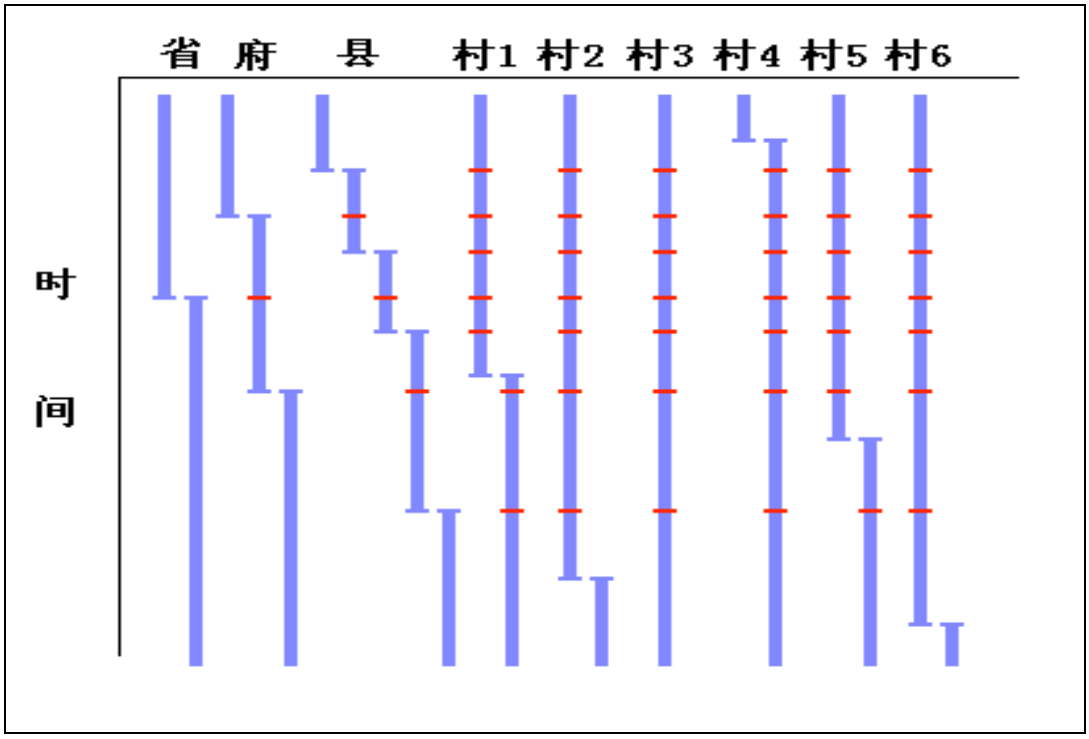

图 6，存在期地名与生存期纪录关系的示意图

生存期描述法中数据库的记录已经不存一个统一的时间截面，许多记录在空间上相互重叠，由此带来的问题是，这样的数据库不能在一般 GIS 软件中按普通的方式显示，必须用查询方法在数据库中选择满足给定时间要求的数据作为一个子集，然后显示这个子集。图 7 是松江府区域数据查询关系的示意图。我们简单的查询功能是用一个基于 MapBasic 的模块来实现的，主要的功能是在点数据和面数据这两个库内查找符合用户时间要求的纪录，并显示这些纪录，此外还有一些界面上的辅助功能，如对话，显示方式等。

