主编　马家奇

传染病流行病学数据共享平台设计与实现

北京大学医学出版社

传染病流行病学数据共享平台设计与实现

主　编　马家奇
审　校　熊光魁
顾　问　杨功焕　曹务春　宫　鹏
编　者　（按姓氏笔画排序）
马家奇　王丽萍　王松旺　李言飞
张业武　赵自雄　戚晓鹏　龚琼宇
彭志勇　鲍一丹　蔡　君　熊光魁

北京大学医学出版社

图书在版编目（CIP）数据

传染病流行病学数据共享平台设计与实现/马家奇主编. —北京：北京大学医学出版社，2010.3
ISBN 978-7-81116-735-1

Ⅰ.①传… Ⅱ.①马… Ⅲ.①传染病－流行病学－数据处理 Ⅳ.R18-39

中国版本图书馆 CIP 数据核字（2010）第 011033 号

传染病流行病学数据共享平台设计与实现

主　　编：马家奇
出版发行：北京大学医学出版社（电话：010-82802230）
地　　址：（100191）北京市海淀区学院路 38 号　北京大学医学部院内
网　　址：http://www.pumpress.com.cn
E - mail：booksale@bjmu.edu.cn
印　　刷：北京溢漾印刷有限公司
经　　销：新华书店
责任编辑：李　娜　　**责任校对**：杜　悦　　**责任印制**：郭桂兰
开　　本：787mm×1092mm　1/16　　**印张**：9　　**字数**：226 千字
版　　次：2010 年 3 月第 1 版　2010 年 3 月第 1 次印刷　　**印数**：1—3500 册
书　　号：ISBN 978-7-81116-735-1
定　　价：29.00 元

前　言

我国在长期的疾病监测工作中逐步建立和完善了从中央到基层的纵向监测网络体系。2003年SARS疫情发生后，中国疾病预防控制中心建立了“以协同工作模式为中心”的传染病网络直报信息系统。随着信息技术的发展，为迎接卫生改革的深化和疾病预防控制需求变化的新挑战，我们必须在实践中不断探索适应新形势发展的传染病流行病学信息资源管理与服务模式。

一方面，由于公共卫生面向服务型转变，一站式健康信息服务公众需求的增加，已经在客观上提出了梳理流程的要求；另一方面，现有条块分割体制又不能为流程再造提供政令依据。这已成为当前我国公共卫生信息化发展的一个主要矛盾。现有的传染病监测信息管理基本上都是孤立、分散的，与其他相关环境、气象、土地等数据资源缺乏整合共享利用的信息平台，而任何传染病的传播流行都不会是孤立事件，因此单一来源的信息资源作用非常有限。

通过国家自然科学基金的支持（项目编号：30590371），我们提出了“我国重要传染病流行病学数据的收集与整合及其共享分析技术平台研究”，建立针对重大传染病分析预测的共享数据模型，并以Web自助服务网站群服务这一新的理念，建立传染病流行病学数据共享平台，实现信息技术服务和信息资源服务的共享。

本书旨在通过对传染病流行病学数据资源的描述、对多源数据总体特征的分析、对数据标准及数据质量控制管理方法的阐述，以及对共享平台信息技术相关理论、技术与方法的介绍，结合传染病监测资料的特征，从专业应用需求出发，通过示例的方式由浅入深地介绍技术平台原型设计与实现、基于WebFax数据采集技术服务、GIS展示技术服务共享过程，并重点收集与整合我国传染病的流行病学数据，以疟疾数据为例研究传染病流行病学数据以及环境因素数据二者之间的关联分析应用模式。力求使读者通过系统学习，能够以决策者、管理者或专业人员的特定需求为依托，实现个性化信息资源和信息技术管理的共享服务。

在本书付梓之际，我们向对本书编写过程给予支持的所有人员表示感谢。特别是在百忙中参与本书审核的中国军事医学科学院曹务春教授、中国科学院遥感所宫鹏教授、清华大学徐彬教授等。

我们衷心希望读者在读完本书后能够获得一个关于传染病流行病学多源数据整合共享管理的坚实知识基础。本书的出版只是抛砖引玉，相信今后会有更多更好的传染病流行病学信息管理技术与方法的著作问世，使我国在传染病监测领域的研究、探索和应用取得更大的成就。限于编者水平，书中难免有疏漏之处，恳请广大读者斧正。

编　者

目　录

第一章 传染病流行病学数据共享平台概述

传染病流行病学数据共享平台适应国家公共卫生对传染病预防控制的需要，重点对传染病流行病学多源数据实施从采集、处理、加工、存储、共享、交换到应用和展示的全过程业务流程处置。各级政府、卫生行政部门、疾病预防控制中心（简称疾控中心）、卫生统计信息部门、卫生监督部门、医疗救护部门及其他相关部门的广大工作人员和医疗卫生人员，通过对平台信息数据的获取、信息管理、信息利用和信息发布，服务于预防控制传染病、创造健康环境、维护社会稳定、保障国家安全和促进人民健康的最终目标。

传染病流行病学数据共享平台的设计和实现，是传染病预防控制应用的重要组成部分，是卫生信息化建设再上新台阶的关键步骤，更是一个需要认真研究和实施的系统工程。

一、必要性

传染病流行病学数据共享平台的规划建设作为一项重要的基础性工作，其必要性表现在以下四个方面：

第一，平台是深化传染病流行病预防控制应用研究的重要举措。近年来，全球新传染病的不断出现、旧传染病的重新肆虐以及生物袭击人为造成传染病的发生和流行，已成为人类必须面对的严峻现实。我国是世界上传染病发病率最高的国家之一，虽然已建立起较完备的流行病学监测网，积累了大量的传染病疫情和媒介生物方面的信息数据，但大多是分散、孤立的资料，技术上也难以实现集成分析与联合处理，极大地限制了其在指导预防策略和制订措施方面应发挥的作用。

