第一章 绪 论

案例导入

某海军陆战队在原始森林进行为期一个月的生存训练，队员们被飞机空投到半径为1 000公里的原始森林的中心地带，要求他们在一个月内走出原始森林，而且每个队员除了身上穿的衣服外，随身只能携带3件物品，每件物品的重量不能超过2公斤。请问队员应该携带哪三件物品最合适？

在这个案例中，理想的答案应该是钢刀、火石和指南针。因为钢刀能够用来获取猎物；火石可以用来取火；指南针可以用来指明方向。物质、能量和信息是人类在世界上生存与发展所必须具备的三大要素。本案例中的钢刀和火石可以帮助海军陆战队员获取生存所需的物质与能量，指南针则是获取信息的工具，有了它，队员们就可以找到走出原始森林的方向，没有它，可能一辈子都走不出原始森林。指南针所指的方向就是一种信息，即方位信息。

第一节 信息、知识、情报、文献及其关系

一、信息（Information）

“信息”一词早在我国古代文献中就已出现过。然而究竟什么是信息？至今仍然没有一个为各界所普遍认同、统一的定义。这主要是由于信息所包含的因素以及认识层次上的差别所造成的。具体说来主要有三点：第一，信息本身呈现出多元化、多层次、多功能的特点，是一个高度复杂的综合体；第二，随着社会、经济和科学技术的不断发展，信息科学及其许多分支学科也随之发展变化，人们很难对其内涵和外延有很确切的把握；第三，由于不同的人有不同的研究和使用目的，因此，从不同的角度出发，对信息概念就会做出不同的阐述。

从本体论层次讲，信息是一种客观存在的现象，是事物的运动状态及其变化方式的表征，不受主体意志的影响；从认识论层次讲，信息是主体所感知或所表述的事物运动状态及其变化方式，是反映出来的客观事物的属性。同一个事物的运动状态和方式对于不同的观察者来说，结果往往不同，有的可以从中获取大量的信息，有的则获取很少的信息，甚至一无所获。

信息的概念是十分广泛的。可以说，世间万物的运动，人间万象的更迭，都离不开信息的作用。神话传说中，诺亚方舟在洪水中飘荡许久，当放出的飞鸽衔回一束橄榄时，它带回的是洪水消退的信息；苏东坡的“大江东去，浪淘尽，千古风流人物。……”传递给人们的则是赤壁怀古的信息；在自然界，山的高度是一种信息，它反映出山的空间特

性；花的香味也是一种信息，它反映了花分子结构的化学特性等。

因此，信息是客观存在着的一切事物通过一定媒介与形式而向外传播、展示或表现的一种迹象、征兆、信号或消息，是自然界、人类社会和人类思维活动中普遍存在的一切物质和事物的基本属性，是物质存在方式及其运动规律特点的外在表现。信息有多种类型，如自然信息、生物信息、社会信息等。

二、知识（Knowledge）

知识是人类主观世界对自然界、人类社会以及思维活动规律的概括和反映，是大量有组织的序列化的信息，是人脑通过思维重新组合的系统化信息的集合，即经过理性化、优化和系统化的信息可称之为知识。

从知识的定义可以看出，知识是大脑通过对信息进行积累、整理和加工而得出的系统化规律、概念、经验，是人类在社会实践中逐步形成的，是对客观现实的反映，它可以通过信息载体或媒介（网络、文献、电视、广播、人等）的传递和交流而促进人类文明进步和社会的全面发展。

三、情报（Information）

在我国科技情报界，对什么是“情报”这个问题一直众说纷纭，莫衷一是。1983 年 7 月，著名科学家钱学森在国防科技情报工作会议上作的“科技情报工作的科学技术”报告中，对“情报”这一概念作了科学而精辟的概括：“情报就是为了解决一个特定的问题所需要的知识。这里包含了两个概念，首先它是知识，不是假的、乱猜的，应该是知识；其次，它是为特定的要求，也就是为了特定的问题，所以，及时性和针对性是非常重要的，人家问的是这个问题，你回答的是另外一个问题，那当然也不行。”钱学森的这一论述，是国内外对情报概念最明确的论述，具有普遍意义，它既明确了情报的科学属性（属于知识范畴），又明确了情报的功能定位（是为解决特定问题服务的）。

情报具有知识性、传递性和效用性三种基本属性。其中知识性指情报是经过加工并为用户所需要的特定知识或信息；传递性指情报是处于动态接受与利用过程中的知识，未经过传递使用的知识不能称之为情报；效用性指情报是激活了、活化了的知识，能够用来启迪人类思维，帮助人类增长见识，改变知识结构，提高认识能力，发挥实用的、社会的、经济的价值。

四、文献（Literature，Document）

文献是用文字、图形、符号、声频、视频等手段记录有人类知识的一切载体。它是固化在一定物质载体上的知识，是记录、积累、传播和继承知识的最有效手段，是人类社会活动中获取情报的最基本、最主要的来源，也是交流、传播情报的最基本手段。文献有三个最基本的要素：一是构成文献内核的知识信息；二是负载知识信息的物质载体；三是记录知识信息的符号和技术即手段。

文献是人类知识宝库的主要组成部分，是人类的共同财富，它在人类科技和社会发展中表现出了十分重要的作用：既是科学研究和技术研究结果的最终表现形式，也是在空间、时间上传播情报的最佳手段，更是促进人类进行研究活动的重要激励因素。

五、信息、知识、情报与文献的关系

知识是人的大脑经过思维加工而形成的有序化信息的集合，是一种信息产品，是信息的一部分；文献则是记录有人类知识的一切载体，是知识的一部分；情报是人们为解决特定问题而被活化了的知识，是知识的一部分，情报也蕴含在文献之中，但不是所有文献都是情报，因而信息、知识、文献、情报之间是一种包含与被包含的关系。

