数值数据库及其情报服务

陈　光　祚

在计算机情报检索服务中，一些新型的机读数据库，例如有关社会调查数据、统计数据、人口数据、公司财政信息、市场行情、科学技术常数、化学命名与结构、材料性能等等的数据库，正以巨大的速度增长，越来越占有重要的地位。它们的基本特征，是向情报用户提供数值型的数据，即事物的绝对值或相对值的数字，并且能在计算机软件的帮助下，进行数值的各种运算推导，例如有外推、内插、填补空缺数据，甚至列出曲线图以及进行各种分析等等的功能。数值数据库是"源数据库"(Source database)的一部分，是情报用户借以取得经过抽取、核实、整理的数值信息的来源和借以进行定量分析的工具。因而可以说，较之书目数据库来说，数值数据库是在更高层次上经过情报加工的产物，尤其受到科技人员与管理决策人员的欢迎。鉴于它们在情报服务中所占比重和重要性的日益提高，情报科学对它们的研究兴趣也与日俱增。

一、数值数据库的基本特征

所谓数据库，是机读化的数据集合。它是任何情报检索系统的必不可少的资源，也是情报用户借助情报系统进行检索而期望获取的目标。

对于数据库，可以根据各种标准进行分类。例如可以根据数据库中信息记录的结构来分类；可以根据数据库得以利用的情报检索系统的类型来分类；也可以按其所含信息内容来分类，等等。目前，情报界倾向于主要是根据数据库所含信息内容作为基本的分类标准。根据这个标准，数据库可划分为：

（一）文献参考数据库 (Reference database)
- 书目数据库(Bibliographic database)
- 指示数据库(Referral database)

（二）源数据库 (source database)
- 数值数据库(Numeric database)
- 文本-数值数据库(Textual numeric database)
- 属性数据库(Properties database)
- 全文数据库(Full-text database)

文献参考数据库主要是二次文献数据库，它包括各种机读版的文摘、索引、目录等。它的价值在于向情报用户指引一次文献。这种书目数据库类似于治病的药方。要解决问题，还必须按照药方找到药物，进行熬煎服用，消化吸收药物中的有效成份，从而有补于身体。同样，情报用户也必须设法找到二次文献中所指引的一次文献，进行阅读，从中筛选出有用的知识或数据，从而满足自己的情报需求。至于文献参考数据中的指示数据库，其内容包括可作为情报取得来源的机构、计划、活动，乃至有特长的个人的介绍，其价值在于指引情报用户找到合适的情报源。它本身并不直接提供用户所需情报，而是起着一种指示、介绍、牵线、

搭桥的作用。而这也正是整个文献参考数据库的共同特征。

源数据库的特点在于它本身含有一次情报——即用户作为检索目的而要求获取的数值、事实或文本。人们往往把这些信息称为"纯情报"、"浓缩情报"。它们类似于经过加工的成药，是直接可以服用和解决问题的东西。为了与文献参考数据库有所区别，在英文中常用"data bank"(译为数据银行或数据库)来特指源数据库，而用"database"(译为数据库)来指文献参考数据库或泛指整个的数据库。

源数据库中，全文数据库是指机读化的文献全文。例如法律条文、文学作品作为数据库全部加以存贮，并能用来检索全文中任何一个字、段、节、章等等。属性数据库是指词典式或手册式的化学、物理数据以及其他数据。文本-数值数据库是指其记录同时包含文本信息字段和数值信息字段的数据库。许多名录型的数据库就属于这种数据库。至于数值数据库包含来自原始文献或调查统计的、并经过处理的各种数值或数据表格。在国外，有人给数值数据库下了这样的定义："它主要是计算机可读的数值性质数据的集合"。

