# 计算机与档案馆

戴　璞　编著

兵器部档案馆编印

一九八五年七月

# 前　言

电子计算机在档案检索和管理方面的应用，证明了计算机不仅是档案工作者的一个基本工具和得力助手，而且为档案文献的管理和利用开创了一个新天地。但是，档案自动化工作较比其它文献工作自动化发展迟缓，在国外尚处于实验应用阶段，在我国还只是刚刚起步。这个实事启示我们，开展档案自动化工作应该十分重视从其它信息管理人员所进行的工作中吸取经验。就我国档案自动化工作而论，更应如此。因此，本文所有介绍的内容主要是国外一些国家档案自动化情况。只要我们善于学习和借鉴国外档案自动化工作经验，又善于结合我们的实际，就会制定出我们的工作方针和开发适合我国档案工作需要的工作方法。

戴　璞

*1984.10*

# 目　　次

第一章　计算机的使用和档案应用的目的
一　档案自动化信息检索
二　档案信息的组织与文档结构
三　计算机软件
第二章　档案工作者和计算机系统
一　档案自动化的产生和发展
二　档案馆的早期自动化
三　档案自动化的近期发展
第三章　档案自动化技术的实施
一　系统目标的考查和确定
二　系统工作流程鉴定和审查的方法
三　计算机资源的选择
四　档案信息的前处理工作
第四章　自动化技术和档案工作的新课题
一　自动化技术和档案标准化工作
二　检索语言和标引工作
§1　检索语言的发展阶段
§2　档案检索语言与档案自动化系统
§3　档案标引工作
三　机读档案
结束语
参考文献

# 第 一 章

## 计算机的使用和档案应用的目的

什么是电子计算机及其工作原理，在很多有关计算机的科普性书籍中都有很清楚的讲解和介绍。本章主要介绍发挥计算机功能的一些方法，重点介绍档案应用计算机技术的一些基本概念，为档案工作者考查和评价一个档案自动化检索系统提供一点基本知识。

### 一、档案自动化信息检索

信息检索这个词，通常是对检索系统和检索过程而言。信息检索可以告诉检索者他们所要查找文件在哪里和目前的状况。实际上进行全文检索的信息检索系统是非常少的，绝大多数检索系统只是提供某种形式的二次文献，例如，文献引文或者文献引文与著录信息，或者文献内容的摘要。按照分类或字母顺序排列的卡片或书本式的主题录目是最常见的信息检索形式。主题目录是先组配式检索系统，这是因为编制主题目录时，一次在所有的文献之间建立起主题关系。检索时，用户不能随意改变主题词或者主题分类之间的关系。在目录的组织结构上也是线性，或者叫树形结构。虽然一个复杂的主题可以由一连串的叙词来表示，但只能在首词下面填写著录项目，所有其它的词都得从属于这个首词。虽然使用根据不同的文件存址进行重复著录方法可以实现多途径查找，但是这样做往往由于受到人力和资金方面的限制而使得著录项的数目和数据更新的频次不能太多。

为了解决传统卡片目录的局限性问题，需要开发允许利用者改变叙词关系的新的检索系统。在四十年代初期，电子计算机问世之前，英国人贝坦和法国人科多尼几乎同时研制了光符合系统。贝坦和科多尼系统实际上是一种穿孔卡片数据处理系统，这种检索后统按照每一个叙词或文件类别对卡片进行归类，就是说每一张卡片上的孔位都专用于一个文件。这样对每一个文件进行分析，后面将它们归类在一个或者多个主题类目之中，然后在主题卡片上用来表示该份文件存取号的位置处穿一个孔。根据文件查找者的提问要求检索文件时，只要把主题卡片重叠在一起，就可以找出主题关系正确的那些文件。采用增加或减少主题卡的方法可以扩大或缩小检索范围。贝坦和科多尼系统是后组配索引系统。检索时，只要主题词之间的逻辑组合合适，允许自由组配主题词，进行主题法检索。

