国外科技动态

世界微生物信息系统——MICRO-IS 简介

马 俊 才

（中国科学院微生物研究所，北京）

笔者于一九八六年在日期间考察了设在日本理化学研究所的世界微生物数据中心（World Data Center of Microorganisms, WDC）、早稻田大学微生物系等单位，并借助于日本科技情报中心的科学数据库，对近年来日本微生物界的计算机应用进行了检索。

（一）世界微生物数据中心（WDC）的现状与职能

1. WDC 的现状： 1970 年 WDC 设立于澳大利亚昆士兰大学[1]。 1972 年经过其所长 Skerman 先生的努力并在联合国教科文组织 UNESCO，联合国环境总署 UNEP 等国际组织的帮助下，出版了世界微生物保存联合会（WFCC）的第一版《世界菌种保藏目录》。1982 年出版了用计算机编辑的第二版《世界菌种保藏目录》。1986 年 5 月 WDC 移至日本理化学研究所。 1986 年又出版了第三版世界菌种目录。 目前 WDC 保存了世界上 56 个国家 327 个保藏单位的生物资源数据，范围遍及细菌、酵母、真菌、藻类、地衣类、原生动物、细胞培养物、动植物病毒、细菌病毒、昆虫病毒等。目前在 WDC 注册的国家中保藏量大于 5000 株菌的保藏机构已有 20 个（表 1）。中国也在这 20 个之内。

表 1 目前在 WDC 注册的国家中保藏量大于 5000 株的机构

序号	单 位	真菌	酵母	细菌	名 称
1	ATCC	13000	5000	12000	美国标准菌株保藏中心
2	CY	0	0	18990	法国
3	NRRL	32000	14000	17000	美国农业部北方研究利用发展部
4	UAMH	4880	189	28	加拿大 Alberta 大学霉菌标本室
5	PDDCC	2000	0	5000	新西兰植物病害菌株保藏部
6	CCFC	8800	0	0	加拿大
7	FGSC	5500	0	0	美国
8	CNCTC	100	50	5000	捷克斯洛伐克国家菌保会
9	CBS	25000	4000	1100	荷兰真菌中心收藏所
10	WRL	300	40	6500	英国
11	UPJOHN	2600	330	6058	美国 Upjohn 公司菌保部
12	IFO	6737	2324	2825	日本大坂发酵所
13	MW	5306	31	0	东德
14	DSM	1200	300	4000	西德菌种保藏所
15	LMG	0	0	7000	比利时
16	LSCC	0	0	6000	加拿大
17	VKM	2221	2498	1845	苏联科学院微生物生化生理研究所
18	CCCCM	4000	1000	3000	中国微生物菌种保藏管理委员会
19	CEPIM	0	0	10000	意大利
20	CCUG	0	300	18000	瑞典

2. WDC 的职能[2]：

（1）收集整理世界各国菌种保藏机构的情况。

（2）出版全部微生物的目录。

（3）不断充实菌株的生理、生化及遗传性状数据库。

菌种目录所包含的主要内容：

（1）系统保藏机构的编号。

（2）系统保藏机构的简称。

（3）系统保藏机构的地址、联系方法及现有设备等。

（4）系统保藏机构的重点领域。

（5）所保藏的生物资源属种名。

（6）对所保藏的生物资源能提供的性状项目。

（7）系统保藏机构是否提供鉴定等服务项目。

（8）收费标准。

（9）有无专利菌株寄存制度。

（10）有无出版的菌种目录。

本文承蒙王大耜、徐浩、赵玉峰先生以及清华大学曹竹安教授悉心审阅，特此致谢。

(11) 菌种数据的最近更新日期。

(二) 微生物性状的编码方式

1. 微生物信息管理系统的几个特点：近年来随着生物技术的发展，微生物生理生化及培养性状急剧增加，使用计算机进行性状数据处理已势在必行。但是微生物信息管理系统与通常的数据管理系统相比有其不同之处。通常的数据管理系统由于其数据项目数不很多，字段长度也不很长，一般都是一个项目为一个字段(FIELD)，而字段长度取该项目的最大字节数即可。而微生物信息管理系统由于其性状数目极大(常在几百以上)，性状类型随菌株类别、研究者的不同差异很大，性状的文字描述复杂而冗长，显然使用一般的数据管理系统模式是不能满足其要求的。从 70 年代开始，国际微生物界提出微生物信息管理系统必须遵循以下条件：

(1) 整个系统设计必须基于一个已经确定了的编码系统。

(2) 编码系统适用于全部菌株的文字型、二值型、数值型数据，并且新编码的追加不能对老编码系统产生影响。

(3) 用户必须能以对话的方式对系统内的数据进行处理。

(4) 系统使用要简便易学。

(5) 系统要适用于所存储的所有数据，并有针对微生物数据特点的检索方式。

(6) 要有数据的保密措施。

满足这些条件的有 RKC 编码系统，以及与之相匹配的 MICRO-IS 管理系统 (Microbial Information System)。

2. RKC 编码的意义：所谓 RKC 编码系统是指 Rogosa (美国国立卫生研究院 NIH)、Krichevsky (美国国立卫生研究院 NIH)、Colwell (马里兰大学微生物系)三人于 1971 年发表的《微生物存储、检索用编码方法》。人们习惯用三人姓名的首字母称这一编码系统为 RKC 编码系统。