第二节 信息获取与信息素养

案例1-2-1

大庆油田是我国20世纪60年代勘探、开发的大型油田。但在当时，大庆油田的具体位置、油田规模及产油量等都是严格保密的。1960年，大庆油田的石油工人王进喜被誉为“铁人”后，国内各大报刊都对他的先进事迹进行了相关报道。但谁能想到，著名的日本商社三菱重工集团的情报专家却从这些公开的信息报道中嗅出了蛛丝马迹，准确掌握了大庆油田的情况。一是关于油田的开采时间，日本人发现，王进喜原来的工作地点是甘肃玉门油田，1959年10月参加国庆观礼后他就销声匿迹了，由此推断大庆油田开发开采时间应该为1959年9月。二是关于油田位置的测算，1966年7月，《中国画报》封面刊出这样一张照片：王进喜头戴狗皮帽，身穿厚棉袄，顶着鹅毛大雪，手握钻机刹把，眺望远方，在他背景远处矗立着星星点点的高大井架。凭着对中国地理的熟知，日本人很快就推断出王进喜的工作地应该在零下30度的东北地区；并根据运送原油的列车上灰尘的厚度，测算出了油田与北京的距离，断定油田应在哈尔滨与齐齐哈尔之间；10月份，《人民中国》也刊登出宣传王进喜的文章，其中提到了一个地名——“马家窑”，综合这些信息，利用日伪旧满洲地图，日本人终于把大庆油田的地理位置搞清楚了。三是关于油田的规模，他们从照片中王进喜所站的钻台上手柄的架式，推算出油井的直径，又从钻台油井与他背后隐藏的油井之间的距离和密度，推算出油田的大致储量和产量。他们还根据炼油厂反应塔的照片，推算出大庆油田的炼油能力。就这样，日本人凭借良好的信息素养，通过精细、准确的信息情报获取，对大庆油田进行了成功的调查分析，并预测中国今后几年必然因为炼油设备不足，会考虑大量引进炼油设备。果然，中国政府不久向世界市场寻求石油开采设备，三菱重工集团以最快的速度和最符合中国要求的设计及设备获得巨额订单，几乎垄断了我国石油设备进口市场。此时，西方石油工业大国还未回过味来呢。

一、信息获取的必要性

当今社会已经是一个信息化生存的时代，掌握信息获取的方法和手段不仅是每一位公民必备的生存手段与技能，更是开展科学研究的向导和获取有效新知识的捷径。20世纪70年代，美国核专家泰勒收到一份题为《制造核弹的方法》的报告，他被报告精湛的设计技术所吸引，惊叹地说：“至今我看到的报告中，它是最详细、最全面的一份”。但使

他更为惊异的是，这份报告竟出自于哈佛大学经济学专业的青年学生之手，而这份四百多页的技术报告的全部信息来源又都是从图书馆那些极为平常的、完全公开的图书资料中所获得的。

科学技术发展到今天，传统教育模式培养出的知识型人才已满足不了现代社会发展的要求，只有具备自学能力和独立研究能力的创造型人才才能适应现代信息化社会的进步与发展。虽然，大学生在校期间已经掌握了一定的基础知识和专业知识，但是“授之以鱼”只能让其享用一时。如果能熟练地掌握信息获取技术方法便可以无师自通，找到一条吸收和利用大量新知识的捷径，就可以在更广阔的知识领域中自由翱翔，对未知世界进行探索，是谓“教人以渔”，才能受益终身。

二、信息素养及其形成

“信息素养”（Information Literacy）一词首先由美国信息产业协会主席 Paul Zurkowski 于1974年提出，具体内容为：“信息素养就是利用大量的信息工具及主要信息资源使问题得到解答的技术和技能”。其后，随着人们对信息素养研究的不断深入，对信息素养的界定也说法不一。在我国，信息素养通常被称为信息素质，它是个体（人）对信息活动的态度以及对信息的获取、分析、加工、评价、创新、传播等方面的能力，它是一种对目前任务需要什么样的信息、在何处获取信息、如何获取信息、如何加工信息、如何传播信息的意识和能力。信息素养主要包括信息意识、信息知识、信息能力、信息道德四个方面。

信息意识是整个信息素养的前提，指的是个体对信息的敏感度。这要求个体具有敏锐的感受力和持久的注意力，能够意识到信息的作用，对信息有积极的内在需求。

信息知识是个体具有信息素养的基础，指的是对信息学的了解和对信源以及信息工具等知识的掌握。

信息能力是整个信息素养的核心。从狭义上来说，指的是个体对信息系统的使用以及获取、分析、加工、评价、传递信息的能力；从广义上来讲，除了上述能力以外，还应该包括语言能力、思维能力、观察能力、判断能力等间接能力。

信息道德是把握个体信息素养的方向，指的是个体在获取、利用、加工和传播信息的过程中必须遵守一定的伦理规范，不得危害社会或侵犯他人的合法权益。因此，无论个体的信息意识如何强烈，信息知识如何丰富，信息能力如何强大，如果他将其才能用在违法犯罪上，那么他的信息素养是非常低下的。

信息素养是大学生终身学习的基础，信息意识的强弱决定了大学生能否有效而及时地获取信息，能否对所做的研究进行扩展，能否对自身的学习进行更有效的控制，使自己具备终身学习能力、竞争能力和创新能力。信息知识不仅体现了大学生所具有的信息知识的丰富程度，而且还制约着他们对信息知识的进一步掌握。信息能力的培养则提高了大学生独立自主学习的态度和方法，使之具有追求新信息、运用新信息的意识和能力，善于运用科学的方法，从瞬息万变的事物中捕捉信息，从易被人忽视的现象中引申、创造新信息的能力。信息道德培养使大学生具有正确的信息伦理道德修养，让学生学会对媒体信息进行判断和选择，自觉地选择对学习、生活有用的内容，自觉抵制不健康的内容，不组织和参与非法活动，不利用计算机网络从事危害他人信息系统和网络安全、侵犯他人合法权益的活动。

信息素养引起了世界各国越来越广泛的重视，并逐渐成为评价人才综合素质的一项重要指标。信息素养的形成不但是人们生存于信息时代的当务之急，更是实现终身学习的必然途径。信息素养的形成可以从以下几方面入手：提升教师的教育观念和自身信息驾驭能力；创设信息教育的良好环境和氛围；紧紧围绕学习的问题，从广泛的信息资源库特别是网络资源独立或协作获取所需信息，并对信息进行分析判断、加工处理。通过交流讨论，在原有信息的基础上再进行信息创新、信息运用；养成正确的思维方法和培养创新精神与创新能力。