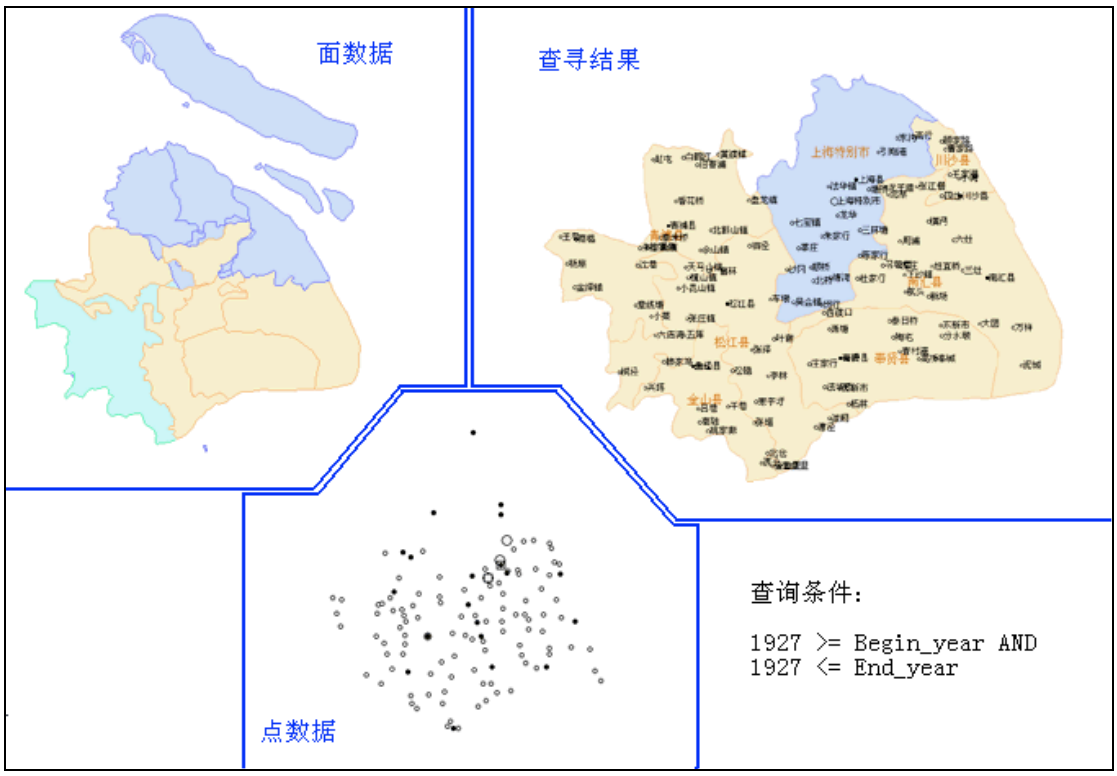

图 7 松江府区域数据查询关系的示意图。

时间属性是松江府地区数据中的重要属性，我们在数据的时间属性类中用 4 个单列的属性来描述，即一个生存期记录出现时间和结束时间，以及这个时间标量的精度和类型。出现时间和结束时间在数据库中用公历表示，仅表示到年。

之所以用年来表示，是基于这样的考虑。由于历史资料现实、以及研究水平和程度问题，实际上是不可能所有的地名纪录都有确切的时间（Beginning and Ending ），根据我们的认识，历史上记载的地名有关的起迄时间有以下几种可能:

1)，精确到大时期，如“秦汉”、“魏晋南北朝”、“宋元”、“先秦”、等

2)，精确到朝代，如“西汉”、“唐朝”、“秦初”、“清末”、“元中期”、等

3)，精确到年号或庙号，如“元始时”、“大历中”、“大中祥符末”、“至元初年”，等

4)，精确到具体年份，如“开元二十九年”、“康熙十三年”、“甲子”、“辛亥”、等

5)，精确到季和月份，如“元丰九年夏四月”、“甲子季秋”等

6)，精确到具体日期，如“1934 年 6 月 12 日”、“嘉靖十四年五月甲辰”等

而考虑到数据检索的要求和检索条件的实现，我们认为以“年”作为基本单位已经足够了。上述 1-3 的情况依据规定，估计到具体的年份，这需要研究者给出一个意见，或仅仅是一个规定。5-6 的情况略去了年以下的信息。但 1-3 个情况的所有文字信息以及规定必须保留在释文中，给用户提供一个参考，以便以后有了进一步的研究后进行修改。而 4-5 类型中有关时间的详细信息也保留在释文中，作为一种文本信息提供给用户。同时在每一个开始和结束时间的精度属性上注明相应的类型标记（用 1-6 表示），如果生存期纪录的是从上级隶属关系中继承的（参见图 6），则不标注精度类型，仅用所继承上级单位等级的第一个拼音字母表示（C、Z、S、F、X 等），这个标记实际上也是一种资料信度信息的表达，可以方便用户很快确定数据的精确程度，并快速地找到相同的精度类型。这样处理方法，一方面考虑了现实查询的需求，对一个涉及到数千年基础历史地理信息的数据库来说，能够直接查到具体的