四十年代后期，卡尔文·穆尔斯 在英国也研制了一个类似的后组配检索系统。穆尔斯的系统使用了不同的文件结构。贝坦和科多尼尔在他们的词表中，对一个词用一张卡片，而穆尔斯的则是一个文件用一张卡片，並且用每张卡片的槽口位置对主题词进行编码。现在把穆尔斯的编排方式叫做“款目项”或叫“款目词”；把贝坦和科多尼尔的编排方式叫做“词项”或“词款目”。至今，计算机信息检索系统中，这两种方法自然是编排索引文件的两种基本方法。

由贝坦和科多尼尔及穆尔斯研制的这两种不同的检索系统的共同特点都是使用受控词（按照标准词表使用单词或词组）。人们称这种检索系统为受控词检索系统。在这种系统中标引人员和检索人员都只限于使用在标引规范文件（主题词表）中出现的那

些词。受控词表通常分为分类表和主题词表两种。

在五十年代初有个叫陶比的英国人建议，人们可以从文件正文中直接抽取单个词作为标引关键词，而不必从标引规范文件（受控词表）中选取叙词或叙词组。这个建议很受欢迎，因为这样标引工作可以交成花钱不多的誊抄工作。但是，这种不受词表控制，完全采用自然语言进行标引的关键词法，由于标引词容易出现一词多义和主题分散致使检索效果和检索质量较低，而没有推广开来，而使用规范化词表进行标引仍占据优势。不管怎样，由于采用自然语言进行标引的方法的问世，在此基础上又产生了三种信息检索的方法，即：上下文内关键词索引（KWIC）、上下文外关键词索引（KWOC）及段落或全文检索和自动标引。

上下文内关键词索引（KWIC）、上下文外关键词索引（KWOC）和其它编排索引的基本原则都是用包含在文件内的关键词，而不是由标引人员赋于的反映某些概念的关键词而制作索引。这里所说的关键词是指那些从文件标题、文摘正文中摘取出来作为标引词的词，在KWIC索引中，关键词是按字母顺序排列出现在其任意一侧的词索的上下文中，从而改进了标引词的标引效果。编制KWIC索引程序通常包括一个删除词表（一个包含诸如介词、连词和冠词等次要词的表）。这个删除词表所包括的词是一些选择关键词时，命令计算机忽略的词，编制KWIC索引主要用于对科技报告和文章的标题产生索引。显然，标题关键词索引的价值取决于标题反映文件内容的确切程度。

用段落或全文检索系统所有文件的全文都须变成机读形式，不须标引就可以进行存贮和检索。这些系统通常按照文内出现的词汇的逻辑组合的能力（不考虑这些词之间相互关系时，必须要考虑两个或更多的词之间的间隔）。目前在国外法律界是这种段落检索的主要用户，使用这种方法检索科技文献的系统也不断增多。但这种方法的很大问题是如果原文不是机读形式，则需要把全文再重新输入一遍，这样建立起来的检索系统所花的费用就很大，同时对有些文献来说也是不可能做到。但是，随着电子照排的机读文件的增加，全文（正）检索系统就会越来越多。

所谓自动标引是用计算机直接抽取或赋于标引词，即按照某一标准从一篇文章的全文、文章的概述或文章摘要中自动选取标引词。运用统计标准选取标引词是常见的方法，也有使用语言学和上下文批断标准的方法选取标引词的。典型的办法是，将一篇文献中的词汇列成表，一般情况下根据各种词在文献中出现的频率进行统计测定。选取那些出现频率较多的词作为这篇文献的标引词．那些在检索时，在所有的文献或大多数文献中出现频率高的词也应该被排除在外。但总的来说，被排除的词应是一些缺乏辨识力和检索意义极小的词。