但是推敲起来，这个定义不是十分清晰的。因为符合这个定义的范围的，不仅有严格意义上的"数值数据库"(即原始调查数据和其它统计数据)本身，而且还应包括文本-数值数据库(其记录既包含文本字段、又包含数值字段)和属性数据库(词典型与手册型数据)。正因为如此，有人认为，除了全文数据库外，所有的源数据库均可列入数值数据库之下。另一方面，即使是"严格的"数值数据库，例如美国预测公司的PTS数据库，其记录也包含书目引文。即使它的数值部分，实际上也是原始文献的更全面的内容摘要。该数据库本身以前的名称就叫做《PTS U.S. Statistical Abstracts》(PTS U.S 统计文摘)。由此看来，需要对数值数据库的含义作更明确的规定。

要更确切地规定数值数据库的范围，应该从数据库所包含的数值信息内容、数值内容的便于存取性、以及库中各有关数值的可运算性等方面来全面予以衡量。也就是说，要真正实现数值数据库的作用，必须同时兼顾数据库本身的内容以及使这种数据库得以利用的计算机系统的能力。具体来说：

第一、数值数据库的基本内容应该是数值信息。提供这种数值信息是编制数值数据库的基本目的，检索这种数值信息是情报用户的主要目标，但不排斥库中含有表明数值属性的文本字段。

第二、数值数据库的结构要便于对数值信息进行分离，便于单项或多数的检索、显示或打印。

第三、数值数据库中的各项有关数值，应能在计算机程序的帮助下进行各种运算和分析，并能进行分类、排序和重新组合。虽然这主要是取决于检索系统的能力，但它可以说是数值数据检索的精髓之所在。事实上，许多数值数据库的发行总是同特定的软件相联系的。用户在使用大多数商业渠道提供的联机数值文档时，必须使用联机系统经营者提供的检索与分析软件包。

这里，还需指出的是，当前由于数据库品种的多样化，也引起了数据库分类概念的新变化。今天的一种新的划分方法是，数据库可分为：①文字型数据库。其下又可细分为文献型、事实型、百科全书型、词库型、全文型等数据库；②数值型数据库；③图象型数据库。这种新的分类概念的出现，一方面反映了图象型数据库的异军突起的局面(尽管目前所占比例很小)，另一方面也更突出了数值数据库作为三足鼎立之一的地位。

二、数值数据库的迅速增长

数值数据库在整个科学进步的基础结构中占有重要地位。数值数据提供了科学研究中试验、测度、计算、记录与传播等活动链条中的基本联系。工程设计、经济分析、未来预测、规划决策等等都离不开数值数据。发达国家的技术优势，部分地是建立在他们掌握的良好的数值数据库基础之上的。

正因为如此，数值数据库的增长要比书目数据库快得多。现在，无论在增长速度和绝对数量方面，它都正在超过书目数据库。1982年出版的《电子出版指南》(Guide to Electronic Publishing)对1982年以前近500个联机数据库作了分析。现引述此材料如下：

数据库类型	所占百分比
文献参考根据库	48%
书目数据库	38%
指示数据库	10%
源数据库	52%
数值数据库	38%

1982年Guadra Associates列举的一千多种商业提供的联机数据库中，将近一半(493种)是数值或文本-数值数据库。

而到1985年，据统计，世界200多个联机情报检索系统提供公众使用的2,200多个数据库中，91%是源数据库。当然，源数据库中主要是数值数据库。

面对这种迅猛增长的势头，国外情报科学界普遍认为，源数据库及其服务是未来的浪潮。

推动数值数据库迅速增长的技术方面的原因，是微型计算机和数据库管理软件的飞速发展。利用这种软件在微型计算机上建立的数据库，大部分是文本-数值数据库。

目前，据估计，数值数据库的90%以上是属于商情方面的数据库，如经济、财政、公司及工业方面的专门数据。在科学技术领域的联机数据库中，大约10%以上是数值与属性数据库。至于在其它社会科学部门如人口调查、民意测验等方面，数值数据库更是占有压倒的优势。