年份，是能够满足绝大部分需要，另一方面也考虑到数据库的编写和查询机制上比较方便实现。

五，其他地区数据

除了太湖地区和松江府地区的数据外，1820 年清代疆域范围内其他地区的数据共有 9832 聚落点，330 个府级和省级的区域。数据的分布如图 8 所示。

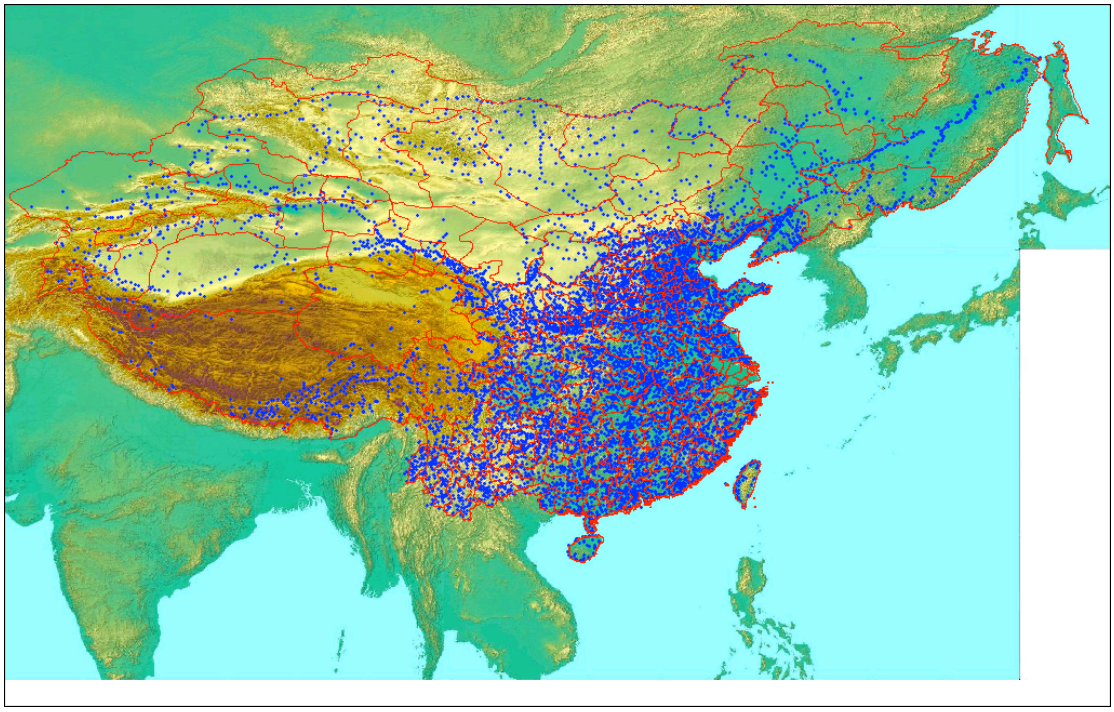

图 8，除太湖地区外的 1820 年数据分布

该部分的数据来源于谭图清朝卷。但在聚落点的空间位置定位校正上，分为两个部分：第一部分，如果在 ArcChina RESPT 图层上有相应的数据，依据 ArcChina 的数据位置，如果没有相应的数据，目前暂时直接用谭图的位置。这两部分数据分别在 Spatial_note 中用 FROM_AC 和 FROM_TAN 分别表识。

谭图绘制时所用的底图主要是 50-60 年代的 1:200 万地形图，但目前留存的资料没有说明这些地图的投影，并且在谭图中也没有作任何投影数据的说明。根据当年研究人员回忆和我们的测试，找到谭图所用的投影是区域等角投影（Regional Conformal Projection），不同纬度的各省分别使用两种投影参数。其中：黑龙江图幅，吉林图幅，新疆图幅，乌里雅苏台图幅，内蒙古图幅使用 Lambert Conformal Conic Europe parallels (42 56)投影，参数为 3, 28, 7, 17, 29.77930555, 42, 56, 2679984.29, -484330。其它各省图幅使用 Regional Conformal Projections (China)投影，参数为 3, 0, 0, 110, 10, 25, 40, 0, 0。使用这些参数可以使扫描后的谭图与标准经纬线很好地重合。

