数据组织的层次体系

任何信息系统都有一个数据组织的层次体系。在该层次体系中，每一后继层都是其前驱层数据元组合的结果，最终实现一个综合的数据库。处于第一层的"位"用户是不必了解的，而其它五层则是用户输入和请求数据时合理的需要。数据是一切信息系统的基础。一个高质量的计算机信息系统的最终用户必须具备数据的组织及其处理方面的知识。

位是主存储器和辅助存储器的基本单位。计算机是电子的，因而只能实现两种状态。从物理上讲，可以通过不同途径来实现这两种状态(电流的方向，开关，涂在带上和盘上的铁淦氧的磁性排列)。由于每一位只能表示两种状态，因此，必须将位组合才能形成字母数字字符。由位组合成的字母数字字符被暂时存放在主存储器中，或永久地存放在辅助存储器中。在主存和辅存中存放的是字母数字字符的内部表示形式(例如，如果采用 EBC-DIC 编码体制，则 11000010 表示字母 A，而 11110001 表示数字 1)。

在输入时，对字母数字字符进行编码以形成若干位的组合，而在输出时进行译码。目前还没有工业标准的编码体制。最为流行的编码体制是六位二进制编码的十进制码(BCD)，七位 ASCII 码以及八位扩充二进制编码的十进制交换码(EBCDIC-发音为 eb-se-dik)。

六位编码最多可以表示 64 个字符(2?6)。七位编码可以表示 128 个字符，而八位编码可以表示 256 个字符。读者可能会问：既然用六位就可以对一个字符编码。为什么还要用八位来编码?这是因为六位码的 64 种可能的组合只够表示字母、数字和 18 个特殊符号。如果希望有表示大写和小写字母，那么六位编码就不够用了。因此，就需要具有 128 种组合的七位编码。

目前还难以想象出对 128 种以上的位的组合需要。引进八位编码体制(EBCDIC)是为了利用这一个事实，即只用 4 位(24-具有 16 种可能的组合)来表示一个数值数据。因此，一个 8 位的编码实际上可以用来表示两个十进制数字。由于所存储的数据多数是数值数据，所以将两个数字的编码压缩成八位可以节省存储空间。EBCDIC 的 8 位组合称之为一个字节。而 BCD 的六位就构成一个字节。在 BCD 和 ASCII 编码体制中，字节是字符的同义词。在 EBCDIC 编码体制中，由于可以将两个数字压缩到一个字节中，所以 EBCDIC 的字节与字符间并不一一对应。然而，在涉及到存储容量时，则经常交替地使用字符和字节。一个磁盘组可以有 800 兆字节容量(即 800 兆字节的永久存储器)，而一台计算机的主存可以有 8 兆字节(作为处理用的兆字节的高速临时存储器)。较小的存储设备用千字节(一千个字节的倍数来度量)。通常将兆和千分别缩写"M"和"K"。

在逻辑上讲，一个 EBCDIC 字节是 8 位，而实际上它有 9 位。由于要将这些位在计算机和外部设备(或远程终端)之间传送，所以在计算机硬件中使用了一种内部校验方法来保证传送数据的准确性。这种构验方法之一是给传送的数据附加一位奇偶校验位，用该位来发现在传送过程中是否丢失了一位。计算机可以采用偶数奇偶校验或奇数奇偶校验法，即每一字符要包含偶数个或奇数个"开状态"位。假定某台计算机采用偶数奇偶校验法，如果要将一个 EBCDIC 的字母 A(它具有奇数个"开"位-11000001)写到磁带上，那么在传送之前为了维持偶校验，则需要增加一位奇偶位(即：111000001—偶数个"开"位)，在将字符写到磁带之前，硬件自动计算"开"位的个数。如果计算机结果是奇数，则说明已经出现了奇偶校验错误，计算机自动向操作员发出警告。

字符(字节)

在通过键盘(光符号识别器或其他输入设备)输入一个字符时，机器直接将字符翻译成某特定的编码系统中一串位的组合。一个计算机系统可以使用不止一种编码体制。例如，某些计算机系统中将 ASCII 编码体制用于数据通信，而将 EBCDIC 编码体制用于数据存储。

数据元

描述数据元的最好办法是举例说明。一个人的社会保险号、姓名、信用卡号、街道地址和婚姻状况等都是数据元。在数据的层次体系中，数据元是最低一层的逻辑单位，为了形成一个逻辑单位，需要将若干位和若干字节组合在一起。一个日期不一定是一个数据元，它可以是三个数据元：年、月、日。对地址来说，也是同样的。一个地址中可以包括州、城市、

街道地址和邮政码这四个数据元。从逻辑上可以把日期和地址都看成是一个数据元，但是输出这种数据元是不方便的。例如，通常在输出时总是把街道地址单写一行，因而应该把一个地址的几个数据元分开。此外，由于姓名和地址文件经常按邮政码排序，因此，需要将邮政码作为一个逻辑实体(数据元)来对待。

根据上下文的需要，有时也把数据元称作为字段(记录中的字段)。数据元是泛指的，而数据项才是实际的实体(或实际的"值")。例如，社会保险号是一个数据元，而 445487279 和 44214158 则是两个数据项。

为了节省输入数据时敲打键盘的时间和存储空间，在输入数据时通常将数据元编码。例如，通常将职工主文件中的"性别"数据元编码，这样，数据录入员就可以简单的输入"M"或"F"来代替"Male"(男)或"Female"(女)。在输出时再将"M"和"F"分别翻译成"男"或"女"。

记录

将逻辑上相关的数据元组合在一起就形成一个记录。表 20.6.2 列举了一个职工记录中可能包含的若干数据元，以及作为职工记录的一个值的若干数据项。记录是能够从数据库中存取的最低一层的逻辑单位。

文件

文件是逻辑上相关的记录的集合。职工主文件包含每一个职工的记录。库存文件包含每一种库存货物的记录。应收帐目文件包含每个顾客的记录。"文件"这个词有时也指某台二级存储设备上的一块已命名的区域，该区域中可以包含程序代码、教材、数据，甚至还可以包含输出报表。

数据库

数据库是一种作为计算机系统资源共享的全部数据之集合。有时根据不同应用领域可将该资源共享数据分成若干段。例如，财会数据库可以划分为一个应用领域，它可以包含六个不同的文件。读者应该注意到：用"文件"来组织数据这种方法将带来数据的冗余。也就是说，为了在处理时使用，必须将某些数据元重复地存放在几个文件中。例如，在一所大学的安置办公室、宿舍管理处、财务支持办公室以及注册处等都有可能保存学生文件。像学生名、校内地址这类数据元几乎在每个文件中都重复出现。在对开发一个综合的学生信息系统进行可行性分析时，一些系统分析员在美国西南部一所规模很大的大学中发现有 75 个计算机文件中都包含学生名和校内地址。采用先进的数据库管理系统比之传统的文件系统有较大的改进，它使得用户可以将存储数据的重复程度减至最小。