RKC 代码是一种可以将微生物性状的自然语言描述转变成数值描述的一种编码方式。使用这种编码可以将微生物性状数据这一原本为文字型的数据转变为数值型数据库，从而大大节省系统资源，特别是当菌株数目很大时，这种节省尤为显著。例如，我们目前正在筹建的微生物资源数据库的细菌性状子库，使用 RKC 代码后(每株菌取 2000 个 RKC 码，每个 RKC 码的描述以 100 个字符计算)，仅占系统存储容量 6.2 mol B。而不采用 RKC 代码(每株菌取 200 个有效性状，每种性状描述以 100 个字符计算)，需占 60 mol B。相差近 10 倍。同时使用 RKC 代码后，由于可以使用列查找方式，在 3000 株菌中检索时，最多只进行 3000 次磁盘 I/O 操作，而不使用 RKC 代码方式，则最多需进行 600000 次磁盘 I/O 操作。相差 200 倍。所以使用 RKC 代码还可以大大节省检索时间。

另外由于微生物性状的文字描述复杂而冗长，很难使性状描述规范化。由于 RKC 代码几乎将微生物的所有性状都进行了编码，从而很大程度上解决了微生物性状的规范化问题。

3. 编码方式：RKC 系统将微生物信息分为 38 个节 (SECTIONS)，每节均为微生物某一个方面性状，第一、二节为菌名、菌株鉴定者、来源等文字型数据；第三节以后为形态和生理生化等性状。每个性状均为一个由 6 位数字组成的编码。整个编码系统的总编码数为 9 千多个。

在 RKC 代码系统中对于可以用 Yes/No 表示的性状，为其设立一个编码。例如，025251 编码为是否能利用果糖产气；对于多态性状，则采用一种状态一个编码的方法。例如第八节《分枝、菌丝、产生无性孢子》一节中的次生菌丝颜色一项，共有 8 种颜色，其编码为：

008377：红色；008378：黄色；008379：绿色；008380：兰色；008381：紫色；008382：白色；008441：黑色；008389：灰色；

对于数值型性状，则采用按区间编码的方法。如第七节《粘孢子、孢囊孢子、配子细胞》中的孢囊长轴一项，孢囊孢子长轴长度：(μm)

007077：<0.5；　　007030：0.5—2.0；

007031：2.1—5.0；　007032：5.1—10.0；
007033：10.1—20.0；　007034：20.1—50.0；
007035：50.1—100.0；　007036：100.1—150.0；
007036：>150；

在 RKC 系统中，所有数据均作成表结构。对于二值型数据，其数据结构如表 2。微生物数据中通常所说的二值数据一般是指“1”(Yes)、“0”(No)，此外还有“Blank”(即 NC，没有或未做)。实际上是三种状态，如表 2 所示。

表 2　二值编码的数据结构

菌名 \ 性状的 RKC 码	025251	025252	025253	025254	025255	025256	025308	025309	025310	025313
LCPR12	1	0		0	1	1		0	1	1
S5.15	1	0		1	1					
RKN	0	1	1	0		1	1	0	1	0
L/N75	1	1	1	1	1	1	1	1	1	1

表 2 中代码 025251 是利用果糖产气。菌株LCPR12，S5.15，L/N75 均能利用，则填入 1；菌株 RKN 不能利用则填入 0；未测定项或在没有数据的时候留下空格 (BLANK)。需要增加菌株和性状时，只需横、纵向延伸这个二维表即可。

RKC 代码系统本身已经提供了数值型数据按区间编码的方式，但是如果用户由于某种原因仍希望直接填入其数值时，可在该代码的后面续若干列 N，如表 3 中的 025251 项，共延续了 2 列 N，则说明该项数值共占了 3 列。四株菌的数值分别为 1.0，0.5，0.0，0.25。

RKC 代码系统还允许增加新的代码。这一工作目前是由 NIH 的系统微生物学部(Microbial Systematics Section) 管理的。需要增加新代码时需先向 NIH 申请，得到认可后方可使用，以保证新代码的增加，对以前的代码系统不产生影响。在得到 NIH 的认可之前可以自己在第 98 节内先设立新的代码，并可在第 99 节内记入有关的注解。

(三) 微生物信息系统 (MICRO-IS) 的功能

表 3　数值编码的数据结构

菌名 \ 性状的 RKC 码	025251	NNN252	NNN253	025254	025255	025256	NNN258	NNN259	NNN260	025263
LCPR12	1	·	0	0		0	4	·	0	0	1
S5.15	0	·	5	0		1	1	2	·	0	
RKN	0	·	0	1	1	0					1
L/N75	·	2	5	1	1	1	0	·	5		1