应用传染病流行病学数据共享平台，不但能动态分析传染病的时间与空间分布特征，而且可以使我们以全新的角度和方式来研究和认识传染病，从其发生和流行的环境来观察传染病；这不但可以深化传染病的监测和预警，有利于发现重点疫区，为制订适宜的防治策略和措施奠定基础，而且将为大面积的监测提供经济、有效的方法，为突发疫情的有效控制提供决策依据。

传染病流行病学数据共享平台为传染病及其环境因素信息资源的开发、管理和服务提供了有效的手段；以共享数据库和信息系统建设应用为基础，使传染病和时空环境因素紧密整合，方便、及时、准确地为传染病研究提供所需要的各种信息，使之成为传染病及其时空环境因素研究的重要桥梁和纽带。当前，传染病流行病预防控制研究对信息深层次的需求越来越迫切，传染病流行病学数据共享平台建设的功能、规模、信息量大小和使用频度已经成为衡量我国公共卫生信息化程度的重要标志，成为衡量传染病传播规律研究水平的重要体现。

第二，传染病流行病学数据共享平台是克服公共卫生信息系统缺陷的有效手段。2003年SARS疫情的发生和蔓延，充分暴露出当时我国公共卫生信息系统存在的严重缺陷。截至SARS暴发以前，我国传染病疫情报告采用传染病报告卡及报表逐级审核汇总上报的方式，易受到人为因素的干扰，不能准确、客观地反映传染病的真实发病情况，不能为传染病控制提供及时、可靠的信息。当时传染病疫情报告信息系统为单机版，系统的安装、维护、更新和培训也存在较多问题。此外，我们还存在一些传染病的专报系统，如结核病专报系统、性

病专报系统、艾滋病专报系统等，使得疫情数据不一致，并造成了资源浪费，增加了各级工作人员的负担。

SARS暴发以前，我国疫情报告、疾病监测实行旬、月报和逐级统计报告制度，即最基层的医疗机构发现传染病后将传染病报告单邮寄到县卫生防疫站，县卫生防疫站统一录入计算机并汇总为报表，再逐级上报，最后到中央。这一报告过程在SARS流行的关键时期，从基层报告到中央平均需要8天时间。报告时间的滞后严重延误了时机，其时效性和敏感性差的缺陷十分突出。

以往我国卫生信息网络的覆盖面较小，虽然疫情报告、疾病监测系统覆盖了全国县、市、省和国家疾病预防控制机构，但疫情和突发公共卫生事件报告单位和报告人是各级医疗机构、城市社区卫生服务中心、农村乡镇卫生院等，仍然未能实现网上直报和个案报告。

传染病流行病学数据共享平台的设计和实现，确保我们有一个统一的传染病预防控制信息管理、交换平台，使之具有强大的传染病信息的整合能力，能够克服疫情报告、疾病监测准确性差、时效性差和覆盖面小的严重缺陷。

第三，传染病流行病学数据共享平台可以从根本上解决传染病研究中从信息数据到硬件、软件的不一致性和不兼容性。SARS肆虐之前，在公共卫生的信息标准开发、数据编码和交换方式等方面，我们几乎没有投入相应的力量去研究，医疗、预防和卫生管理之间的信息链条也很不完整。在推进传染病研究信息化过程中，人们发现现有的众多系统都建有自己的数据库，从硬件、软件到数据的不一致性、不兼容性，系统相互之间的信息封锁已成为阻碍传染病研究信息化进程及传染病现有信息价值进一步发挥的关键因素。

特别是在传染病研究中，跨领域的研究缺乏共享标准或标准不统一，造成领域概念差异，从而导致对数据理解的差异，数据很难在概念层面上实现融合和共享；对专题研究则由于基础数据来源异构，由此导致衍生数据的重复和缺失；同时，更缺乏面向专题分析的统一数据模型和数据集成模式；且非标准化编码也是造成不同系统之间数据共享问题的重要原因。

这里的根本原因是缺乏统一的数据结构标准和分析技术平台，没有可以进行比较分析的共同的时间和空间基础，难以进行宏观的分析。这就需要把多个“信息孤岛”的传染病数据集成，以建立一个统一的数据模型和数据交换标准，使信息集中成为资源为各种用户共享。还要对传染病信息进行整合，做好资源的规划，使之便于集中维护管理，同时也使其成为消除条块分割、解决数据冗余和实现一致性、兼容性的最有效手段。怎样结合我国传染病及其时空环境研究实际应用的战略规划目标、历史、现状，构建为相关各部门、各组织所共享、连接便捷、资源丰富、功能强大的传染病流行病学数据共享平台，就成为当前传染病信息化建设中必须要着力解决的重要课题。

第四，传染病流行病学数据共享平台有利于统筹各地方、各单位传染病信息资源的规划、管理、交换和使用，建立有序的传染病信息资源共享机制。在过去旧的管理体制、机制、技术水平和观念意识之上建立起来的传染病信息系统，越来越难以满足人们的需要，网络、系统重复，数据冗余、条块分割、功能落后、相互脱节等矛盾日益突出，阻碍了传染病研究信息的快速流动和共享。由于不同的专家、不同的部门和组织对于传染病信息都有着各自的要求和格式，在信息传递不够顺畅的时代，人们往往会重复采集信息，不可避免地造成业务延迟、输入错误和额外费用的负担。随着新技术（网络系统、大型数据库检索、存储等技术）的发展，构建传染病流行病学数据共享平台可以统筹发挥其资源整体优势，为各信息