各种自动标引方法已经进行过大量的试验。试验的主要目的在于提高标引效果和大大节约经费和时间，并使自动标引的质量达到由训练有素的专职主题标引员的标引效果。试验结果

表明在提供标引词的有效性方面已取得了进展。但是自动标引仍然多限于进行标引试验，在实际工作很少使用。

虽然信息检索中自然语言仍在继续使用着，但是受控词表的应用仍然很盛行。并且在日趋增加和完善。这种增加的最重要特征是主题词表（叙词表）的应用和发展。主题词表是一个具有相关结构的受控词表、词表中的每一个词都与各种相关相关联。第一个主题词表是美国E。I人杜邦在1959年为信息检索创建的，尽管这个表不够完善，但是杜邦的主题词表在结构上具有层次清楚和词间相互参照的特点，杜邦使用主题词表上的一个词（主题词）和其它的主题词组配起来进行标引和检索。这个主题词表的模式一出现，产生了很大影响。在近二十年中，主题词表发展很快，它已成为后组配系统中，为人们所普遍接受和使用的词汇规范化的模式。现在很多学科领域都有它们自己的主题词表。由于迫切需要解决迅速增长的文献资料的及时存取问题以及科学技术交流的不断扩大，使得主题词表的变化也很快。

档案自动化信息检索实际上是检索方法和检索语言的问题。任何一种技术方法都必须通过一定的技术手段，利用一定的工具来实现。两者的有机结合，相辅相成，逐步发展，构成了不同时期的从低级到高级的各种档案自动化检索系统。上面对档案信息检索主要进展所做的简要综述，试图介绍一些有用的概念，希望能对档案工作者在改进智能管理和检索方法等方面所进行的工作有所帮助。

## 二、档案信息的组织与文档结构

一个档案馆使用计算机时，必须为计算机提供要处理的档案数据或信息及计算机运行的程序。处理数据的各种详细指令由程序设计人员编写。但是，要使计算机发挥作用，达到档案工作应用计算机的目的，就必须要求档案工作者密切配合，协助确定处理什么信息和怎样处理这些信息。

传统手工编制的档案目录是由条目组成的；条目是由著录项目组成的。当把这种手工编制的档案目录通过输入设备输入到计算机里，便组织构成了计算机（目录）文件。这种计算机（目录）文件是由记录组成的；记录是由字段组成的。如同手工编制的目录构成必须按照一定的编排次序、排列顺序及书写表达方式一样，计算机（目录）文件更需要有固定的排序顺序和独特的表达方式。如它的记录就是作为一条长长的字符串提供给计算机。并且还需要一种特殊的方法来告诉计算机什么地方是一个字段的结束和另一个字段的开始。因此，档案工作者必须把要处理的数据或信息组织到这种称之为计算机文件的数据结构之中，这种工作方式是计算机化系统的一个重要的特征。

正如前面所讲的，一个文件是由许多记录组成的。这些记录是相互关联、结构相同的数据项集合。每个数据项都可作为一个单元来处理。在一个档案馆建立起来的包括案卷级或文件级的目录数据的各种计算机文件中，一个记录应该记录着有关该案卷内容的所有著录信息（例如标题、作者和日期等）。这些信息的一部分记录在一个字段或称数据字段（一个用来记录特定的类目或数据单元的特

定区域）之中。在档案文件中，每个记录可能包括一个作者姓名字段和一个含有每个手稿文集的作者命名的标题的附加字段。在一个文件级的目录中，可能有多个字段，其中包括文件标题、日期、通信者姓名或主题词等。一个系统可能要求一个记录的某些字段担当特殊职能，这时绝大多数职能可由系统用户设置。