扫描后的谭图中清代各省图幅，经上述投影注册后与 ArcChina RESPT 数据比较，可以发现部分地区相同聚落的差异较大，进一步比较分析（主要是利用河流比较），可以发现谭图在西部地区的局部误差较大，直接影响到聚落点的定位。图 9 是两个局部的例子。从这个例子可以看出，CHGIS 数据如果需要统一到 1:100 ArcChina 数据基础上，就不能直接从谭图上数值化，而需要对点的位置作一定的调整。

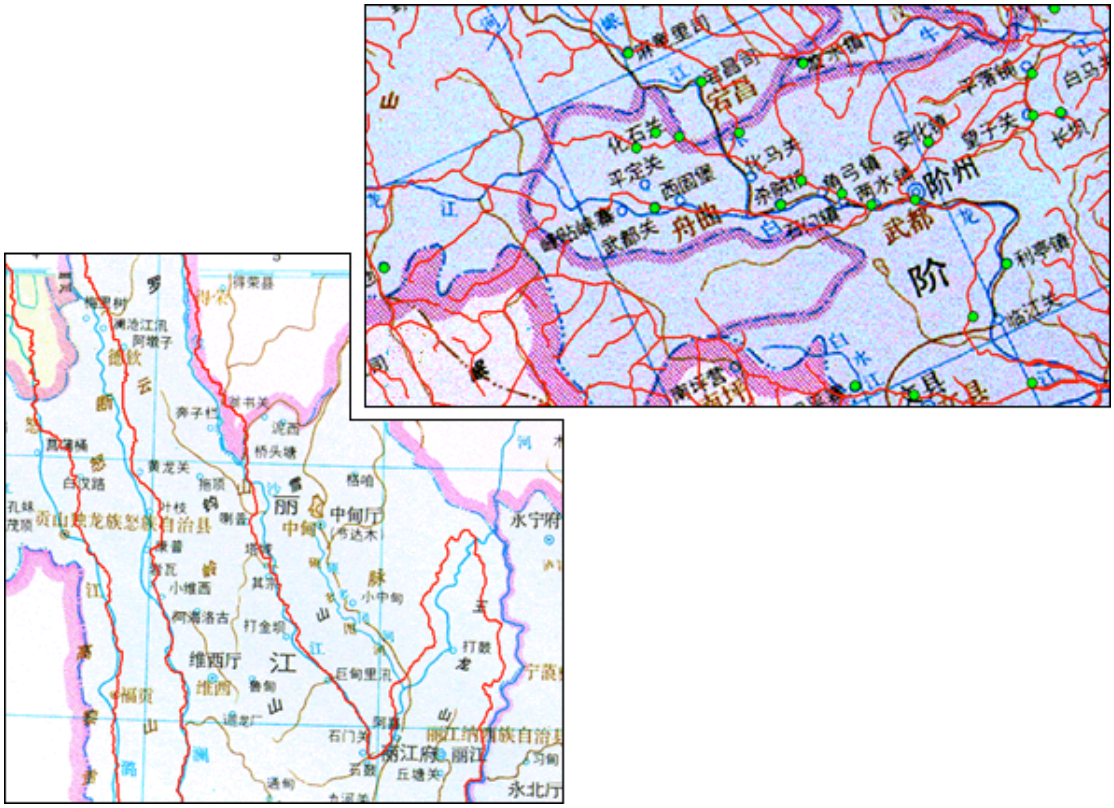

图 9，谭图图幅中河流与 ArcChina 河流的比较

六，其他几个有关数据的问题

1，简体与繁体汉字的转换

CHGIS 数据的工作语言是简体汉字，由于需要在数据库中保留繁体地名，这涉及到简繁体汉字的转换问题，但这个转换是在《CJK 统一汉字编码字符集》中的转换，不是 GB 向 BIS5 的转换。一般简体向繁体的转换较容易，目前有许多现成的工具可以使用，如 Windows2000 所带的代码转换工具就是一个不错的工具。但简体与繁体汉字之间存在一对多的关系，如简体“云”不能简单地全部转换为“雲”。这需要在转换后，用人工的办法检查修改。