商业与经济方面的数值数据库的发展之所以如此迅速、独占鳌头，是由于这方面的信息是属于"高价值"的信息。目前在西方流行的一句话是："关于金钱的情报已变得几乎同金钱本身一样重要。"他们认为最高梯级的情报就是商业需要的情报。它们最迫切需要转化为机读数据库，以便能进行及时的检索和准确的分析。

三、数值数据库的结构与编制特点

在自然科学和社会科学领域内，有各种类型的数值数据。它们在准确性、可靠性、可重复测定性、对客观条件的不依赖性等方面存在大量的差别。例如，物理与化学领域的数据，一般是在实验室中、用充分定义的结构和在细致控制的条件下完成测定的。这种测定是可以重复的，可以在不同的时间、地点进行重复测定，也可以由完全独立的研究者进行重复测定。而在工程与材料科学领域中的数据，尽管类似于物理与化学数据，但也有所区别。虽然它也是可以重复测定的，但要大大依赖于测定的条件准备与确切的技术细节，并且要依赖于工程材料的市场产品的可靠性。在生命科学领域中，更多的是观察数据，少量的是可以作定量测定的，而大部分还需用语言文字来描述，即所谓"事实数据"。在地球科学、空间科学、

环境科学、气象科学等领域，通常是观察数据，即涉及单个事件（如火山喷发、飓风、大气中CO_2的分布）的数据，往往是不能确切地进行重复测定的。在社会科学领域中的数据，例如物价、人口、失业等数据，对于时间地点的依附性，它的随机性和不可重复测定性是比较明显的，它所受到制约的条件也更为复杂。

另一方面，情报用户对于不同类型数值数据的利用方式与目的也是不同的。有的只是简单的查找和引用，以应用于科学计算和工程设计；而有的则须对查到的数据进一步地进行解释与分析，找出适当的数学模型，并在历史和现有数据的基础上进行预测。

上述两个方面，即数值数据的不同类型和用户的不同使用方式，决定了数值数据库在结构与编制方面有自己的、有别于书目数据库的若干特点。

数值数据库内的数据结构，要比书目数据库复杂和多样。一般来说，机读的数值数据既可以以单元形式存贮，也可以以表册形式存贮。后者是统计表格的机读模拟，而前者是原始性材料或未经加工的数据的模拟。此外，数据可以是纵向的或时列性质的，也可以是横断式的。一旦数值数据库被建立之后，所有这些文档都可被检索，其中的所有数据都可进行运算。民意测验数据、化学属性数据等一般以单元形式存贮，即每个民意测验的回答者、每种化学物质作为一条记录。这种数据可称为“微数据”（microdata）；而表册式的数据或聚集式的数据，则以N维的表格形式或矩阵形式存贮。每个记录包含一个行或列的所有信息。大多数经济数据、健康数据等都是以这种形式存贮的。这种数据可称为“大数据”（macrodata）。横断式的数据，例如有十年的人口调查数据等，它提供了一个时间的片断。时列数据是历史性的，即以一定的测度单位来表明一定的变量在一定时期内的连续的值，如月度生产或失业指数等。

数值数据库的编制，一个非常重要之点是核实数据的有效性和可靠性，并需说明在什么条件下、用什么仪器、以什么计量单位、用什么标准、什么数学计算方法与统计方法获得的。这些在建库中都是必要的步骤，以便在原则上使这些数据能够重复测定，即使有的在实际上不能作重复试验。

数值数据库往往是高度专门化的，它的编制工作依附于从事相应领域科学研究工作的机构。也就是说，建库中的对数据进行的评价和数据的保证工作必须由科学家来做，并且要依靠实验室的良好设备。

数值数据库的编制，还须找出包含数据的一次文献或二次文献，对其进行标引与再标引（如果文献是来自已标引的书目的话），以便为数据的评价作好检索的准备。这意味着，编制数值数据库工作的第一批产品是一定学科主题的细致标引了的书目。换句话说，数值数据库的编制需要文献数据库的支持与配合。