资源权威发布者提供规范、科学的共享发布手段，为各医疗卫生部门提供资源的检索、定位与服务。通过与传染病及其时空环境因素共享数据库信息交换平台提供的目录服务相结合，可以解决各部门重要信息资源管理难和共享交换难的问题，是克服各自为政、系统重复、数据冗余的必由之路。这不但符合公共卫生信息化建设的需求，而且对推动医疗卫生各部门改革、提升工作效率、提升预防控制传染病的科学决策能力，都有着重要意义。

二、平台的设计目标和内容

设计和实现传染病流行病学数据共享平台的目标，主要是通过现有的大量传染病流行病学数据和较完善的传染病流行病学监测和报告系统，面向我国各种传染病流行病学、媒介生物及环境的相关数据，研究并提出共有公共模式，制订科学的数据描述方法以及数据质量控制规范，建立传染病流行病学数据的共享数据模型和共享标准；针对全国、省、地（市）、县和社区等多个层次，面向主题、按照时空进行数据组织，有效集成相关数据，提出集成数据管理的物理组织方法，实现重大传染病流行病学数据的有效整合，并搭建数据共享分析平台，在此基础上实现同一信息数据基底上多层数据的叠加和可视化。该平台将主要用于分析不同类型传染病的流行规律及其主要的自然和社会因素，建立传染病的流行病学动态模型。鉴于现有的传染病监测信息基本上都是孤立、分散的，但任何传染病的传播流行都不会是孤立事件，因此现有系统作用非常受限。通过传染病流行病学数据共享平台的设计和实现，我们可以建立针对传染病分析预测的共享数据模型，利用数据仓库技术，在实现传染病相关数据集成、调度和融合的基础上，实现面向事件的传染病信息挖掘和预案关联，从而建立为快速反应和辅助决策服务的传染病流行病学数据共享分析平台，并利用平台的支持进行相关的模型研究及预测分析。

吸收国际最新的研究成果和经验，以我们已有的研究和实践为基础，立足于创新和突破，传染病流行病学数据共享平台的设计内容可以归纳为以下几点：

1. 传染病信息共享与标准化规范的规划与设计　①建立传染病、环境因素与其他相关数据的共享信息标准；②建立传染病、环境因素与其他相关数据的质量控制规范。

2. 传染病流行病学数据共享模型的规划与设计　①建立传染病、环境因素与其他相关数据的共享数据结构；②以本体、知识库为支撑，建立病患信息、传染病流行病学数据以及环境因素三个库之间的关联模式。

3. 传染病流行病学数据共享平台的规划、建设与应用　①考虑全国、省、地（市）和社区等多个层次，对我国传染病流行病学数据进行收集与整合；②整理我国传染病历史资料；③整理分析我国传染病流行趋势、流行规律与防治模式；④构建我国传染病的标准化信息数据库/数据仓库。

4. 软件平台与工具套件开发　①实现基于 GIS 环境的、同一信息数据基底上多层数据的叠加和可视化，具体而言就是实现如图 1－1 所示的一个软件平台；②研发传染病流行趋势空间分析工具套件；③实现传染病流行病学数据门户；④利用平台的支持，通过数据挖掘进行相关的模型研究及预测分析，建成传染病流行病学数据共享分析平台。

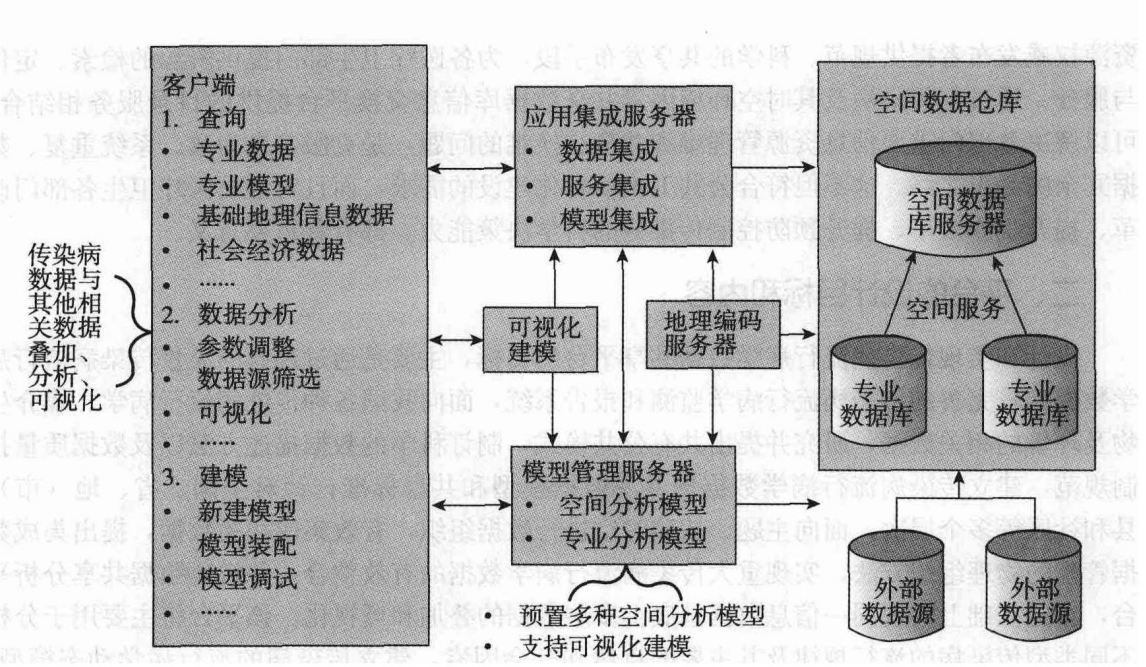

图 1-1　平台示意图

三、平台设计基本要点

传染病流行病学数据共享平台的设计要求是通过已有的大量传染病流行病学数据和较完善的传染病流行病学监测和报告系统，建立针对重要传染病的综合分析技术方法和信息共享技术平台，分析在我国快速发展过程中不同类型传染病的流行规律变化及其主要的自然和社会因素，建立重要传染病的流行病学动态模型，为传染病预防体系的有效运作提供科学的决策依据。其整体设计的研究思路如图 1-2 所示。主要需要进行三个层面的研究：