2， 汉字与拼音的转换

当然最原始的办法是逐个用手工的办法输入，但效率低下，且容易出错。批量转换是在处理大量数据时应该考虑的。但目前还没有见到有效的转换工具。在南极星汉字平台中带有一个汉字与拼音的转换工具，可以它只支持 GB2310-80 标准，也就是它只能处理常用的 6763 个汉字。不适合 CHGIS 数据处理，因此需要自己编制合适的转换工具。简单的转换工具可以借用文字处理软件的替换功能，也就是说对特定的汉字，用相应的拼音来替换。这样的工具有许多，其中比较理想的是 Office 套件中的 Word。替换功能是 Word 中常用的工具，只要把需要输入替换的字或词以及替换后的字或词，执行后就可以自动完成通篇文字的查找和替换。这是一个基本的方法，但显然逐字替换的效率低下，我们利用 Word 所支持的 Visual Basic Application 来编制“宏”，利用“宏”自动执行所有的替换，替换效率很高。只是在编制“宏”的时候需要作基本材料的准备工作，需要有汉字与对应的拼音，这个材料可以从 Windows 中文版的“全拼输入法”的码表中找到。

3，罕见汉字的处理方法

在计算机有关处理中文汉字的标准目前有三个：1，《信息交换用汉字编码字符集 》（GB2312-80），实现了汉字窄计算机上的运用，该标准简称 GB 码，包括 7445 个图形字符，其中 6763 个是简化汉字。2，《CJK 统一汉字编码字符集》（ GB13000.1）完全等同于国际标准《通用多八位编码字符集 (UCS)》ISO10646.1，在中国常常简称为 GBK。GBK 中最重要的也经常被采用的是其双字节形式的基本多文种平面。在这 65536 个码位的空间中，定义了几乎所有国家或地区的语言文字和符号。对东亚文字最重要的是，在 0x4E00 到 0x9FA5 的连续区域包含了 20902 个来自中国（包括台湾）、日本、韩国的汉字，称为 CJK (Chinese Japanese Korean) 汉字。CJK 是《GB2312-80》、《GB12345-90》、《BIG5》等字符集标准的超集。但实际上仍然有相当部分的汉字不能包括在上述两个标准中，尤其是人名、地名和古文献中的一些汉字不在上述标准之内，也就是说在计算机上无法输入和显示这些汉字。2000 年 3 月，国家颁布 GB18030 标准，建议支持的汉字增加到 27564 个[1]。国际标准化组织（ISO）在 ISO/IEC 10646-1:2000 的基本平面（简称 Unicode）编入了 GB18030 建议的汉字，其中超出 GBK 部分的 6582 个汉字，又称为扩展 A。同时 ISO 还在 ISO/IEC 10646-1:2000 第二平面上又扩展了 42711 汉字，也称为扩展 B。[2]由于这 42,711 汉字编排在 ISO10646-2000 的第二平面，所以编码需要 4 个字节。为了能够存取处理这些 4 字节字符，在 Unicode 中引入了 Surrogate 机制(在 ISO10646-2000 中命名为 UTF-16)。根据这样一种机制，在 Unicode 中用两个 16 位编码就可以对扩展 B 中的汉字进行存取。也就是说目前从标准上来说已经解决了 60275 个汉字在中文平台上的操作。

但有了标准仅仅是第一步，要真正实现6万余个汉字方便地运用，还需要其他几个条件：1，计算机操作系统支持上述汉字编码机制；2，相应的字符集字库；3，相应的汉字输入法；4，应用软件支持上述编码机制。目前计算机操作系统 Windows2000开始对这样一种机制提供支持(Windows 98, Windows ME 及 Windows NT4 中没有)。在这个版本中对扩展A的汉字支持很好，汉字基本字库（宋体）已经包括了27564个汉字，并在部分应用程序中能很好的显示（在字符映射表不能显示扩展A的汉字），但缺少相应的输入法。图10是《中国历史地图集》中位于扩展A的部分汉字。

