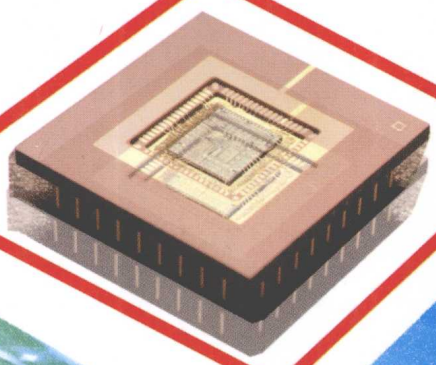
信息资源管理系列丛书

信息采集

XINXI CAIJI

侯延香 王霞 编著



知识产权出版社
全国百佳图书出版单位

信息采集

侯延香 王霞 编著

知识产权出版社
全国百佳图书出版单位

内容提要

本书全面论述了信息采集的相关理论、方法和技术，包括信息采集的概念与类型、内容与程序、信息需求分析、信息源的筛选、主要采集方式与途径、信息采集的效率与质量等内容。本书内容完整、系统性强，注重理论与案例分析相结合，具有较强的可读性和可操作性。既可作为信息管理、情报学、图书馆学、企业管理等专业的教材，也可供广大政府及企事业单位信息管理工作者及其他对信息采集有兴趣的人士阅读参考。

责任编辑：于晓菲

图书在版编目(CIP)数据

信息采集/侯延香，王霞编著. —北京：知识产权出版社，2011.9

ISBN 978-7-5130-0772-6

Ⅰ.①信… Ⅱ.①侯… ②王… Ⅲ.①信息学 Ⅳ.①G201

中国版本图书馆 CIP 数据核字(2011)第 169323 号

信息采集

XIN XI CAI JI

侯延香 王霞 编著

出版发行：知识产权出版社

社　　址：北京市海淀区马甸南村 1 号　　**邮　　编**：100088

网　　址：http://www.ipph.cn　　**邮　　箱**：bjb@cnipr.com

发行电话：010-82000893 82000860 转 8101　　**传　　真**：010-82000893

责编电话：010-82000860 转 8127　　**责编邮箱**：yuxiaofei@cnipr.com

印　　刷：北京富生印刷厂　　**经　　销**：新华书店及相关销售网点

开　　本：787×1092mm　1/16　　**印　　张**：19.5

版　　次：2012 年 1 月第 1 版　　**印　　次**：2012 年 1 月第 1 次印刷

字　　数：308 千字　　**定　　价**：42.00 元

ISBN 978-7-5130-0772-6/G·432(3672)

前　言

当今社会，信息已渗透进社会的每一个角落，其重要性已逐渐被人们所认识。正如美国前总统卡特所说："对于我们，信息就像阳光和氧气，它点燃创造智慧的火花，它照亮了通向未来的道路。"然而，随着网络和科技的发展，信息总量呈爆炸式增长趋势。品种繁多、形态多样、分布不均、良莠不齐的非结构化信息铺天盖地，给人们的信息采集和利用带来了困难，甚至可能遭遇"无用信息遍地皆是，有用信息芳踪难觅"的尴尬。因此，如何采用科学的方法，快速、及时、准确地采集到新颖可靠的信息，是人们密切关注的焦点。

信息采集是信息管理的首要环节，是管理利用信息的基础。然而，迄今为止，有关信息采集的理论著述散见于《信息资源管理》、《信息管理学》等著作中，直接论述信息采集理论的著述较少。1995 年，孟雪梅老师主编的《信息采集》是目前可以查到的较为全面和专业的信息采集著述，但随着时代的发展，其中许多内容已经过时。2002 年，张安珍教授出版了《信息采集、加工与服务》，较为专业地论述了信息采集的相关理论，但其中部分内容，尤其是网络信息采集部分已较为陈旧。2005 年以来出版的《网络信息采集与应用》、《网络信息采集》、《网络商务信息采集》、《市场信息的收集与处理》等著作，或侧重于网络信息采集，或侧重于市场调查信息采集。因此，目前尚缺乏系统、新颖的信息采集理论著述。针对上述情况，本书力求在系统阐述信息采集理论的基础上，结合信息采集发展现状与趋势，打造一本内容丰富、理论系统、实践性强的信息采集著述。