总之，信息素养是信息社会人的整体素养的一部分。信息素养的教育关系到人们如何立足于信息化社会这一基本点。它不是所谓的超前教育观，而是教育界必须要面对的现实问题。只有加强信息素养的教育，教育的职能才会充分发挥作用。对于学生信息素养的培养不是短时间内、几个人就可以完成的，需要大量的时间以及人力物力，更需要教师们首先具有这种信息素养。信息技术教育的路需要我们在实践工作中去开辟，信息素养的形成，需要学生与教师的共同努力。

第三节 信息资源的类型与特点

人们通常使用的信息资源有传统信息资源和网络信息资源，传统信息资源（印刷型文献）主要有期刊、图书、报纸、学位论文、会议文献、专利文献等。上述传统信息资源经数字化加工后上网，就成为网络信息资源。那么，信息资源究竟应该如何进行分类？信息资源又有哪些特点呢？

信息同能源、材料并列为当今世界三大资源。“信息资源”（information resource）这一名词最早是由美国一批经济学家提出来的。广义的信息资源包括信息内容以及提供信息的设施、设备、组织、人员和资金等，即信息源及其他与之有关的各种资源的总和，是人类社会信息活动中积累起来的信息、信息生产者、信息技术等信息活动要素的集合。狭义的信息资源是指信息内容构成的信息有序化集合，是人类社会经济活动和科技活动中加工、处理、有序化并积累的有用信息的集合，如科学技术信息、政策法规信息、经济信息、金融信息、市场信息、社会发展信息等，亦即信息载体及其信息内容的集合。2009年出版的《医学图书馆学》一书对信息资源的解释是：信息资源就是经过人类筛选、组织、加工并可以存取和能够满足人类需求的各种媒介信息的集合。从这个定义来看，信息资源包括了信息内容、信息载体（媒介）、信息活动（筛选、组织、加工）和信息价值（满足人类需求）等诸要素。信息资源广泛存在于经济、社会各个领域和部门，是各种事物形态、内在规律和其他事物联系等各种条件、关系的反映。随着社会的不断发展，信息资源对国家和民族的发展，对人们工作、生活至关重要，成为国民经济和社会发展的重要战略资源，它的开发和利用是整个信息化体系的核心内容。

一、信息资源的类型

对事物进行分类，是人类认识事物的基本方法。对信息资源的认识也是如此，要方便快捷地获取信息资源，首先必须了解信息资源的类型。目前国内信息资源管理专家和相关学者提出了20多种有关信息资源划分的标准与划分方法（方案），以下是比较常见的信

息资源的分类方法和类型。

（一）依据信息资源的载体分类

可分为体载信息资源、文献信息资源、实物信息资源和网络信息资源。

1. 体载信息资源

体载信息资源是指以人体为载体并能为他人识别的信息资源。按其表述方式又可分为口语信息资源和体语信息资源。口语信息资源是人类以口头语言表述出来但被记录下来的信息资源。如谈话、授课、讲演、讨论、唱歌等。体语信息资源是以人的体态表述出来的信息资源。如表情、手势、姿态、舞蹈等。

2. 文献信息资源

文献信息资源是以文献为载体的信息资源。文献信息资源依据其记录方式和载体材料又可分为刻写型、印刷型、缩微型、机读型、声像型等五大类。

3. 实物信息资源

实物信息资源是指以实物为载体的信息资源。依据实物的人工与天然特性又可将实物信息资源分为以自然物质为载体的天然实物信息资源和以人工实物为载体的人工实物信息资源（如产品、样品、样机、模型、雕塑等）。

4. 网络信息资源

网络信息资源是指计算机技术、通信技术、多媒体技术相互融合而形成的网络上可查找到的资源。网络信息资源可分为：①非正式出版信息。指流动性、随意性较强、信息量大、信息质量难以保证和控制的动态性信息。如电子邮件、专题讨论小组和论坛、电子会议、电子布告板等工具上的信息，以及博客（Blog）、微博客（Miniblog）上发表的信息等。②半正式出版信息。又称“灰色”信息，是指受到一定产权保护但没有纳入正式出版信息系统中的信息。如各种学术团体和教育机构、企业和商业部门、国际组织和政府机构、行业协会等单位介绍宣传自己或其产品的描述性信息。③正式出版信息。指受到一定的产权保护，信息质量可靠，利用率较高的知识性、分析性信息，用户一般可通过万维网查询到。如各种网络数据库、联机杂志和电子杂志、电子图书、电子报纸等。

（二）依据信息资源的存在状态分类

依据信息资源的存在状态分类可分为潜在的信息资源和现实的信息资源。潜在的信息资源是指个人在学习、认知和实践过程中储存在大脑中的信息资源，其特点是只能供个人所用。现实信息资源是人类获取并表述出来的，能够为公众所利用的信息资源，如文献型信息资源、实物型信息资源、网络信息资源。

（三）依据信息资源的内容和用途分类

1. 数据库资源

数据库资源包括：文献型数据库（题录文摘数据库和全文数据库）、数值事实型数据库（主要包括基因库、核酸序列、蛋白质结构库等分子生物学数据库，以及毒理学、药物方面的事实型数据库）、多媒体数据库（包括化学物质或药物三维立体结构数据库、各种医学图谱库、医学影像库（X线片、CT片、核磁共振图像）、病理切片库等）。