数值数据库的编制，应了解用户群及其使用数值数据的模式，如单个的值、数据序列、图表表示、分析形式、使用的单位和数学表达式及参数等等。

数值数据库的编制，应该使数据评价员（技术专家）和图书馆与情报科学家作到密切配合。由于技术专家不一定了解情报处理与数据传播的计算机信息方面的问题，因此这两部人之间取得共同语言是十分必要的。也只有这样，才能使数据的获得、核实与传播构成一个有机的整体，并建立起数据中心同它所服务的用户界之间的联系。

数值数据库的编制需要特别考虑保护数据的完整性与安全。数值数据库包含的数值量很大，数值数据库的用户面也很广泛，因此数据库的设计应体现保护措施的严密性。例如，美国地质调查局的“水数据存贮与检索系统”从三个级别上采取对数据库的保护办法：①在计算

机主机上，②在进入数据库的入口点上，③在数据内的记录一级上，采取层层设防，防止对数据库的未经允许的使用，限制对数据的检索，预防对库内数值的修改。

数值数据库的编制，尤其需要国家情报政策的指导、全国范围的协调和国际合作。正如前面已经指出的，数值数据库的编制是一个国家科学基础结构中的一个重要部分，加之数值数据的获得与鉴定核实需要花费大量的人力物力，数值数据库的编制与利用需要涉及科学界、信息处理与计算机界、及图书馆部门的通力合作。作为国家一级的协调机构是完全必要的。某些数值数据，如地球物理、气候及环境科学等领域的数值数据，其收集编纂工作还需全球范围的合作。然而另一方面，涉及国家利益与安全的一些敏感的数据却需要保密和控制。有关数值数据库编制与利用的国家情报政策是必不可少的。

四、数值数据库情报服务的发展道路

由于数值数据库是以最终用户(即本身提出情报需求的科技人员和规划管理人员)为服务对象与市场推销对象，而不是以图书馆为对象的，因此，数值数据库的生产与情报检索服务，在其历史传统上保持与书目数据库的平行发展道路。这种数据库的情报服务，多数是由所谓"分时公司"(timesharing firms)经营的。这些公司一贯从事数据处理、为顾客提供计算服务。近二十年来，它们开发了各种应用软件，开展咨询服务，包括数值数据库的检索服务。在这些公司看来，这种检索服务只是为顾客设立的一种附属性质的服务，而不是一个独立的部分。

数值数据库与书目数据库的服务在发展过程中的上述区别，造成了互相脱节的情况。图书馆与情报服务部门对数值数据库及其服务比较陌生。尽管图书情报人员经常使用数据手册与统计年鉴之类的出版物作为参考咨询的工具，但对于计算机化的数值数据库却没有表现出他们对书目数据库那样的接受与热情。这里的原因可能是，数值数据库的使用更依赖于各学科的专业知识，图书情报人员在对于这类数据的解说与分析方面显得不那么得心应手。这类数据库的检索语言也往往采用有别于书目数据库所采用的"常规"检索语言的特殊检索语言。

另一方面，数值数据库的经营者也处于图书馆与情报服务的主流之外。他们自己组织协会，自成系统。

但是，对情报用户来说，其目标往往是既想获得二次情报(书目情报)、又想获得一次情报(包括原始数据)。无论是书目数据库还是数值数据库，对他们来说都是取得所需情报的来源。对这两种数据库的利用渠道没有必要进行严格的区别。这两种数据库及其服务在互相独立的轨道上并行发展，对用户来说无疑造成了人为的障碍。用户要求打破人为障碍。也正因为如此，近年来若干经营书目数据库联机服务的系统，如美国DIALOG，BRS，日本科技情报中心联机系统，法国 Telesystem Questel 等已把包括数值数据库在内的"非书目"数据库(此外还有全文数据库)纳入自己的服务范围。例如，DIALOG 在1982年新增加的 43 种数据库中，至少有12种被认为是数值数据库。目前，仅商业与经济方面，DIALOG 提供了诸如"美国市场情报索引"、"美国统计索引"、"国际商业预测"、"国际商情报道"、"世界经济与人口报告"等等数值数据库的检索服务。因此，书目数据库服务与数值数据库服务正在逐步结合，形成一体化的情报检索服务。看来这种发展趋向是合乎逻辑的。