本书全面系统地阐述了信息采集的相关理论，全书共分 10 章，第 1 章为信息采集概述，由侯延香撰写；第 2 章为信息采集的内容与程序，由王霞撰写；第 3 章为信息需求分析，由沈洪杰撰写；第 4 章为信息源的筛选，由周秀霞和丁莉撰

写;第5章为信息采集的方式与途径,由梁宏撰写;第6章为基于因特网的信息采集,由卢文锋撰写;第7章为基于数据库检索的信息采集,由张立新撰写;第8章为基于社会调查的信息采集,由侯延香撰写;第9章为信息采集的效率与质量,由宗蕾、徐秀杰撰写;第10章为大学生与信息采集,由韩宇撰写。全书由侯延香拟定写作大纲,负责统稿;由侯延香和王霞负责稿件审核;最后由侯延香定稿。

本书的写作,得到了山东建筑大学管理学院领导的支持和教研室主任邓晓红教授的鼓励,在此表示衷心的感谢!在本书写作过程中,我们参考和借鉴了大量的中外文书刊和网站资料。借此机会,我们向这些参考文献作者表示诚挚的谢意。由于篇幅所限,我们未能一一列出所有参考文献,还有部分文献来自网络,无法列出作者的姓名,在此,我们对未能列出的参考文献作者表示深深的歉意和诚挚的谢意!感谢我的家人对我工作的支持,感谢我的博士导师王知津教授的谆谆教诲!感谢知识产权出版社于晓菲编辑的辛苦劳动!

由于信息采集是一个快速发展和不断更新的领域,加之编者的学识、水平和能力有限,缺点、疏漏和错误在所难免,恳请各位专家、学者和广大读者批评指正,以便在本书修订时加以补充、更正和完善。

侯延香

2011年6月

目　录

第1章　信息采集概述

【本章提示】

本章主要介绍信息采集的基本知识，包括信息的概念与特点及信息采集的概念、类型、意义、发展与趋势等。通过本章的学习，学生应当掌握信息采集的类型，理解信息采集的意义，了解信息的特点、信息采集的发展与趋势。

1.1　信息的概念与特点

1.1.1　信息的概念

随着科技的进步和互联网的发展，大多数人对“信息”一词已不再陌生。早在春秋战国时期，著名军事家孙武在《孙子兵法》中提出，“知彼知己者，百战不殆；不知彼而知己，一胜一负；不知彼，不知己，每战必殆”，充分说明了信息的采集和利用在军事领域的重要作用。在信息时代，微软创始人比尔·盖茨(Bill. Gates)在《未来时速》中指出：“将您的公司和您的竞争对手区别开来的最有意义的方法，使您的公司领先于众多公司的最好方法，就是利用信息来干出色的工作。您怎样搜集、管理和使用信息将决定您的输赢。”更加体现了信息在竞争中的重要性。在日常生活中，信息也与我们息息相关：医生靠号脉、观察舌苔、询问等方式来采集病人的病情信息；母亲从婴儿的啼哭中，得到孩子饥饿的信息；到集市上走一遭，便收集了蔬菜的价格信息……可见，从古至今，从军事到企业竞争再到人们的日常工作生活，信息的影响已渗透到各个领域。

然而，对于什么是信息，目前还没有一个统一的定义。有人认为，信息是世界上事物存在方式和运动状态的反映，一切能够表示一定物理形式和物理量的代码、符号、声音、光亮、颜色等，都可以称为信息。还有人认为，只有那些能被

人们发现、理解、接受，能对人们解决问题有用的事物表征，才能称为信息，即信息能消除人类认识中的“不确定性”。《辞海》中对“信息”的解释是：“信息是指对消息接受者来说，预先不知道的报道。”根据这种认识，那些对人们解决问题没有意义的，已被人们熟知的知识内容不能叫做信息。我们认为，信息有广义和狭义之分。广义上，信息是事物存在的方式和运动状态的表现形式，可用代码、文字、符号、声音、光、颜色等多种符号表达。狭义上，信息是为了特定目的而被人们发现、理解、传递、交流的事物表现形式。即：对其接受者来说是预先不知道的消息。广义的信息定义表明，信息是普遍存在的，在时间、资金、人力、设备等条件允许的情况下，应尽量拓展信息采集的范围，并注重对多媒体信息的采集利用。狭义的信息定义表明，信息是有针对性的，在时间、资金、人力、设备等条件有限的情况下，应明确信息采集的主次目的，并根据需求有目的有计划地开展信息采集。