设计字段是计算机应用设计中最为重要的一个方面。因为计算机查找的是字段的位置，而不是其内容，所以字段必须总是出现在一个记录（固定位置）之中的同一个物理位置上，或者出现在用一个特定标识符为首的相对位置上。（把一个字母／数字代码称为一个标识符，或称为字段标识符，数据标识符）。整个文件必须使用固定的字段设置。例如，在一个登记文件中，必须用每一个记录的相同字段填写进馆日期。否则，要产生一个进馆年代一览表就可能非常困难。如果为多个组织机构建立一个数据库（计算机顺序存取的机读形式的累积信息的大型文件）。使用这种数据库的所有用户必须使用相同的字段设置和标识符结构。这时，应当根据字段记录信息的种类和这些信息的可能应用再很好地定义字段。

设计标识符结构的工作要认真细致。要做到这点，设计者就必须熟悉档案学方法论，透彻地了解人们对系统的要求。设计者还必须熟悉自动化系统的特点。这些特点往往决定了一个特定系统满足档案工作者要求的工作效率。

记录长度是系统的一个重要特征。几乎所有的系统都规定了每个记录所能容纳的最大字符数。虽然档案工作者有时很鲁莽，但他们要求足够的记录容量还是合乎情理的。

足够的字段长度也很重要。比最大的记录长度和字段长度都更

有意义的是字段长度可变（长度可以变化直到一个最大值）。有些系统要求所有的记录是定长的，并且由定长（固定长）的字段组成。定长记录和定长字段一般用于商业系统。而档案工作者通常需要具有变长记录、变长字段和灵活出现字段（在每个记录中有些字段可以出现，也可以不出现）的系统。因为档案文件没有统一的题录数据，而且档案工作者並不希望以同样详尽的程度去描述各种馆藏文件。

系统结构的另一方面的问题是数据库数据的排列和存取。数据存取方法直接影响着数据库建立者的使用效果。既使在一些专用的系统中，文件不同，存取方法也可能不同。数据排列横式虽然各有不同，但基本上有两种类型；一类用于顺序存取，一类用于直接存取。

顺序文件通常以数字或字母顺序排列，必须按顺序串行检索和处理这种文件。设计这种文件是为了进行批处理（一种脱机数据处理的方法）。采用这种方法处理数据时，须事先编写好计算机程序。程序编好后，计算机执行一组程序（执行中采取执行完一个程序，便调入该组中的下一个程序开始执行的方法）。典型的批处理方式是把数据按顺序存贮在磁带上，按照予先确定的顺序，提取磁带上的数据运行一组程序，並按事先设计好的格式打印结果。在档案馆用这种方法编制各种不同的打印目录和索引是有效的，並且经济实用。因为大型文件的联机顺序检索可能非常慢，而且不经济。所以人们很少设计允许在交互式终端上进行联机检索和信息显示的顺序存取信息的系统。

直接存取信息的系统可在一个存贮器的任何一个位置上存取信

息。存取时间与数据存贮的位置无关。直接存贮器上的存贮数据的每一个位置用具有一的数字地址识别。磁盘是最常用的直接存贮器。存贮在磁盘上的数据记录在扁平圆盘磁表面上。许多称之为磁道的同心圆中一个磁化点表示一个二进制数位。由若干个二进制数位组成字符；再由若干个字符组成字段……。

磁盘与密纹唱片很相似。两面都可以存贮数据。通常（10～20个）磁盘串在一根旋转轴上，使磁盘串（磁盘组）的每一面都可以由各自的活动臂（查找臂或存取臂）存取信息。每一个查找臂上装有一个读／写磁头（一种可以在磁性记录介质上读写信息的装置）。计算机系统总保持着关于磁盘上每个文件位置的索引，从而可以使查找臂直接移动到磁盘相应的存放位置上。待需要检索的记录旋转到该磁头下面时，便从该文件上存入或取出数据。