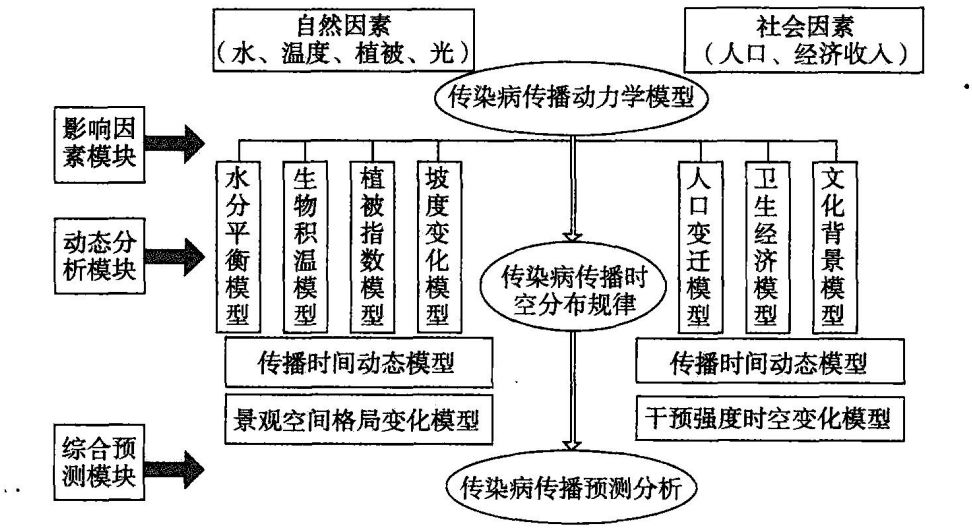

图 1-2　总体研究框架

1. 数据收集、整合体系研究　主要解决对已有的重大传染病流行病学数据进行系统、专题的收集和组织整理的问题。

2. 重大传染病流行病学共享数据平台建设　包括制订数据质量控制规范、数据共享标准，研究共享数据模型和数据关联模型，以及具体的数据库建设。

3. 软件平台和共享数据门户研究建设　软件平台主要是集成常用的统计分析和空间分析模块，并针对专题研究，开发部分专用分析模块。这些软件模块可供其他子专题调用，接口拟采用 Web service 方式。共享数据门户主要提供数据发布、维护、检索和下载功能，可供其他子专题使用，同时逐步建设成为一个可长期运行、可扩充的业务系统。

传染病流行病学数据集成共享平台的设计和实现是一项全面、细致、逐步深入的系统工程。它的研究设计内容复杂、牵涉面广，全过程涉及七大基本要点，这里我们尝试从宏观综合层面上予以介绍，本书还将在后面的章节中分别对这些关键环节加以阐述。

（一）传染病流行病学数据资源规划

数据资源规划是指对我们工作中所需要的数据，从产生、获取，到处理、存储、传输及利用进行全面的规划。在传染病流行病学数据共享平台的设计和实现过程中，首先要考虑的是传染病流行病学数据资源规划，它是信息化建设的基础工程和先导工程。

传染病的发生、发展是多因素综合作用的过程，因此与其相关的数据资源也来自多个领域，包括来源于病例的信息、来源于标本的信息、基础信息和其他信息。病例信息主要是以病例为基础的监测数据及其汇总，也包括以事件为基础的监测数据和计划免疫信息、病例的症状信息等；标本信息包括常规实验室监测以及暴发流行时现场采样获得的标本的属性信息，如标本的检测结果等；基础数据如各种标准、人口统计学信息、监测系统的基础编码等；其他数据如动物、病原生物、食品、环境、气象、地理、经济等信息，这些数据资源从不同角度作用于传染病的发生、发展、预防和控制活动。

传染病数据资源规划需要传染病的数据管理机构和 IT 咨询机构来统一规划，通过分析、整理用户的实际需求，针对具体问题搭建规划平台。在规划平台之初，首先要整合传染病的数据资源，建立起平台内数据共享和交流互通的机制，采用与现行机构相适应的数据管理模式（内部集中管理、外包服务管理等），然后结合先进的传染病专业数据分析和数据挖掘方法，从内部渠道上保证数据信息传递的高效性和准确性。

（二）传染病流行病学数据标准

传染病流行病学数据标准是传染病流行病学数据共享平台上系统互操作的基础。数据标准是一个经商定的、共同的和一致的记录信息的方法，它允许数据在不同的信息系统之间交换，并且保证数据在不同的系统、程序和机构之间都有相同的理解。数据标准对于提供无歧义的数据含义、形成使数据能够汇总的基础以及数据挖掘应用中的数据判别显得非常重要。标准化和规范化研究工作是一项最基础性的建设工作，如果没有统一的数据标准，没有统一的编码方法，没有统一的数据交换协议，各地区、各部门形成的数据将很难汇总分析和共享。实践表明，数据标准方面的差异已经成为当前各国信息技术应用发展最主要的障碍之一。

经常被人重复引用的一句话是“标准的可爱之处是总是有许多可供选择”，在医疗和公共卫生信息学领域中同样如此[1]。最著名的卫生信息标准是 WHO 发布的疾病与健康相关问题国际分类（目前为 ICD－10）以及损伤、失能和残疾国际分类（ICIDH）。与传染病有关的数据标准可分为四类，包括术语（如 ICD－10）、消息机制（如 HL7）、业务/需求（如