与此相适应，在情报科学领域内，对包括数值数据库在内的非书目数据库及其服务的承认，也是一个逐步演化的过程。1975年美国情报科学学会成立了数值数据库组。《情报科学年

度评论》(Annual Review of Information Science and Technology) 1977年卷发表了"数值数据库及其系统"的文章，成为情报科学对数值数据库研究的发展过程中的一个里程碑。

事实上，为最终用户提供数值数据库的服务，正在纳入图书馆和情报单位的参考工作范围之内。这是对传统的参考服务的一种延伸。虽然某些数值数据库是相应印刷版数据手册之类出版物的重复，但是更多的是扩展了甚至取代了印刷版的数据手册。数值数据库即使是重复印刷版的出版物，但它要比后者更加灵活，使用起来更为快速和准确。看来，作为联机检索中间人的图书情报人员，为最终用户既进行文献检索，又进行数值数据检索，使两者相结合而成为完整意义上的情报检索服务，将是确定无疑的趋势。

五、数值数据库检索系统的功能

为了满足情报用户对于数值数据库检索与利用的要求，计算机检索系统除了提供一般的联机检索功能外，还应该具有下列基本的功能：

1. 准确的数值数据处理能力，如运算、查找、限定范围的检索、输入编辑等。例如，可以要求系统只查找一定年代范围的经济时列；可以要求系统根据历史增长率的模型为基础来预测某些数值；可以要求系统以某年作为基础年，重新确定时列的指数；可以要求系统进行不同计量单位的换算；可以要求系统在计算时精确到小数点之后多少位，等等。

2. 交互能力。如学习、敦促用户、允许用户转储以利用私人文档和软件包。

3. 图形处理能力。例如可以要求系统提供分子结构或晶体结构图形，要求系统能处理非线性座标，等等。

4. 分析能力，包括进行模型化的能力。

数值数据库的研制往往带有特定的应用目的与用户对象。对于不同类型的数据，采用不同的处理标准是最为有效的。因此当一些不同的专门化数据库出现时，它们一般是不能彼此兼容的。这些不同类型数据库的检索服务，往往采用不同的计算机，使用为专门处理某种单独数据而研制的软件包。因此这种情况给用户带来接口和存取方面的问题。它阻碍了数据利用的进步，也阻碍了整个科学的进步。如果有人想利用20种数据库，就必须学会用20种不同的方法来对付这些数据库，否则就不能有效地进行检索与利用。

因此，数值数据库情报服务的一个发展趋势是，从过去的单数据库的、单机处理的、少量专业用户存取的系统，正在发展成把一批数值数据库集合在一起，用一个软件包，把几个程序连接起来，向用户提供单一接口语言的所谓"数据系统"(Data system)。数据系统可以使用户通过一个计算机系统存取两个或更多个的数值数据库，以满足那些想同时获得多种类型数据的用户需求。例如，分析化学家可能希望得到某种化学物质的融点、颜色、晶体结构及 C^{13} 核磁共振谱等等的数据。这些本来可能分属于多个数据库的信息，现在却能够在数据系统中集合在一起。数据系统的建立，正在克服数据库中对物质命名的不一致性和数据格式的不一致性所造成的困难。要求同一事物在不同数据库中命名的统一性；制定统一的标准格式，使各种不同数据库转换成数据系统所要求的统一格式。国际原子能局和各国核数据中心使用的 EXFOR 系统就是用这种办法来成功解决数值数据格式统一性的一个例子。然而，许多科学数值数据在格式上的变化是如此之大，很难想象有一种普遍的格式能适应各种不同的情况。因此解决数据格式问题的第二种办法，就是编制不同的软件以适应不同数据的格式，并实现综合处理。例如美国的NIH/EPA化学情报系统的作法便是一个例子。