在联机磁盘存贮器上存取数据只须花几个毫秒的时间。对于进行交互式应用的用户，这个存取数据的过程几乎是瞬间发生的。在交互系统中通常允许运用布尔逻辑运算（这种方法要求只限回答是”或“非”，包括的逻辑运算符有“与”（and）“或”（or），“非”（not）“禁”（except），“如果”（if）和“则”（then）。这些符号可以用许多种方式进行组合检索数据库。（例如，显示有关联帮监督“与”黑手党“与非”联帮调查局的文件的所有案卷标题）。这种后组起检索方式允许查询者自行确定其提出的检索词之间的关系，使系统按照其特殊的需要进行检索。此外，查询者还可以修改查询要求，扩大或缩小查找范围。档案工作者使用直接存取信息的系统，可以通过对话修改或补充文件的数据，使近期档案资料信息很快就能得以利用。

档案工作者如果把直接存取信息的系统和间接存取信息的系统的各自优点加以比较，很快地就可以得出直接存取信息的系统更好些的结论。尽管对直接存取信息的系统功能更新这点几乎没有任何人提出争议。但是，一些档案工作者还是认为档案文件没有必要用直接存取信息的系统。既便是采用了这种直接存取的信息的系统，使用这种直接存取信息的系统，使用这种系统所需的附加费用也会没有保证。一般来说，直接存取信息的系统更费钱。直接存取信息系统的程序设计通常比较复杂，成本较高。除此之外，直接存取信息的系统的数据存贮器成本较高，其原因有下列三点：(1)磁盘存贮器存贮介质的成本比磁带或键控穿孔卡片高；(2)要实现直接存取系统所存贮记录的随机检索需要进行附加的记录管理和保存系统索引数据，所以通常需要较大的存贮空间；(3)为了进行交互性检索，联机文件的维护费用常常比批处理高。

另一方面，虽然某些顺序存取信息的系统的计算机使用费用可以大为减少，但是直接存取信息的系统使用起来却是简便、效率高。在维护一个高质量的文件时，交互式数据输入和文件编辑的效率可以减少人工费用。这样做可能使数据库的结构不够理想，但评价系统的性能和经济效益应该从能否满足档案工作者的要求的角度去考虑。

档案工作者无论考虑或着手应用任何一种系统时都应该看它是否实际可用，并且避免建立没有足够的档案数据或信息可提供的系统。此外，有效的系统支持也是很重要的，因为它可能决定任何一个软件包的长远使用价值。系统支持包括完备和有水平的编写程序的文本（系统中所有程序的操作的介绍和使用说明书）以及可靠的

系统开发和维护。缺乏可靠的现行程序文本，就会设制一个系统从一种类型的计算机移植到另一种类型的计算机。而且会防碍软件包的有效使用。几乎最新设计的所有软件包都会有一些错误（缺陷）。所以，良好的系统维护是十分必要的。用户发现这些错误时，应予以纠正，並适当地修改文本。完全有必要努力改进和扩大系统功能，並将这些进展情况通知给系统的用户。服务周到的计算机公司所拥有的和所提供的系统通常拥有最完善、最可靠的系统支持。但这些系统往往价钱最贵。如果协议适当和资源充足，大的非商业性机构可以提供良好的系统支持。为了合作互利，也可以组织用户协会。采用一个没有固定可靠的系统支持的系统是靠不住的。应该避免采用那些依靠私人协议和工作关系去进行维护和开发的那些系统。

## 三、计算机软件

计算机运行是由软件（程序的集合）指挥的。软件由一系列告诉计算机一步一步地如何工作的指令组成。每一条指令使计算机执行一个基本的功能。包括加法、减法、乘法、除法、打印、逻辑判别（比较）是／不是、请求输入和请求输出等八种。

如果这些指令排序正确、书写无误，计算机就会有效地按规定步骤运行。把输入的数据转变成正确的输出结果。计算机令人满意，也会令人沮丧。如果人们不向它提供非常简明清楚的指令，计算机可能一事无成。