HIPAA 法令）和数据内容（如核心数据集）。美国国家生命统计中心（NCVHS）和统一卫生信息学（CHI）项目提炼出了应用于传染病信息学的数据标准（见表 1-1）。当前美国联邦政府的信息系统要求采用这些标准。在公共卫生领域中，美国疾控中心还通过国家电子疾病监测系统（NEDSS）和公共卫生信息网络（PHIN）推动了数据的标准化。这两个项目定义了一系列词汇表、消息传输标准、消息和数据的格式以及公共卫生信息系统所需的一些组件[2]。

表 1-1　应用于 IDI 领域的 CHI 标准

CHI 采用的标准	域
美国卫生信息传输标准（HL-7）	信息传输
观测指标标识符逻辑命名与编码系统（LOINC）	实验室及临床观测指标
医学术语系统命名法－临床术语（SNOMED CT）	实验室结果内容，非实验室干预和过程，解剖、诊断，描述临床药物
临床药学标准术语（RxNORM）	描述临床药物
HL-7 临床免疫程序（CVX）和厂商编码（MVX）	免疫接种登记，术语

当前，网络技术、信息技术的发展为解决数据交换和传输提供了可以依靠的良好的技术支撑手段。在信息数据资源规划的基础上，建立传染病信息数据资源的基础标准、平台功能模型和数据模型，并在这些标准和模型的指导、控制和协调下，进一步实施传染病信息化建设的网络工程、数据库工程和应用软件工程，我们就能保证传染病流行病学数据共享平台建设的高起点、低成本，实现传染病信息数据资源整合和共享的最终目标。

（三）传染病流行病学数据收集与整合

准确、及时、综合的信息是实现传染病预防和控制的前提条件，也是制订策略和实施评价的基础。传统的信息采集方法是纸质的报告卡或调查表，如传染病报告卡和流行病学现场调查表。随着信息来源和种类的增多以及相关技术尤其是互联网的发展，目前传染病数据的采集方法和手段得到了较大提高。个人手持设备如掌上电脑（PDA）和智能手机、GPS（全球定位系统）设备等已逐渐应用于传染病信息的采集，在相应软件的支持下能够快速获取现场调查数据，在具备无线网络的条件下也可实现信息的实时发送，使信息采集变得更简单、快速。如中国疾控中心开发的汶川地震灾区应急手机疫情报告系统作为网络直报系统信息采集手段的重要补充，经过现场使用，有效替代了原有网络直报系统的信息采集功能。

系统间数据交换是另一个应用热点。由于传染病数据来源的复杂性，涉及大量不同机构，所需的信息往往以电子文件的形式存储在这些机构现存的信息系统中，通过建立数据标准、开发数据接口使这些系统间的通讯成为可能。然而，传染病信息数据的海量性和无序性却与我们使用数据的选择性形成尖锐对立，只有经过细致的整合才能供我们使用。这里的数据整合，就是将收集来的大量传染病原始数据进行筛选和判别、分类和排序、计算和研究、著录和标引、编目和组织，使之成为可用的数据，这恰恰是传染病流行病学数据共享平台的必备功能，是平台设计不可缺省的一步。

（四）传染病流行病学共享数据库

传染病流行病学共享数据库是传染病流行病学数据共享平台设计中的一个重要环节，是平台数据存储的组成部分。共享数据库存储需要共享的数据，实现数据集成，并提供一个统

一的数据模式，以供客户端订阅自己需要的数据。一方面，它不同于数据仓库，它不是面向分析和决策支持的，其中的数据也不是按照维来存储，不维护历史数据，因此其建设周期短，维护代价小；另一方面，共享数据库也不同于传统的数据库，它存储着需要共享的数据，其目的是实现数据共享，并且由于它提供一个全局的数据模式，有效地实现了数据的共享，在实际应用中取得了较好的效果。

由于传染病流行病学数据是来自多个数据源的异构数据，数据库的异构体现在计算机体系结构、基础操作系统、数据库管理系统本身三个方面。建立的共享数据库可分为四个层次，即基础性共享数据库、应用性共享数据库、中间性共享数据库和管理性共享数据库。异构数据库的集成处理主要侧重于异构分布式数据库集成研究。这种集成技术是将参与数据库的有关信息在逻辑上集成为一个属于异构分布式数据库的全局概念模式，以达到信息共享的目的。实现策略首先是通过模式翻译器将局部数据库模式以某种公共数据模型为基础映射成局部集成模式，然后通过模式集成器将各个局部集成模式根据用户的需要采用全局数据模型来定义，最终成为全局概念模式。

实践表明，数据模型和接口不同，即缺少标准化，是限制数据库间信息交换与共享的瓶颈。为实现不同数据库间信息的共享和交换，国外已经开发了大量系统，按照架构模式的不同可以分为三类：导航（navigator）、中介（mediator）和数据仓库（data warehouse）[3]。导航是基于链接的对一些数据源的整合，这样的一个门户通常并不整合数据本身，只是向用户提供一个外部数据源的导航页面。中介是通过重新定义用户对外部数据源的查询来提供对分布式数据的访问，这两种解决方案的主要缺点是可行性和效率较差。数据仓库是通过一个综合的数据模型将大量外部数据源在语义上完整地整合为一个本地数据库，由于此方法避开了其他方法可能遇到的如网络瓶颈、外部数据源的短时不可用及外部数据源的变化等典型问题，因此可以实现高效查询。

（五）传染病流行病学共享数据网络系统

传染病流行病学数据共享平台运用IT技术的新进展，对大规模、分布式、异构的共享数据库中海量的数据资源进行整合，实现全方位、深层次的资源共享，并在高性能环境的支持下开发基于网络系统的先进应用系统，使其能够满足未来疾病预防控制研究的需要。显然，传染病流行病学共享数据网络系统是传染病流行病学数据共享平台的重要核心组成部分，是必不可少的支撑环境。