2. 会议信息资源

会议信息资源是指与会议召开相关的一些信息资源，如会前信息通报、会议征文通知、大会报告、大会发言及大会交流资料以及会后编辑出版的会议论文集等。会议信息资

源是人们快速获取信息的一种比较重要的信息来源。

3. 专利信息资源

专利信息资源指与专利申请相关的信息资源，如专利说明书、专利全文等。

4. 电子出版物资源

互联网提供的电子出版物包括电子期刊、电子报纸、电子图书、手册、法规、指南、图谱、百科全书等。其特点是成本低廉、发行速度快、发行面广、更新速度快、检索功能强。

5. 网络新闻资源

在网络上发布的各种时事新闻、天气信息、世界大事要闻等。

6. 网络教育信息资源

网络上发布的与教育教学有关的信息资源，如学习网站、在线信息素质教育课程、多媒体学习视频、在线考试系统和精品课程网站等。

7. 市场信息资源

市场信息资源是指反映市场供求情况的相关信息，如医疗器械需求信息、药品供应信息、医务人员招聘信息等。

8. 软件资源

可以利用的统计软件、计算软件包和文献管理软件等资源。

9. 联机馆藏目录

联机馆藏目录是揭示馆藏内容、提供馆际互借、资源共享的现代化服务手段，互联网上已有 600 多个著名公共图书馆、大学图书馆，400 多个学术机构的馆藏目录通过网络向公众开放，并通过 OPAC 提供服务。

10. 其他信息资源

主要有产品样本、技术标准、学位论文资源和基金申请信息、求职信息等。

（四）依据信息资源加工的层次和深度分类

1. 零次信息资源

零次信息的载体形式及其信息内容之和称为零次信息资源，是指未经正式发表或未形成正规载体的一种文献形式。如：书信、手稿、会议记录、笔记、实验记录等。内容新颖，具有原始性、客观性、零散性、不成熟性的特点，难于检索和获取。零次文献一般是通过口头交谈、参观展览、参加报告会等途径获取，不仅在内容上有一定的价值，而且能弥补一般公开文献从信息的客观形成到公开传播之间费时甚多的弊病。

2. 一次信息资源

一次信息的载体形式及其信息内容之和称为一次信息资源，有时也称为原始文献。它是以作者本人的工作经验、观察或者实际研究成果为依据而创作的具有一定发明创造和一定新见解的原始文献，如期刊论文、研究报告、专利说明书、会议论文、学位论文、科技报告、技术标准等，内容先进、新颖，叙述具体、详尽，数量庞大、分散。具有新颖性、创造性和系统性等特点，理论上比较成熟可靠，有较高的参考利用价值。

3. 二次信息资源

二次信息的载体形式及其信息内容之和称为二次信息资源，它把分散无序的一次文献筛选后，按其内容特征和外表特征进行加工编辑而成的系统化有序文献体系，目的是方便

查找和利用，如书目、索引、文摘等，具有浓缩性、汇集性、系统性、有序性等特点，不仅可以报道信息的内容，更重要的是可以提供一次信息资源的线索，以方便人们获取全文文献。这类信息资源在信息获取时使用频率较高。

4. 三次信息资源

三次信息的载体形式及其信息内容之和称为三次信息资源，它是对一次信息资源进行综合分析、研究和评述而编写出来的成果，如手册、百科全书、年鉴及其他综述和述评性文章，三次信息资源源于一次信息资源，又高于一次信息资源，是一种再创造资源。具有浓缩性、指引性、针对性、参考性、系统性、综合性、知识性、概括性等特点，人们通常所称的医学综述就是一种非常重要的三次信息资源。

信息资源是一个发展着的有机体。信息资源的类型也不是一成不变的，随着科学技术的发展，新的信息资源类型将不断涌现，科学的信息资源分类体系应及时吸纳、涵盖这些新兴类型。另外，随着信息资源内涵与外延的深化、拓展，信息资源的分类标准与分类方法也可能随时发生新变化，信息资源类型体系亦应及时地予以调整，以保持信息资源类型与其定义的一致性。

二、信息资源的特点

信息资源是人类自身挖掘、创造、存取与积累而成的一种社会智力资源。信息资源与自然资源、物质资源相比，具有以下几个特点：能够重复使用，其价值在使用中得到体现；信息资源的利用具有很强的目标导向，不同的信息在不同的用户中体现不同的价值；具有整合性，人们对其检索和利用，不受时间、空间、语言、地域和行业的制约；它是社会财富，任何人无权全部或永久买下信息的使用权；它是商品，可以被销售、贸易和交换，具有流动性。可见，信息资源具有综合性、增长性、传播性、共享性和效用性（具有开发利用和价值转化性）。作为信息资源重要组成部分的医学信息资源，还具有其他一些特点。

（一）数量巨大、来源广泛

（1）网络信息资源数量有海量之称，每时每刻都在不断增加。有人估计，Internet 每天的信息流量达万亿比特，全网提供的信息在 20TB 以上，WWW 网址每 6 个月增长 1 倍，仅在 Yahoo 搜索引擎的医药卫生（health）栏目下就有 2 万 6 千多个信息节点。Google 中文搜索含有医学或药学的网页就有 7 100 万页，含有 Medicine 一词的网页有25 300万页。Google 数据库拥有的网页总数在 100 亿页以上。

（2）联合国教科文组织（UNESCO）统计信息生产量 2 000 印张/秒。

（3）全世界最常用的连续出版物有 21 万种（乌里希国际期刊目录）。其中生物医学约占 30% ，每年至少刊登 600 万篇学术论文。

（4）2010 年 2 月 25 日，新华社授权发布《中华人民共和国 2009 年国民经济和社会发展统计公报》。公报显示，2009 年全国出版各类报纸 437 亿份，各类期刊 31 亿册，图书 70 亿册（张）。

（5）信息来源广泛而复杂，有政府的、研究机构的、大学的、学会的、企业的、个人的，造成了数据质量参差不齐，信息的广度和深度不一，而且相互交错，也有许多与信息本身无关的、多余的、重复的，甚至无用的信息。但专业数据库还是很受欢迎的，使用

率较高的主要是美国的专业数据库。数据库类型与数量众多，有动态信息，如政府机构发布的信息、政策法规、会议消息等；有电子期刊，仅生物医学就有几千种，其中相当一部分可免费提供全文或摘要；网上有600多个公共图书馆和大学图书馆、400多个学术机构图书馆的联机馆藏目录；有各种专业数据库，如DNA序列库、蛋白质序列库、核苷酸数据库、代谢库、PubMed及各临床学科的专业数据库。