计算机有两类基本软件。一种是应用软件，包括为特定用户服物的指令；另一种是系统软件，它是属于整个计算机系统，而不是

属于任何个别用户的程序的集合，通常执行支援性功能。操作系统（系统软件的主要类型）是一些计算机存贮器的单个区域内组合在一起的专用程序。这些专用程序，用户不能改变。当很多应用程序需要同时处理数据时，这种管理性软件就可以在内部协调计算机资源。这使分时（把机器时间分成许多小的时间段使用）安排（通常是在大型计算机中心）成为可能。这种具有分时操作能力的计算机允许多个用户几乎同时从一个中央处理器（CPU）中存取信息。

系统软件是大规模使用计算机的基础。它使电子数据处理（EDP）设备经济实用成为可能。不过，档案工作者更多地关心应用软件，以及下面将提到的关于与应用软件有关的那些程序系统的说明。

档案工作者应该对软件系统（软件包）有所了解。需要了解的第一件事可能是并非所的程序在各种计算机上都能运行。要求一个软件系统在各种计算机上都能运行的想法听起来好象是很合理。但要具正做到这一步是不可能的。

需容问题上受到限制一般是由于编写软件的语言。计算机程序由程序设计语言或称为源语言编写。在使用这些语言时，须用计算机系统软件把指令翻译成该中央处理机能够进行处理的二进制机器代码（目标代码）。程序设计语言有很多种，用的最广泛的是FORTRAN（公式翻译语言）和COBOL（面向商业的通用语言）它们可以用于不同类型的计算机。其它高级语言有的是为了满足某些特定用户的一致需要，有的是为了某一种型号的计算机设计的。汇编语言就是厂家为了某一种特定的计算机设计的。它与机器语言很相近。虽然，在一些系统中，可以用几种语言写程序，但是在大

多数系统这样做是不行的。由于一个系统可以包括很多个程序，因此，用另外一种语言重写一个系统的程序，就要花费很多心血。由此可见，一个软件系统可移植的灵活程度受到编写程序所使用的语言的影响。所谓软件系统的可移植性是指软件系统可以在不同型号和类型的计算机上使用。

向计算机的主存贮器传送一组程序的方法和传送其它数据或信息的方法相同。程序一经存储在存贮器中，程序就会告诉计算机数据是怎样组织的，怎样识别它以及用它进行各种运算。一个自动化信息检索系统由一个数据库和软件系统组成。数据库中包括了档案馆和它的行政管理和业务工作需要所使用的全部信息。用软件；或叫计算机指令进行检索分类打印部分或全部数据。这样一个系统在工作人员的努力下能做很多工作。

本章除了对在档案工作中有特殊定义，同时在计算机检索系统中也有特定定义的一些名词术语作了如上的一些解释外，也希望能对档案工作者结合自己的工作实际，着手考虑计算机应用的途径和方法等方面的问题有所启示和帮助。

# 第 二 章

## 档案工作者和计算机系统

本章主要介绍一些国家的档案自动化系统和档案工作者使用这些系统的情况。在简要地记述早期档案自动化状况之后，本章将综述档案自动化在一些国家的近期发展情况和现状。

### 一、档案自动化的产生和发展

在二十世纪，档案文件产生的速度是惊人的。这种情况导致了档案库数量增加和档案本身业务工作量的增加以及档案这门专职业务的发展。虽然增加档案工作者的人数可以使这个问题得到缓和，但是人员的增加速度根本跟不上档案文件数量、库房数量和档案业务量的增加速度。而且也不能从根本上解决问题。为此，档案工作者不得不发展新技术。

档案工作者面对档案文件数量急剧增加的同时，还遇到了档案种类的增多。在档案利用方面，随着人们利用档案进行研究各种问题的广泛和深入，档案利用的数量增加了，利用的范围也扩大了。使用传统的查找工具常常不能满足利用者和档案研究人员的需求。他们试图超越档案机构的传统界线，希望能够在一个地方，以一些共同的检索角度出发在一个或几个档案库中找到保存的不同类型的档案资料。档案馆馆藏量的增加和档案利用范围的扩大也使得档案工作者不得不发展新的技术，寻求利用的工具。