1.1.2 信息的特点

与物质、能量等不同，信息有其独特的本质属性，其特点主要表现在以下几个方面：

1. 时效性

信息的时效性是指信息从产生、接收到利用的时间间隔与其价值存在一定的关系。这种比例关系在大多数情况下表现为一种正比例关系，即信息提供和利用的时间越早，信息的价值就越大；反之，就越小。如：股票市场上的交易信息瞬息万变，谁能及时掌握股票行情，谁就能获得直接的经济利益。当然，这种比例关系有时也表现为相反的情况，随着时间的推移，某些信息可能像陈年老酒一样不断增值，如基础理论的研究成果等。

2. 传递性

信息的传递性是指信息可以通过一定的传输工具和载体进行时间上和空间上的传递。人类之所以能够提供、接收和运用信息，就是因为信息具有可传递性。信息可通过多种渠道进行传递或交流。古代人类借助于声光、符号等载体传递信息。文字的发明和使用，突破了信息在时间和空间上传递的局限性，人类可以跨时代、跨地域地传递信息。纸张的发明，使信息的传递更加便捷。现代信息技术的应用扩大了人类传递信息的范围，提高了人类传递信息的速

度。由计算机技术与现代通信技术相结合而形成的信息网络，高速度地处理、高密度地存储和远距离地传输信息，使信息传递发生了质的飞跃。

3. 依附性

信息的依附性是指信息的存储、传递和交流必须依附在一定的物质载体上。信息本身是看不见、摸不着的，它只能附着在某种载体上，并以一定形式表现出来，才能为人们所利用。一场激动人心的演讲，通过语言而催人泪下，借助声波而广为传递，利用磁带方可存储，而语言、声波和磁带只是信息载体，离开了这些物质载体，人们就无法采集利用信息。实际生活中，人们要运用语言、文字、图表、声像、实物等记载或反映信息，并要用纸张、胶卷、磁带、光盘等物体存储信息。人们要采集信息，首先要获得载有信息的载体，通过对载体的利用，才能解析出其中的信息内容。

4. 共享性

信息的共享性是指信息可由不同个体或群体在同一时间或不同时间共同享用。通常，物质的交换与转让，一方有所得，必使另一方有所失，而信息产品的使用价值可以同时被若干个用户所使用，任何一个用户不会因为信息资料的提供和传递而失去它。如：主持人为观众播报新闻，其掌握的信息不仅不会在播报中遗失，相反会在播报中得以巩固。英国文学家萧伯纳说过："倘若你有一个苹果，我也有一个苹果，我们彼此交换苹果，那么，你和我仍然各有一个苹果；但是，倘若你有一种思想，我也有一种思想，我们彼此交换思想，那么，我们每个人将各有两种思想。"也充分说明了信息的可共享性。

5. 可伪性

信息的可伪性是指信息在其衍生过程中可能发生变化，产生虚假信息。在信息衍生过程中，由于信息失去了与源物质的直接联系以及人们在认知上的差异，对同一信息，不同的人可能会产生不同的理解，形成"认知伪信息"，如：盲人摸象。由于传递过程中的失误，会产生"传递伪信息"，如：计算机乱码。另外，也有人会出于某种目的，故意采用篡改、捏造、夸大、假冒等手段，制造"人为伪信息"，如：虚假广告。人为伪信息会带来社会信息污染，具有极大的危害性，必须严加防范。

6. 效用性

信息的效用性是指使用信息能带来经济或社会效益。科学合理地应用信息会使信息增值。值得注意的是，信息的效用性是相对信息需求主体而言的，同一条信息对不同的用户，其价值不同。如：一份职称外语等级考试的通知对已是教授职称的教师没有多大价值，但对于正要申报高级职称的教师则是必备信息。