（二）内容全面、质量参差不齐

（1）信息载体多样化：包括印刷型、缩微型、声像型、机读型等。

（2）信息内容包罗万象：医学信息涉及整个医学领域，既有基础医学、临床医学信息，又有预防医学、药学、生物医学材料与工程、遗传学、免疫学等专业信息。

（3）随意性和自由度很大，内容良莠不分，质量参差不齐：正式出版物与非正式出版物交织在一起，科技、学术、商务及个人信息与一些色情、暴力、种族歧视、反科学、反人道等污染信息混为一体；既有高水平的研究成果，又有许多不雅之作和虚假信息。

（三）传播迅速、更新频率不一

信息可以随时发布、随时修改或删除。传播速度快，范围广，影响大。更新频率没有统一规定，半衰期缩短。

（四）形式多样、类型齐全

包括了原始论文、电子报刊等一次文献信息；文摘、题录、索引等二次文献；综述、述评等三次文献信息；网上医学会议、聊天等零次文献信息。信息类型齐全，有文本、表格、图形、声音、图像、程序软件、超文本、多媒体等信息。

（五）分散无序、缺乏组织

Internet是全球分布式网络，信息散布于各个国家的主机与服务器上，但主要集中在美国、加拿大、澳大利亚、日本等国家和西欧地区，80%～90%的节点位于英语国家，其中美国的数据量最大。网络信息资源没有一个中心点，分散在世界各地，而且处于无序状态。许多信息资源只是时间序列的堆积，缺乏组织加工和整序。组织形式松散，有规范化文本，也有非规范化文本，任何人均可在网上发布信息。对于网络医学信息资源，要经过精心筛选、仔细甄别和全面评价后才能使用。

（六）电子资源将逐步取代纸质文献

一个比较明显的特征就是目前世界一些著名大学图书馆在订购纸质期刊的费用和种数上呈线性递减，而电子信息资源则呈指数化增长。以哈佛大学医学院图书馆为例，目前该馆订购的纸质期刊已由1 800种骤减至400种。而该馆的电子期刊则达到12 000余种。电子资源正在成为图书馆的核心馆藏和利用率最高的信息资源。

此外，国内外有关印刷型的检索工具已逐步被光盘和数据库取代。如：印刷版《中文科技资料目录：医药卫生》、《中文科技资料目录：中草药》已被中国生物医学文献数据库（CBM）取代，中国医学文摘系列已被《中国期刊全文数据库》（CJFD）取代。《Index Medicus》（IM，美国医学索引）已被MEDLINE光盘数据库和PubMed网络数据库取代，《Excerpta Medica》（EM，荷兰医学文摘）被EMBASE取代，《Chemical Abstracts》（CA，美国化学文摘）被CA on CD取代，《Biological Abstracts》（BA，美国生物学文摘）被BIOSIS Proviews取代，《Science Citation Index》（SCI，美国科学引文索引）被Web of Science取代。

（七）分布的不均匀性

各国医学信息资源在地域的分布与国家的发达程度和信息化水平基本一致，即一个国家的科学技术发展水平越高、信息化程度越高，其拥有的医学信息资源就越丰富。目前美国、西欧、日本等发达国家和地区拥有的医学信息资源比较多，而亚洲、非洲、拉丁美洲的发展中国家和落后国家的拥有量相对较少，这体现了医学信息资源在地域分布上的不均匀性。

（八）获取的复杂性

获取医学信息资源不仅难度大，而且复杂程度高，特别是循证医学信息资源的获取更是如此，因此需要花费更多的时间和精力，需要掌握娴熟的方法和技巧。

三、信息资源管理与开发利用

信息资源管理经历了三个主要阶段：传统管理阶段（20 世纪 50 年代 ～ 70 年代，以图书馆、情报所为代表的文字信息资源管理）、信息管理阶段（20 世纪 70 年代末 ～ 20 世纪末，以计算机应用和数据处理为典型代表）、信息资源管理阶段（21 世纪初 ～ 未来 20 年，以网络平台、海量数据库、信息处理技术为代表，信息交换、信息共享、信息应用为内容，视信息资源为主要经济资源进行管理的信息资源管理）。

信息资源是无限的、可再生的、可共享的。其开发利用会大大减少材料和能源的消耗，减少污染。人类和地球所在的宇宙在其存在的无限时间和无限空间内，生成了海量的物质、能量和信息。人类在其存在的有限时间和有限空间内，消耗了大量的物质和能源，也生成了大量的信息。大力推动信息资源开发利用，要以需求牵引，与信息化应用相结合，特别要注重实效。

（1）发布和实施与国家信息资源开发利用相关的法规，制定相应的规划，加强信息资源开发利用的统筹管理，规范信息服务市场行为，促进信息资源共享。

（2）积极开展试点示范工程，在国民经济和社会各领域广泛利用信息资源，促进信息资源转化为社会生产力。

（3）建设若干个国家级数据交换服务中心和一批国家级大型数据库，形成支撑政府决策和社会服务的基础资源。

（4）加大中文信息资源的开发力度，鼓励上网应用服务，鼓励信息资源的共享。

（5）协调信息资源开发利用标准的制订工作。

第四节　数据库的类型、结构与检索技术

数据库是依照某种数据模型组织起来并存放于二级存储器中的数据集合，这种数据集合具有如下特点：尽可能不重复，以最优方式为某个特定组织的多种应用服务，其数据结构独立于使用它的应用程序，对数据的增、删、改和检索由统一软件进行管理和控制。从发展的历史看，数据库是数据管理的高级阶段，它是由文件管理系统发展起来的。数据库是人们查找信息的重要工具。数据库有哪些类型，数据库结构怎样，数据库又有哪些可以利用的检索技术呢？

一、数据库的类型

对于数据库的分类，比较常用的分类方法是按照被处理数据的类型、所含信息的内容及数据是否关联等标准进行划分。

（一）按照被处理数据的类型划分

1. 模糊数据库

模糊数据库指能够处理模糊数据的数据库。一般的数据库都是以二值逻辑和精确的数据工具为基础的，不能表示许多模糊不清的事情。随着模糊数学理论体系的建立，人们可以用数量来描述模糊事件并能进行模糊运算。这样就可以把不完全性、不确定性、模糊性引入数据库系统中，从而形成模糊数据库。模糊数据库研究主要有两方面，首先是如何在数据库中存放模糊数据；其次是定义各种运算建立模糊数据上的函数。模糊数的表示主要有模糊区间数、模糊中心数、模糊集合数和隶属函数等。