一般说来，传染病流行病学共享数据网络系统要适用于主流网络拓扑结构，支持主流网络协议，并能适应不断发展的网络技术的需求，支持数据通信、语音通信、多媒体通信以及各种控制信号的传输；网络设计应能有效地避免单点失效，在设备的选择和关键设备的互联时，要提供充足的冗余备份；网络系统要能适应较复杂的空间使用环境，保证在信息集成网络系统中传输的各类信号之间互不干扰；系统应提供完备的安全防护策略，能防止非法访问，保护重要业务系统不受非法入侵；网络要提供较强的系统管理能力，可以有效地进行系统管理、系统维护、系统故障的排除；系统实施过程中要保证系统间的协调配合，要遵循开放性标准，网络互联设备应支持多种协议，能与其他厂家设备互操作；主要网络设备可以采用目前成熟的和国际先进的标准化产品。

通过建立集共享数据库、海量存储、超级计算机、高速网络等软硬件于一体的共享数据网络平台；建立专业领域数据分中心，规划、设计和集成专业领域数据库资源，组织不同病种标准规范的制订和实施，传染病流行病学数据共享平台才能够应用于专业领域，提供专业

领域综合研究信息的共享与服务。建立完善的网络体系，建设面向传染病数据共享需要的网络软硬件支撑环境，将为我国传染病及环境因素共享信息服务体系的形成和先进应用系统的发展打下坚实的基础和提供有力的支持。

（六）传染病流行病学数据的展现和服务

传统的流行病学数据展现方法通常与数据的分析方法密切相关。如人、时、地三间分布数据的频数表、汇总表格等，也可使用绘图工具制作二维图、三维图，如直条图、饼图等。数据的图形展现使人们不再局限于通过数据表来观察和分析信息，而能以更直观的方式看到数据背后隐藏的信息，识别疾病流行、传播的特点，起到决策支持的作用。由于疾病的发生和暴发与病例发病地点和疾病的传播空间因素有很大的关系，因此将疾病的空间因素绘图展示出来，将有助于加深人们对疾病暴发的理解。与原始的疾病数据不同，疾病地图提供了视觉展示的方式来识别病因以及患者和他们所处环境之间的关系，使医务人员和一般公众能够直观和形象地交流疾病的分布情况。

在利用传染病数据向专业人员和公众提供服务方面，既有简单的数据发布，如以美国的MMWR和欧盟的Eurosurveillance为代表的传染病公告板及我国卫生部发布的疫情数据等，也有部分研究与实践将各类传染病的数据发展成数据产品，提供给公众使用，满足专业人员和公众对疾病信息的需求。如美国疾控中心的CDC WONDER在线共享数据库分为艾滋病、出生、婴儿死亡、结核病等若干主题以供查询相应数据；加拿大公共卫生署在其网站上列出了21个公共卫生监测系统的汇总数据，包括传染病、慢性病、伤害等的周报、月报和年度汇总数据。部分监测数据还提供交互式的查询，使用者可自定义查询和显示的条件。此外，如欧盟的Health-EU和EUPHIX门户网站也都提供了大量的信息和数据查询服务。

地理信息系统（Geographic Information System，GIS）提供了一个有效的方式来管理、存储、分析和展现疾病信息。它不仅在空间数据的绘图和分析上有很强的能力，而且对非空间数据也是如此，表现在它能够整合并表达多种类型的数据，如人口统计学和环境数据等。GIS具备很多有用的功能，如疾病监测地区的网络分析、缓冲分析和统计分析。当疾病出现时，GIS能够快速描绘出疾病信息并动态分析疾病传播。Foody[4]通过比较不同时间区间的专题地图，反映出疾病的时空变化，包括在时间上的聚集、媒介传播的速度、易感人群的发病率。互联网的快速发展影响了基于Web的GIS的普及，当前的GIS应用已经可以实现通过Web访问卫生信息，出现了大量定制的在线交互式地图。通过结合GIS地图和其他与传染病发生、发展有关的因素变量（如气候、环境、人口密度等）来实现预警，估计疾病可能的传播地区。另外，还有针对普通公众的GIS应用，可由用户设定查询条件，如年龄、性别、病种、年份等，在地图中进行展示。

有文献提出了使用GIS在线展示疾病信息时面临的四个挑战，即疾病数据在语义、语法和结构上的异质性；当前的应用系统尚难以重用和整合；使用中缺少不同疾病服务之间的互操作；缺乏合适的绘图方法与病例隐私保护。为了解决数据异质性和系统间互操作的问题，人们提出了图1-3所示的架构，它包含了四层：数据存储层、分类引擎层、标准的卫生服务层、地图和动画层。基于此架构，人们在加拿大的新不伦瑞克省和美国缅因州建立了一个在线的面向服务的传染病绘图和共享系统，结果表明该系统能够满足传染病流行病学数据展现的需求，其分析功能也能够早期发现疾病暴发[5]。