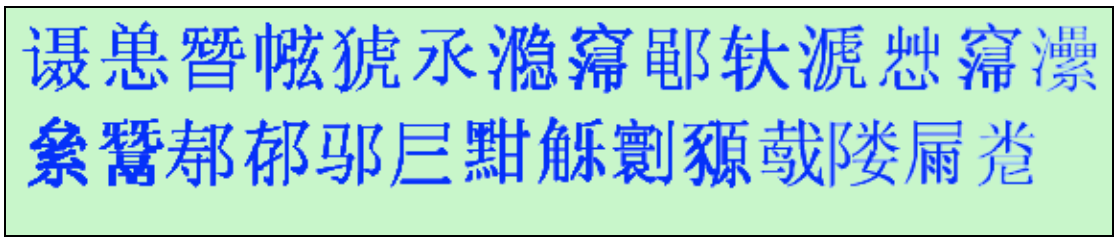

图10，部分位于扩展A中的汉字

最近微软公司发布的中文简体版Office2002（又称Office XP）及相应的多语言包中专门开发了包括全部6万余个汉字的超大字符集中文字体(宋体-方正超大字符集)，以及用于输入这些汉字的增强型区位码，该超大字符集包括了GB18030标准支持的全部27564个汉字，以及在第二平面(扩展B，42711个汉字)中选出的36,862个在中国大陆，香港特别行政区(以及部分台湾地区)使用的汉字。因此包括西文等常用字符在内，宋体-方正超大字符集共包括了65531个字符。但由于一般Windows应用程序并不支持Surrogate 机制，因此扩展B内的汉字除了在Office2002系列应用程序中可以使用外，在其他GIS应用程序中并不能支持。

[1] http://www.chinagb.org
[2] http://www.iso.ch

考虑到目前的技术条件，因此我们认为，在CHGIS数据中不能包括扩展B中的汉字，显然这些汉字除了在Access数据库中运用外，不能在GIS平台上正确显示。同时考虑到目前还有许多用户的操作系统仍然是Windows98/Me，扩展A中的汉字也不适宜用在数据中，因为如果数据中包括这些汉字后，必须规定用户在使用CHGIS数据时一定要使用Windows2000中文版，否则不能显示这些汉字。显然这个要求对用户太高了，同时GBK加上扩展A的27564个汉字也未必一定能覆盖历史地名中所有汉字，势必还需要使用其他的辅助手段。因此我们在编入数据中的汉字仅以windows/Me所支持的为限，在地名中用代码表示相应的汉字，加括号，并建立相应的文字代码。同时编制一对一文字数据表，建立文字代码与图形文字的对应关系，图11是地名表、文字数据表的格式和相应的关系。待以后汉字平台能真正支持ISO/IEC 10646-1:2000标准后，再把数据中的代码换回相应的汉字。

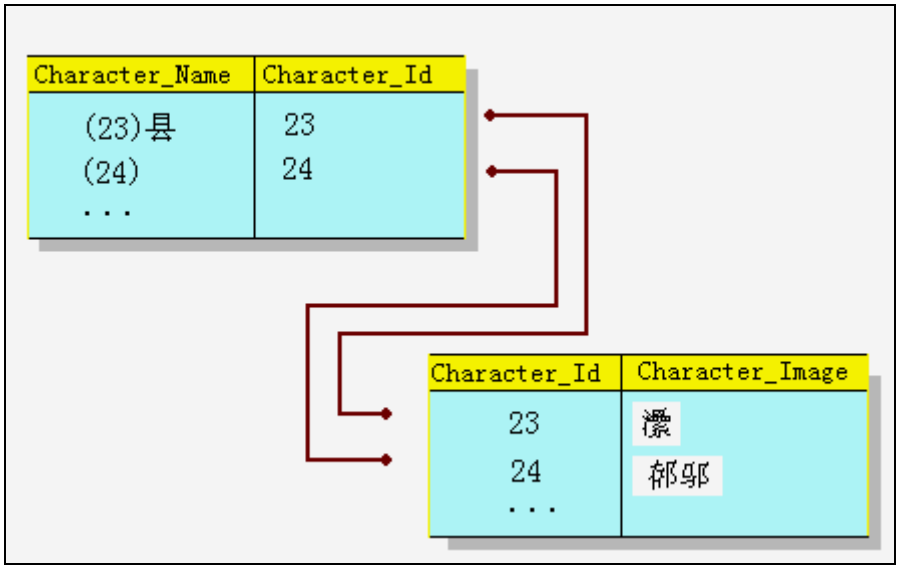

图11，CHGIS数据中罕见汉字的代码方案

聚落是最基本的信息（其它内容与此相似），从空间上来看，它是一个点（一定比例尺的条件下），表征这个点的属性数据有许多，但按照上述整个数据库的需要达到的效果和要

求来分析，可以包括：名称、行政隶属关系、空间属性、时间属性、数据库属性、注释等 6 大类。

名称　　任何聚落都有名称，但考虑到整个 CHGIS 数据库需要向全世界开放，在表达上应该包括三种类型，即简体、繁体和拼音。此外，不少地名有不同的俗称或异名，而不仅仅是一个官方或较正式的名称，因此俗称或异名也构成了这个聚落在名称上的一种属性。同时，异名有时是用其他语言系统表达的，这个语言系统是什么？也需要注明。