2. 统计数据库

管理统计数据的数据库系统。这类数据库包含有大量的数据记录，但其目的是向用户提供各种统计汇总信息，而不是提供单个记录的信息。

3. 网状数据库

处理以记录类型为结点的网状数据模型的数据库。处理方法是将网状结构分解成若干棵二级树结构，称为系。系类型是两个或两个以上的记录类型之间联系的一种描述。在一个系类型中，有一个记录类型处于主导地位，称为系主记录类型，其他称为成员记录类型。系主和成员之间的联系是一对多的联系。网状数据库的代表是 DBTG 系统。1969 年美国的 CODASYL 组织提出了一份“DBTG 报告”，以后，根据 DBTG 报告实现的系统一般称为 DBTG 系统。现有的网状数据库系统大都是采用 DBTG 方案的。DBTG 系统是典型的三级结构体系：子模式、模式、存储模式。相应的数据定义语言分别称为子模式定义语言 SSDDL、模式定义语言 SDDL、设备介质控制语言 DMCL，另外还有数据操纵语言 DML。

4. 演绎数据库

演绎数据库是指具有演绎推理能力的数据库。一般来说，它用一个数据库管理系统和一个规则管理系统来实现。将推理用的事实数据存放在数据库中，称为外延数据库；用逻辑规则定义要导出的事实，称为内涵数据库。主要研究内容是如何有效地计算逻辑规则推理。具体为：递归查询的优化、规则的一致性维护等。

5. 书目数据库

书目数据库是指存储二次文献信息的数据库，也称二次文献数据库。

6. 目录数据库

目录（Catalog）是以完整的出版单元（如一种图书、一种期刊）为单位，按照一定次序编排的对文献信息进行描述和报道的工具，也称书目。目录对文献的描述比较简单，每条记录的字段主要包括：文献题名、责任者、出版事项、分类号、主题词等。一种出版物经过如此描述后形成一条记录，将所有的记录组织起来就形成了目录，存储目录的数据库即称为目录数据库。

7. 分发数据库

分发数据库是指分发服务器上的数据库，存储用于复制的数据，包括事务、快照作

业、同步状态和复制历史信息。

（二）按照数据库所含信息的内容划分

1. 文献书目数据库（Bibliographic Databases）

文献书目数据库是存储某个领域原始文献的书目，即二次文献数据库，记录内容包括文献的题目、著者、原文出处、文摘、主题词等。大多数是印刷本检索工具的机读版，如美国工程索引数据库（Ei Compendex）、英国科学文摘数据库（INSPEC）、美国化学文摘数据库（CA Search）等。

2. 信息指南数据库（Dictionary Databases）

信息指南数据库主要是记录一些机构、人物、产品、项目简述等事实数据，通过该类数据库可以查到公司、机构的地址、电话、产品目录、研究项目或名人简历等信息。这类数据库也称为事实数据库。

3. 数值型数据库（Numeric Databases）

数值数据库是专门提供以数据形式表示信息的一种源数据库。主要记录科学研究中试验、测量、计算、工程设计、经济分析和工业规划等方面的数据。这类数据库主要包含数值数据，有的也包含文字，文字是用来定义数据所需的最小量的文字，有时称为文本—数值数据库（Textual－numeric Databases）。

4. 全文数据库（Complete Text Databases）

全文数据库是存储文献内容全文或其中主要部分的数据库，简称全文库。它是将经典著作、学术期刊、重要的会议录、法律法规、新闻报道以及百科全书、手册、年鉴等的全部文字和非文字内容转换成计算机可读形式。全文数据库可以解决用户获取一次文献所遇到的困难，能向用户提供一步到位的查找原始文献的信息服务。近年来，全文数据库发展很快，在各类数据库建设中异军突起。据统计，在美国，全文数据库从1985年的28%增加到1995年的52%，其数量是书目型数据库的一倍，而书目型数据库则从57%下降到24%。在我国，已有《中国学术期刊全文数据库》、《书生之家数字图书馆》、《超星数字图书馆》和《中国博士学位论文全文数据库》等全文图书、期刊、学位论文数据库和报纸、会议、专利全文数据库等建成并投入使用。

除了上述四种基本的数据库类型之外，还有多种混合型的数据库形式，如“数值—全文型”数据库，“书目—数值—全文型”数据库等。特别是随着多媒体技术的迅速发展和广泛应用，将图形、图像、文字、动画、声音等多媒体数据结合为一体，并统一进行存取、管理和应用的多媒体数据库已经问世，并受到人们的普遍欢迎。随着超文本、多媒体和光盘驱动器技术的发展和普及，多媒体数据库的数量会越来越多。

（三）按照数据是否关联进行划分

按照数据是否关联可分为：层次数据库、网状数据库、关系数据库。其中层次数据库和网状数据库都是非关系型数据库，非关系型数据库在20世纪70年代至80年代初非常流行，在数据库产品中占据了主导地位，现在已逐渐被关系型数据库取代。在美国等一些国家，由于早期开发的应用系统都是基于层次数据库或网状数据库的，因此，目前仍有层次数据库或网状数据库系统在继续使用。

从上述数据库分类方法来看，按照数据库所含信息的内容划分代表了当前数据库分类的主方向，也是信息获取的重要前提和基础，集中反映了当前信息获取时使用最为频繁的

一系列数据库。事实上，这些数据库都是信息获取的重要工具。

二、数据库的结构

（一）物理结构

数据库的物理结构分三个层次，反映了观察数据库的三种不同角度。

1. 物理数据层

物理数据层是数据库的最内层，是物理存储设备上实际存储的数据的集合。这些数据是原始数据，是用户加工的对象，由内部模式描述的指令操作处理的位串、字符和字节组成。

2. 概念数据层

概念数据层是数据库的中间一层，是数据库的整体逻辑表示。指出了每个数据的逻辑定义及数据间的逻辑联系，是存储记录的集合。它所涉及的是数据库所有对象的逻辑关系，而不是它们的物理情况，是数据库管理员概念下的数据库。