传染病流行病学数据共享平台依托内容管理系统，结合数据仓库和商业智能技术，综合运用上述各类展现方式，利用成熟的公众信息发布体系，可以建立信息门户系统，将与业务

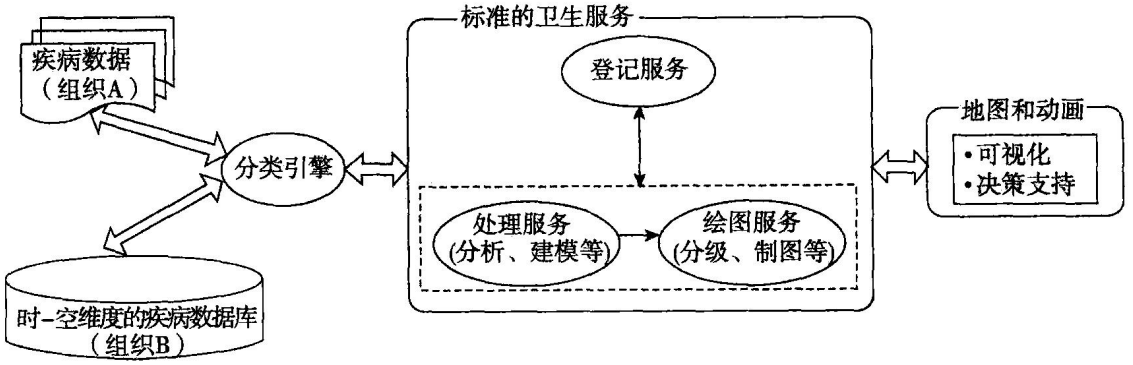

图 1-3　传染病流行病学数据的 GIS 展现平台架构

流程关系紧密、展现内容动态、形式多样的高质量的传染病流行病学数据信息和预测、决策信息对外充分展现，以满足公众的实际需求。

（七）传染病流行病学数据仓库和数据挖掘技术的进一步应用

数据仓库作为先进的 IT 技术，在条件成熟和具备的时候，应该作为核心手段在传染病流行病学数据共享平台中广泛深入应用。由于传统数据库技术不能很好地支持分析型处理，数据仓库（Data Warehouse）的概念应运而生。20 世纪 90 年代初，W H Inmon 将数据仓库定义为“数据仓库是面向主题的、集成化的、稳定的、随时间变化的数据集合，用以支持管理层的决策过程”。这个定义指出了数据仓库的几个特征：面向主题，即数据按照一定的主题域进行组织；集成的，数据仓库中的数据是在对原有分散的数据库数据抽取、清理的基础上经过系统加工、汇总和整理得到的；相对稳定，一旦某个数据进入数据仓库以后，一般情况下将被长期保留，很少被修改和删除。数据仓库最根本的特点是物理地存放数据，这些数据并非是最新的、专有的，而是来源于其他的数据库。其建立并不是要取代原有的数据库，而是建立在一个较全面、完善的信息应用的基础上，用于支持高层决策分析。

数据仓库已发展成为当今信息管理技术的主流，通过建立数据仓库可集成多个来源的异构数据并将其转换为综合的、高质量的信息，在分析工具和模型的帮助下识别出问题和机遇，作出关键决策，形成战略并评价其实施情况。国内外已出现了大量应用于卫生领域的数据仓库，如美国国家生命统计中心数据仓库，集成了生命统计、死亡等数据；美国多个州的卫生部门也建立了综合的数据仓库，从生命统计、电子健康档案、监测数据等多个应用系统中抽取数据；加拿大公共卫生信息网络（CNPHI）和澳大利亚的生物安全系统（BSS）也都建立了数据仓库用于多源数据的集成；欧盟的 INFTRANS 项目通过建立数据仓库收集人群迁移和人口统计学信息来分析新发传染病的传播模式。

建立在传染病数据仓库基础上的数据挖掘（Data Mining）属于对传染病信息数据的分析利用。它是指从数据库中提取隐含在其中的、人们事先未知的、潜在的有用信息和知识。所提取的知识可表示为概念、规则、规律、模式等形式。其目的是通过提供工具辅助发现大量数据中的关联和模式以及基于从已知数据获得的信息来预测未知数据的价值，进而获得传染病预防控制收益。它与数据库中的知识发现（Knowledge Discovery in Database，KDD）是等价的概念。常用的数据挖掘方法有统计方法、机器学习、神经网络计算和可视化等。传染病流行病学信息来源较广，既包括日常监测数据、现场调查数据、实验室检验等结果，又包括 GIS 影像数据、遥感图片等非结构化数据。这些数据具有时间性和冗余性等特点，由

此也决定了传染病流行病学数据挖掘的特殊性。数据挖掘技术可利用传染病流行病学数据仓库，实现传染病暴发的预警、预测、分级、发现病例聚集和传播模式，发现病例之间的关联，也可通过数据挖掘实现数据的可视化展现等。

数据仓库和数据挖掘技术在传染病流行病学数据共享平台上的进一步应用，将使我们的传染病流行病学研究能力大幅度提升、如虎添翼，将为传染病流行病学研究打开快捷、方便之门，使未来的传染病预防控制工作效益显著、突飞猛进。

四、国内外研究发展现状

随着当前医疗卫生事业的不断发展，在传染病流行病学多源数据集成共享方面，人们的研究工作取得了长足的进步，突出表现在传染病信息学的研究和以美国的 BioPortal 系统为代表的传染病信息共享分析系统的应用，以及我国的公共卫生科学数据共享工程和《基于现代信息技术研究传染病时空传播与流行规律》国家自然科学基金项目中《我国重要传染病流行病学数据的收集与整合及其共享分析技术平台的建立》子课题的研究。

（一）传染病信息学

传染病信息学（Infectious Disease Informatics，IDI）是一个新兴的研究领域。它系统地研究传染病预防、探测、管理相关的信息管理和分析等问题。IDI 研究性质上是跨领域的，需要很多领域的专家，包括但不限于信息技术领域的大量分支，如数据整合、数据安全、GIS、数字图书馆、数据挖掘和可视化，以及其他领域如生物统计学和生物信息学等。传染病信息学研究框架如图 1－4 所示。