这种数据系统已取得令人鼓舞的进展。例如美国化学情报系统(CIS)，可以允许用户利用一个远程终端，同系统连接一次就能存取一批独立的数值数据库。由于研制出相应的软件系统，用户可用单一的方式询问系统，以存取下列各种数据库：NBS质谱、NBS单晶文档、粉末衍射联合委员会的粉末衍射模型、美国国立职业安全与卫生研究所的化学物质毒性效应登记，以及美国化学文摘社登记号等等文档(或数据库)。这个系统可提供一定的数据分析程序和数学模型化方法。尽管数据系统的软件系统的研制是十分复杂的，但证明是可行的。数据系统又称为"超级市场系统"。它的出现与发展，无疑大大提高了数值数据库的计算机系统的功能。

六、国内外主要的数值数据库及其服务

目前在国外，许多学科专业领域出现了数值数据库。在社会科学领域，有财政统计、人口调查、选举结果、劳动、民意测验、国内暴力行为等数据库。在自然科学领域，有化学结构、谱、结晶、环境、资源、材料、实验室动物、死亡记录等数值数据库。究竟目前有多少数值数据库，无从统计，但其数量至少是以万计的，绝大部分是"内部"性质的。

以商业渠道提供服务的联机数值数据库，极大部分是商业与经济方面的。这是由于这类数值数据库有较广泛的使用面。其它的数据库往往是由实验室或工业部门研制的，服务于较小的用户界。在整个数值数据库中，大部分是以批处理方式进行存取的，以联机方式提供服务的只占小部分。在 Cuadra Assoiates 公司编的《联机数据库指南》(Directory of Online Database)中列有商业渠道提供服务的联机数值数据库的目录。《联机评论》(Online Review)杂志也定期公布新增加的非书目数据库的目录。

国外有许多提供数值数据库检索服务的系统。例如美国地质局"地球科学情报系统"(ESIS)、美国Baffelle的数值数据情报分析中心的有关工程技术的数值数据库系统。此外还有下列系统：如ADP(美国，财政、人口统计，外汇汇率、经济学、化学品)；Business International 公司(美国，70多个国家的经济与市场活动)；Data Resources 公司 (DRI) (美国，经济、财政、能源、天气)；IP Sharp(加拿大，71个国家的经济与人口统计、飞机意外事故)；GEISCO(美国，美国及其它九十多个国家的矿物资源)；SIA(英国，欧洲共同体国家的旅行和运输、核电站的建设、139个国家的财政信息)；Chemical Information System (美国，化学化工)；GIDSFT(西德，自然科学)；F·A·C·T(加拿大，自然科学)；Thermodata(法国，自然科学)，等等。

我国在八十年代也开始了数值数据库的研制工作，如石油开发、石油地质勘探、化学、生物资源、稀土材料、土壤信息、市场技术动态等方面数值数据库都已经和正在进行。我国已参加了国际科技数据委员会(CODATA)，成立了该委员会之下的中国委员会。中国科学院成立了数据库办公室。我们相信在不久的将来，我国也会出现数值数据库及其情报服务的繁荣。

参考文献

① Ching-chih Chen and Peter Hernon, Nummeric Databases, Ablex Publishing Corporation, 1984。

② Judith Wanger and Ruth N. Landau, Nonbibliographic Online database services, Journal of the American Society for Information Science, 31(3), May 1980, P. 171—180。

③ 龚国伟：《联机检索策略》，1986年，湖北科技出版社出版。

④ 陈光祚：《国外情报存贮与检索的若干发展动向》，载《现代图书情报技术》1986年第4期。