3. 逻辑数据层

逻辑数据层是用户所看到和使用的数据库，表示了一个或一些特定用户使用的数据集合，即逻辑记录的集合。

数据库不同层次之间的联系是通过模式的映射进行转换的，从而保证了数据库的逻辑以及物理独立性。数据库具有以下主要特点：

（1）实现数据共享。数据共享包括所有用户可同时存取数据库中的数据，也包括用户可以用各种方式通过接口使用数据库，并提供数据共享。

（2）减少数据的冗余度。同文件系统相比，由于数据库实现了数据共享，从而避免了用户各自建立应用文件。减少了大量重复数据，减少了数据冗余，维护了数据的一致性。

（3）数据的独立性。数据的独立性包括数据库中数据库的逻辑结构和应用程序相互独立，也包括数据物理结构的变化不影响数据的逻辑结构。

（4）数据实现集中控制。文件管理方式中，数据处于一种分散的状态，不同的用户或同一用户在不同处理中其文件之间毫无关系。利用数据库可对数据进行集中控制和管理，并通过数据模型表示各种数据的组织以及数据间的联系。

（5）数据一致性和可维护性，以确保数据的安全性和可靠性。主要包括：①安全性控制：以防止数据丢失、错误更新和越权使用；②完整性控制：保证数据的正确性、有效性和相容性；③并发控制：使在同一时间周期内，允许对数据实现多路存取，又能防止用户之间的不正常交互作用；④故障的发现和恢复：由数据库管理系统提供一套方法，可及时发现故障和修复故障，从而防止数据被破坏。

（二）数据结构

将信息标引、著录后形成的信息记录，按一定格式依次录入计算机，并存储在磁带或磁盘上，形成供计算机检索用的数据库。数据库是被收集在一起的一组有序的信息单元，每个信息单元由若干个独立的结构单元组成，数据元存储在字段中，每个数据元描述信息单元的一个特性。每个信息单元由著者、标题、出版日期等数据元（字段）组成。

文献数据库大多是书目型数据库，这类数据库里存储的并非是原始文献的全文，而是

经过加工的二次文献，即文献的题录或摘要。数据库是一个包含大量反映文献外表特征的著录项目的集合。随着电子技术的日益发展和信息资源的数字化，也逐渐出现了一些全文数据库，如 AIAA Meeting Paper 全文数据库、中国学术期刊全文数据库等。

（三）数据库的记录格式

数据库的记录是构成数据库顺排文档（主文档）的基本单元，是对某一实体属性进行描述的结果。在书目数据库中，被描述的实体是某一特定的文献，这类记录通常被称做文献记录。一个数据库可能包含几千条甚至几十万条记录，一条记录又包含若干个数据字段。这些数据字段就是手工检索工具正文部分的文摘款目中的若干著录项目，例如原始文献的篇名、著者、文献出处、出版时间、文摘、主题词、语种等。它们是构成记录的最小信息单元。为了方便计算和检索，每一个字段都有自己特定的标识符，称为字段名，如 AB 代表文摘字段、TI 代表篇名字段、AU 代表著者字段等。数据库记录的著录项目（字段）往往比手工检索多得多，这就决定了计算机检索能够提供比手工检索更丰富的检索途径。下面以 PubMed 数据库的 MEDLINE 显示方式为例介绍数据库的记录格式。

PMID - 19151878
OWN - NLM
STAT - MEDLINE
DA - 20090119
DCOM - 20090305
IS - 0026 - 1270 (Print)
IS - 0026 - 1270 (Linking)
VI - 48
IP - 1
DP - 2009
TI - Biomedical and health informatics in translational medicine.
PG - 4 - 10
AB - OBJECTIVES: To discuss translational medicine advances challenging biomedical and health informatics. METHODS: Reviewing material presented at the Heidelberg 35th Anniversary Workshop, summarizing results from the 1st AMIA Summit on Translational Bioinformatics and discussing the opportunities, difficulties, and ethical dilemmas confronting researchers, practitioners, and healthcare managers in transitional bioinformatics. RESULTS: The first results in translational medicine are appearing in the biomedical literature. All rely on bioinformatics methods for analysis. CONCLUSIONS: Translational medicine introduces new problems of interpretation and application to healthcare. Applying results to complex human - machine systems raises ethical issues, which are augmented in healthcare informatics. Bridging biological, medical, and informatics knowledge requires new epistemological approaches.
AD - Department of Computer Science, Rutgers University, New Brunswick, New Jersey 08901, USA. kulikows@cs.rutgers.edu
FAU - Kulikowski, CA

AU - Kulikowski CA
FAU - Kulikowski, CW
AU - Kulikowski CW
LA - eng
PT - Journal Article
PL - Germany
TA - Methods Inf Med
JT - Methods of information in medicine
JID - 0210453
SB - IM
MH - * Biomedical Research
MH - Curriculum
MH - Decision Making
MH - Educational Status
MH - Ethics, Medical
MH - Evidence - Based Practice
MH - Humans
MH - * Medical Informatics
EDAT - 2009/01/20 09: 00
MHDA - 2009/03/06 09: 00
CRDT - 2009/01/20 09: 00
AID - 09010004 [pii]
PST - ppublish
SO - Methods Inf Med. 2009; 48 (1): 4 - 10.