案例 1-1　左撇子用品专营店

日本人渡边曾经是个打工仔，被老板解雇的几次经历使他萌发了自己当老板的愿望。开始，他想在东京开家小商场，但经过调查了解后，知道东京的商场很多，竞争激烈，自己如再挤进去，没什么独特优势，很难生存。一天，他在一份报纸上看到：美国人中有1/4、日本人中有1/6、英国人中有1/7是左撇子。此信息令他忽生灵感：开一家左撇子产品专营店。因为当时众多厂家均以右手习惯来设计产品，几乎没有人考虑左撇子的习性和生活、工作需要。于是，他立即说服一些厂商专为他的商场设计、生产一些左撇子专用产品，如：汽车驾驶盘，网球、高尔夫球用具等，结果这些产品大受世界各地左撇子消费者的欢迎。不久，其左撇子用品专营店成为东京最有实力的大商场[①]。

1.2　信息采集的概念与类型

1.2.1　信息采集的概念

信息采集是根据用户的特定需求，通过各种途径对相关信息源进行科学地收集、检索、调查、采访、获取、鉴别、整理、分析，并最终形成所需有效信息的过程。信息采集实际上是一个泛指的概念。从使用的信息源来看，它囊括了文献、实物、口语、视觉资料等多种信息源；从适用的对象来看，各类决策人员、管理人员、研究开发人员、技术人员、统计人员、策划人员、调查人员、咨询人员、传播人员、政府工作人员等都要采集信息；从利用的工具和方法来看，它既包括利用专门的检索工具从序化信息中收集信息，也包括利用问卷、访谈等方法收集整理社会信息的过程，还包括利用扫描仪、读卡器等设备收集结构化信息。值

① 任洪润. 市场信息的收集与处理. 北京：电子工业出版社，2006. 8：5.

得注意的是,信息采集不等于信息检索。信息采集包含信息查找、检索、访谈、调查、获取等多种内涵。而信息检索是仅根据特定的需求运用某种检索工具寻找序化信息的过程。

案例 1-2　日本人对大庆油田的信息采集

1959 年 9 月 25 日,中国石油勘探队在东北松辽盆地找到了大庆油田,摘掉了中国贫油的帽子。当时,国家对大庆油田的位置和生产规模严格保密。然而当许多中国人(包括一些中上层国家干部)还不了解大庆时,日本却准确测知了大庆油田的所在位置和生产规模。日本人是怎样采集到这些重要信息的呢?首先,他们从 1960 年《人民日报》关于"铁人"王进喜的报道中发现,油田在 9 月末就已寒气逼人,因而确定了油田所在的地理纬度;其次,1966 年 7 月,《中国画报》上曾刊载王进喜头戴厚棉帽的照片,再结合运到北京的油灌车上剥离的黑土,得到油田位于零下三十摄氏度左右的东北地区。根据运原油的列车上灰尘的厚度,日本人还测算出了油田与北京的距离——油田应在哈尔滨与齐齐哈尔之间。另外,他们从 1966 年 10 月的《人民中国》杂志上看到,王进喜率井队到大庆,首先在马家窑车站下火车,利用地图,很快就测知了大庆油田的所在位置是在黑龙江省安达市车站附近。他们还从高空侦察照片上发现了大庆的储油罐,并根据照片上的人和油罐的比例关系,测算出 1971 年大庆石油年产量为 1200 万吨。日本人利用《人民日报》的报道、《人民中国》杂志的文章及实物黑土等信息源得到了他们想要的信息资料,并在实地调查的基础上,根据当地的气温、湿度等气候条件为大庆油田设计炼油设备,在相当长的一段时间内几乎垄断了我国的石油设备进口市场。

1.2.2　信息采集的类型

按照不同的标准,可以将信息采集划分为不同的类型。

1. 按采集内容划分

(1)文献信息采集

文献信息采集是指以文献作为采集对象,收集包含用户所需信息内容的文献。包括文献线索信息采集和文献文本信息采集。通常,可利用书目、索引、摘要及各类信息资源数据库等检索工具获取文献的线索信息,可通过购买、复制、检索、交换、接收、征集、申请等方式获取所需的文本信息。文献信息采集是最