行政隶属关系　　在国家社会中，聚落，无论是村镇、县治或其他，都受一定的行政管理体系的管辖和节制。以水口镇为例，属清朝、浙江省、湖州府、长兴县，以上的朝代、省级单位、府级单位和县级单位构成了水口镇在行政隶属关系上的几个属性。从上述水口镇的例子来看，我们只要有 5 个属性就可以描述这个聚落行政关系，但由于数据库不光是描述清代的地名，它的结构还必须容纳其他朝代与上述结构关系不同的行政体系。因此行政关系属性需要扩展。我们目前在 CHGIS 数据库中规定的行政关系属性一共有 9 个，朝代、政权、省级、二级、三级、四级、五级、县级、县以下，以保证任何一个聚落的行政关系属性有相应的属性描述位置。表 1 是几个地名的行政关系属性和描述。需要指出的是，CHGIS 数据的目的并不是编制完善的行政区划和隶属关系，有些行政关系在数据库中需要加以简化或省略，如县以下的聚落仅归并为一类，而边地的一些特殊内容也作了简化。完善的行政区划和隶属关系应该有专题数据来描述。此外，从聚落所处的行政地位可以对聚落的性质进行分类，但分类体系需要考虑分类的使用范围和用处。如县以下的地名有许多种，镇、市、集、圩等都有一定内涵的差异，可以作为分类的依据。但从数据查询的要求来看，分类不能太复杂，专门的分类应该是专题内容的工作，CHGIS 数据基础目的是提供基础历史地理信息，分类主要考虑数据检索的需要。

名称	朝代	政权	省级	2 级	3 级	4 级	5 级	县级	县以下	类型
水口镇	清朝		浙江	湖州府				长兴县	水口镇	村镇
武进县	清朝		江苏	常州府				武进县		县
江宁府	清朝		江苏	江宁府						府

表 1，行政隶属关系属性的几个例子

空间属性　　空间属性包括点的经纬度（如水口镇，经度 119.865753，纬度 31.093875）、今地（水口镇，今浙江长兴县北水口镇）、位置来源，定点依据。尽管在地理信息系统中，基本的空间属性已经由点的位置所决定，但这还不够，考虑到 CHGIS 数据不光是在我们工作中所用的 GIS 系统表现，我们还希望用户可以利用这些数据在其它 GIS 中运用，直接用数字方式描述聚落的位置，可以方便的其他 GIS 系统中生成相应的空间点的位置。今地这个属性对聚落的描述并不是必要的，但在编制数据库时，可以方便地帮助找到该聚落在底图上的相应位置。位置来源用于描述任何一个聚落在技术上是如何确定在目前位置上，这个信息是一个技术基础，一方面表明了位置的数据是从何来的，同时也描述了如何来的，可以为以后修改数据的人员掌握数据的技术基础提供帮助，并发现可能存在的问题。以下是一个描述的例子：

FROM_FD 用于表识聚落点 （Point) 的来源是依据研究人员绘制的分府草图。

各县聚落点是研究人员用手工的方法绘制在 1：50 万地形图底图上，没有 1：

50 万地形图的地区，利用 ARCCHINA 的打印地图作为工作底图。县界画法的资料

依据，参见各县聚落点数据的 Note 中的说明。每个研究人员分别承担分府的聚落点与县界的草图绘制。

ARCCHINA 是 ARCINPO 格式，在 ARCVIEW3.0a 中分别读入各分幅 RESPT 层的数据，并输出为 SHAPE 格式保存。在 MAPINFO 的 Universal Translator 中把各分幅的 SHAPE 文件转换为 TAB 格式文件。转换的投影参数为：

"Longitude / Latitude (Pulkovo 1942)\p4284", 1, 1001

在 MAPINFO 中把各分幅的 TAB 文件合并为一个 RESPT 文件。

各府草图扫描后，在 MAPINFO 中注册。

1：50 万地形图的投影参数如下：

"--- Gauss-Kruger (Pulkovo 1942) ---"