这条记录由若干个字段（著录项目）组成。每个字段标出字段名称，如 Title、Author 分别代表篇名字段和著者字段。为了计算机在检索时能够顺利地识别字段，对每个字段又给予一个相应的字段代码标识符，如用 TI 表示篇名字段，AU 表示著者字段。有的字段又由若干子字段（Subfield）组成，这些子字段彼此是同等关系，在内容上有一定联系，但相互独立。例如，主题词字段（MH）中的各个主题词，分别是主题词字段中的子字段；文摘字段中的每个句子分别是文摘字段中的子字段。记录中的字段标识及对应名称说明如下：

PMID（PubMed Unique Identifier）：PubMed 记录存取号。在 PubMed 中，每篇被收录的文献都给出了一个 8 位数的顺序号，如 19151878，表示这是第 19151878 条记录

OWN（Owner）：提供引文数据的机构缩写，数据库编辑出版者

STAT（Status Tag）：状态标签，NLM 内部使用。包括：MEDLINE、in process、pubmed not medline、publisher 等

DA（Date）：日期

DCOM（Date Completed）：文献记录创建完毕的时间

IS（ISSN Type）：国际标准刊号，分印刷版刊号和网络版刊号

VI（Volume）：卷

IP（Issue）：期

DP（Publication Date）：出版时间

TI（Title）：篇名

PG（Pagination）：页码（起始页码到终止页码）

AB（Abstract）：文摘，英语文摘直接取自已出版的论文

AD（Affilication）：第一作者单位或通信地址

FAU（Full Author Name）：作者全名

AU（Author）：作者

LA（Ianguage）：语种

PT（Publication Type）：文献类型，主要有期刊论文、综述等

PL（Place of Publication）：出版地，指出版国家或地区

TA（Title Abbreviation ）：刊名缩写

JT（Full Journal Title）：期刊全名

JID（NLM Unique ID）：NLM 进行图书、期刊、音频视频资料编目时使用的唯一期刊标识符

SB（Subset）：数据子集，代表特定专题的期刊或文献子集

MH（MeSH Terms）：医学主题词

EDAT（Entrez Date）：文献被 PubMed 收录日期

MHDA（MeSH Date）：文献给予主题词标识并进入 MEDLINE 的时间

CRDT（Create Date）：文献记录第一次创建的时间

AID（Article Identifier）：文献标识码。Pii：出版物件标识符（Publisher Item Identifier）。出版商提交的文献标识符，用作数字对象标识符，主要用于构建 PubMed 外部链接

PST（Publication Status）：出版状态

SO：文献出处，包括刊名缩写、年、卷（期）、起讫页码

上述字段分为两类：一类称为基本索引字段，包括 TI、AB、MH 三种字段，都是反映文献主题内容特征的，提供从主题内容特征查找文献的途径；另一类称为辅助索引字段，包括 AU、SO、DP 等其余所有字段，都是反映文献的外部特征的，提供从文献的外部特征查找文献的途径。值得指出的是，不同的数据库，其记录的字段种类、数目、名称、代码不尽相同，在检索时，可根据每个数据库的说明查询字段的设置情况及使用方法。

（四）数据库的编排结构

所谓数据库的编排结构，就是计算机检索系统中数据库的每条记录数据项的编排方式，有顺排文档和倒排文档两种。

1. 顺排文档

顺排文档存入了数据库的全部记录，文献记录按照存取号的大小顺序排列，类似于检索刊物中按文摘号排列文摘款目。每一篇文献为一条记录单元，一个存取号对应一条记录，存取号越大，对应的记录就越新。由于它存储记录的最完整的信息，所以，又把它称之为主文档。如果在顺排文档中进行检索，计算机就要对每个检索提问式逐一扫描数据库

中的每一条记录，存储的记录越多，扫描的时间越长，这样检索效率就会很低。

2. 倒排文档

倒排文档是将主文档中的可检字段（如主题词、著者等）内容抽出并在其后赋予相应的文献记录号，然后将字段内容按某种顺序重新排列起来所形成的一种文档。不同的字段组织成不同的倒排文档（如主题词倒排文档、著者倒排文档等）。倒排文档可以按主题词的字顺排，也可以按分类号的大小排。按表达文献内容特征的主题词排列的文档称为基本索引文档；按表达文献外部特征排列的文档称为辅助索引文档。倒排文档只有文献的标识、文献篇数及文献存取号。因此，在实施检索时，必须和顺排文档配合使用，先在数据库的倒排文档中查得文献篇数及其记录存取号，再根据存取号从顺排文档中调出文献记录。倒排文档类似于检索工具中的辅助索引。

三、检索技术

信息技术包括了信息存储技术、信息组织技术和信息检索技术。信息检索技术是信息获取的重要技术，它包括了布尔逻辑检索（Boolean Searching）、位置检索、命令检索、截词检索、指定字段检索、加权检索、跨语言检索等检索技术。

（一）布尔逻辑检索

布尔逻辑检索是检索系统中应用最为广泛的检索技术，是最早建立的检索理论，是最简单、最基本的匹配模式。其理论基础是集合论与布尔逻辑。它采用布尔逻辑表达式来表达用户的检索要求，并通过一定的算法和实现手段，从经过标引的文献信息集合中检出所需要的文献信息。布尔逻辑表达式是通过布尔逻辑运算符（Boolean Operators）连接检索词进行逻辑组配，表达词与词之间的关系。

1. 逻辑“与”（AND）

这是具有概念交叉关系和限定关系的一种组配，可表示为“A AND B”（A、B 分别代表不同的检索词）。检索时，“A AND B”命中的信息同时含有 A 和 B 两个概念。其作用是缩小检索范围，提高查准率。例如，检索“阿司匹林治疗高血压”的有关文献，就是要检出那些同时出现有“阿司匹林”和“高血压”两个概念的文献信息。AND 所组配的两个检索词可在同一记录的任一字段中出现。

2. 逻辑“或”（OR）

这是属于概念并列关系的一种组配，可表示为“A OR B”。检索时，“A OR B”表示检出含有检索词 A 或 B，或同时包含 A 和 B 的文献。其作用是扩大检索范畴，提高查全率。例如，检索“所有关于阿司匹林（乙酰水杨酸）”的有关文献，就是要检出所有关于“阿司匹林”、所有关于“乙酰水杨酸”以及同时包含这两个概念的文献信息。

3. 逻辑“非”（NOT）

这是概念包含关系的一种组配，是从原检索范围中排除一部分。“A NOT B”表示所有包含检索词 A 但不含检索词 B 的文献，其作用是缩小检索范围。例如，检索“番石榴叶治疗轮状病毒（非细菌）所致腹泻”的有关文献，就是要检出那些同时出现有“番石榴叶”、“腹泻”、“轮状病毒”这三个概念的文献信息，但有关“细菌性腹泻”的文献信息必须去除。

在编制检索式时，AND、OR、NOT 在一个检索式中可以混合使用。