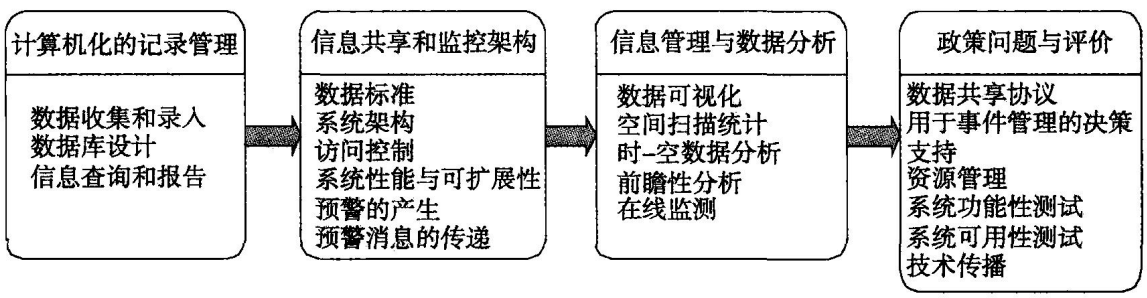

图 1－4　传染病信息学研究框架

随着传染病信息学研究的深入开展，建立于其理论基础之上的传染病信息系统，在形成高效和综合的方法来预防、探测、响应和管理传染病暴发的应用上，起到了重要作用。不同层次的实验室、医疗保健提供者和政府机构收集了大量的传染病数据。此外，很多机构已经在传染病信息学这个重要领域中开发了不同复杂程度的信息访问、分析和报告系统。例如，作为在美国应对人法定报告传染病的关键机构，疾控中心已经开发了计算机化的报告系统用于地方和州的卫生部门。类似的，美国农业部（USDA）正在升级某些动物疾病的数据系统。美国地质调查局（USGS）通过它的野生生物健康中心（NWHC）和大量的协作者来管理野生生物疾病。在其他的联邦、州和地方的卫生、农业、环境/野生生物机构和实验室也能找到这样的数据库。然而，对这些数据源的访问和相关的搜索与报告等功能可能只局限在开发系统的机构中，这就降低了传染病数据在国家和全球范围内的可用性。此外，实时的数据共享，尤其是跨领域的数据共享能够扩展专业人员的视野，并使用多个政府机构和协作者

提供的信息作出快速响应。

信息的集成、整合和共享是传染病信息学研究框架中的一个重要部分，也是解决“数据孤岛”问题，保证数据可用性和有用性的有效方法。当前，从传染病信息学的角度考虑，信息集成和共享的实现在技术和政策上还面临如下挑战[6]：

1. 现存的传染病信息系统不能完全互操作　大多数现存的系统都是孤立开发的并具体针对不同的领域。当疾病控制机构要共享不同系统的信息时，它们可能需要使用电子邮件的附件、传真、电话或者重新手工输入数据。此外，大多数系统的搜索和数据分析功能仅供内部用户使用。

2. 缺少一个跨组织边界的有效的报告和预警机制　从国家安全和公共卫生的角度看，某些传染病相关信息需要通过公共卫生机构快速传播，这样的信息可能也需要及时与国家安全机构共享。在公共卫生机构内已存在一些模型，但一般来说，当前的报告和预警机制还远远没有做到完整和有效，并可能包含了大量人为的和错误的干预。

3. 用于分析大量传染病数据和开发预测模型的信息管理环境需要很大改善　当前的传染病信息系统对专业人员分析数据和开发预测模型提供了非常有限的支持。一个整合的分析环境应能提供诸如地理编码、先进的数据挖掘和汇总能力以及可视化的展示等。

4. 数据的所有权、准确性、安全性及其他法律和政策相关问题需要仔细检查　当传染病相关数据跨部门共享时，需要解决涉及数据提供者和用户之间的访问控制与安全问题。

这一系列的挑战和难点还有待于传染病信息学研究的继续深入，逐步予以攻克和解决。

（二）美国的 BioPortal 系统

BioPortal 是一个可扩展的传染病信息共享和分析系统，其研究团队包括亚利桑那大学、犹他大学、加利福尼亚州和纽约市卫生局以及美国地理调查局等。该系统的目的是阐明和评价传染病信息共享（跨领域和地域）、预警和分析框架的技术可行性和可扩展性；开发和评价先进的数据挖掘和展示技术用于传染病数据分析和预测建模；识别在开发国家传染病信息基础架构（NIDII）中遇到的技术和政策方面的挑战。2002 年 9 月，美国国防部（DOD）、能源部（DOE）、国立卫生研究院（NIH）、疾控中心（CDC）、国家航空航天局（NASA）等 18 个机构合作建立了传染病信息学工作委员会（IDIWC）。2003 年 10 月，IDIWC 启动了 BioPortal 系统原型的开发工作。其目标用户包括公共卫生研究人员和各级执业医生、分析员和决策制订者、一般公众、公共卫生或相关领域的学生、国家安全机构、反恐和突发事件应急反应的相关部门[6,7]。

目前，BioPortal 项目已建立了一个网站（http://www.bioportal.org/）并向公众提供不同的访问权限，可查询西尼罗病毒感染、肉毒中毒、手足口病和 BioWatch 系统的监测数据，包括人、禽、动物和地理的数据等。系统可实现多源数据的查询、展现和分析等功能。当前已开发较为完善的是 BioPortal 的一个子系统 WNV-BOT 门户系统，用来整合来自纽约、加利福尼亚和其他一些联邦水平上的西尼罗病毒和肉毒中毒的传染病数据集，同时提供一套适合这两种疾病的数据分析、预测建模和信息展现工具。图 1-5 概括了 WNV-BOT 门户系统的数据来源和目标用户。

从系统的角度看，WNV-BOT 门户系统与州的公共卫生信息系统是松耦合的。州公共卫生信息系统使用双方商定一致的协议通过安全连接将信息传送给 WNV-BOT 门户系统。这些信息将存储在由 WNV-BOT 门户系统维护的数据存储中。系统也自动核查来自于外部数据来源如美国地质调查局（USGS）的数据项并把它们存储在内部数据存储中。