"GK Zone 21 (Pulkovo 1942)\p28421",8,1001,7,123,0, 1,21500000,0

ARCCHINA 来源的底图注册投影参数见上。

注册的扫描底图与 RESPT 图层叠加比较，如果研究人员绘制的聚落点位置和地名与 RESPT 图层没有相应的点，则用 MAPINFO 的点工具，依据草图的位置直接在聚落点图层商绘制，该图层的投影设置与 RESPT 图层相同。并在记录中表识为 FROM_FD 。

定点依据实际上是地名释文的一部分内容，数据库中的任一聚落，必须有相应的文献依据和相应的判断意见，其中重要的是如何定位。只有描述了这些信息，才能正确地把聚落确定到经纬度系统中，形成电子地图。以下也是一个例子。

龙窟镇

乾隆《武进县志》卷二驿站，北塘九里铺，南接府总铺，北接龙窟铺，计程俱十里，又北塘龙窟铺，在龙窟镇，北接江阴火庄铺，计程十里，可见龙窟镇在县北二十里。光绪《武进阳湖合志》卷三水道龙窟荡，俗名龙虎塘，今有龙虎塘，在武进县北近二十里，接江阴县界，符合上述记载，当即清之龙窟镇。

时间属性　时间属性是涉及到历史地理信息的一个重要的属性，因为 GIS 数据中引入了“时间坐标”的概念，也使得 GIS 数据可以用来描述地理信息的历史变化过程，相关概念已经有了较多的讨论[3]。但如何在属性数据上对一个历史地理对象的描述，可以有多种方法，我们将在下面专门讨论有关思路和概念，这里仅给出我们在 CHGIS 数据运用的方法和规定。时间属性类用 4 个单列的属性描述，即一个独立对象记录出现时间和结束时间，以及这个时间标量的精度。出现时间和结束时间在数据库中用公历表示，仅表示到年。之所以用年来表示，是基于这样的考虑：

由于历史资料现实、以及研究水平和程度问题，实际上是不可能所有的地名纪录都有确切的时间（Beginning and Ending ），根据我们的认识，历史上记载的地名有关起迄时间有以下几种可能:

1)，精确到大时期，如“秦汉”、“魏晋南北朝”、“宋元”、“先秦”、等

2)，精确到朝代，如“西汉”、“唐朝”、“秦初”、“清末”、“元中期”、等

3)，精确到年号或庙号，如“元始时”、“大历中”、“至元初年”，等

4)，精确到具体年份，如“开元十九年”、“康熙十三年”、“辛亥”、等

5)，精确到季和月份，如“元丰九年夏四月”、“甲子季秋”等

6)，精确到具体日期，如“1934 年 6 月 12 日”、“嘉靖十四年五月甲辰”等

而考虑到数据检索的要求和检索条件的实现，我们认为以“年”作为基本单位已经足够了。上述 1-3 的情况依据规定估计到具体的年份，这需要研究者给出一个意见，或仅仅是一个规定。5-6 的情况略去年以下的信息。但 1-3 个情况的所有文字信息必须保留在释文中，给用户提供一个参考，以便以后有了进一步的研究后进行修改。而 4-5 的情况也保留在释文中，作为一种文本信息提供给用户。同时在每一个开始和结束时间的精度属性上注明相应的类型标记，这个标记实际上也是一种资料信度信息的表达，可以方便用户很快确定数据的精确程度，并快速地找到相同的精度类型。这样处理方法，一方面考虑了现实查询的需求，对一个涉及到数千年基础历史地理信息的数据库来说，能够直接查到具体的年份，是能够满足绝大部分需要，另一方面也考虑到数据库的编写和查询机制上比较方便实现。

数据库属性　数据库属性包括数据的 ID 编号，任何一条记录在 CHGIS 数据库中都有一个 ID、用于表识一条地名记录，唯一性是它的主要特征。ID 主要用于连接位于不同数据库中的记录。数据库属性中还包括记录的编写者和核对者。

释文　是有关历史地理信息记录的文字描述，如原始资料、出处、作者定位和定时间的依据和理由等文字信息。释文即包括了历史地理信息属性中的内容（实际上属性内容就是根据释文编写的），还包括那些在属性数据中无法表示或特殊的情况和说明。

[3] Ian Johnson，Mapping the fourth dimension: the TimeMap project，http://www.archaeology.usyd.edu.au