1. MICRO-IS 的构成[3]：MICRO-IS 主要由文件生成模块、检索模块及数值分类、鉴定等统计处理模块组成(见图 1)。

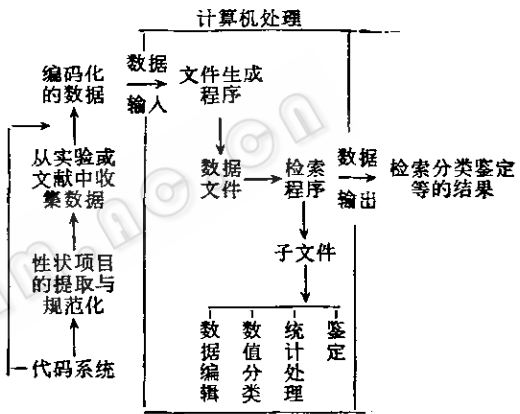

图 1　MICRO-IS 数据流程图

2. 数据格式的生成方法及存储格式：首先使用计算机所提供的文本编辑功能，把按照 RKC 系统代码化了的数据，以每行 80 字符的形式作成卡片。并使用 DATA. DES 程序，将数据的项目、属性、在数据文件中的存放形式等，作为格式文件记录下来。

CREATE 程序把输入的内容同格式文件相比较，如果两者间的属性及存放形式都正确的话，则生成数据文件。如果两者间有不同之处的话，就把有关的信息输出到终端。使用 CRDATE 程序还可以给数据文件定义口令，以防止数据文件被改写或随意使用，使数据得到保护。如表 4 所示，数据文件是按照菌类的不同以表格 (FORM) 形式存放的。表格中顺序存放着菌株的记录 (STRAIN RECORDS)。文

表4 数据文件的结构

项目名称	项 目 内 容
文件指针	表格数及在机内的属性
表格指针	口令，表格编号，表格内的数据位置
表格字段	第一节中的字符型数据(420字28项)
表格二值型性状	第一节中的二值型性状值(70项)
二值型性状代码集	该群二值性状的 RKC 代码(600项)
数值型性状代码集	该群数值性状的 RKC 代码(300项)
(菌株A的记录)	
菌株指针	该菌的数据长度，二值数值数据的位置
菌株字段数据	第二节的字符型数据(140字41项)
菌株二值型性状	该菌的二值型性状值
菌株数值型性状	该菌的数值型性状值
(菌株B的记录)	
菌株指针	该菌的数据长度，二值数值数据的位置
菌株字段数据	第二节的字符型数据(140字44项)
菌株二值型性状	该菌的二值型性状值
菌株数值型性状	该菌的数值型性状值
(菌株C的记录)	
菌株指针	该菌的数据长度，二值数值数据的位置

件指针 (FILE HEADER) 中记录着表格数以及在机内的属性等，表格指针记录着口令、表格编号、表格内的数据位置等参数，然后是记录着与各类群有关的信息，即第一节(一般信息)中的字符型数据 (FORM FIELDS) 和二值型数据值 (FORM BINARY ANSWERS)。在此字符型数据长度被限制在420个字符(28个项目)，二值型数据被限制在70个项目以内。

接下来是表示这个类群菌株的全部二值性状和数据性状的 RKC 代码。二值性状代码集限制在600个项目以内，数值型性状限制在300个项目以内。然后是顺序记录的各个菌株记录。其后是记录着各个菌株数据长度及二值型、数值型数据位置的菌株指针，接下来是菌株字段数据 (STRAIN FIELD DATA)。它以字符型数据的形式记录第二节(特定菌株信息)的内容，这个数据限制在140个字符(44个项目)以内，最后记录第三节以后各节的二值数据代码值和数值型数据代码值 (STRAIN NUMBERICAL ANSWERS)。一个表格中的菌株数不得超过3280株。

这种结构使得数据文件具有比输入文件更压缩、更适于检索的结构。例如，一个二值数据在输入文件中占一个字节(8个 BIT)，而在此数据文件中只占2个 BIT，大大节省了存储空间。并且针对每个数据文件、表格及菌株记录都设立了指针，具有相当于数据目录的功能。因此检索效率很高。

用 QUERYN 程序对以上构造进行检索。

3. 检索条件式与检索命令：

(1) 检索条件式的组成：

①二值型数据的检索条件式：

对某个 RKC 码所代表的性状(如025251)为1(Yes/+)的菌株检索时，条件式为：

(25251)

对某个 RKC 码所代表的性状(如025251)为0(No/−)的菌株检索时，条件式为：

(NOT 25251)

对某个性状为空(未测定)的菌株检索时，条件式为：(BLANK 25251) 在此 RKC 码最左端的0可以省略。

②数值型数据的检索条件式：由 RKC 码、BLANK、大小关系符(>=<)和数值(整数、实数及带指数的实数)组成。如：(525251<2)，(525251 > 4.6E12)，(BLANK 525251)，其中(525251 > 4.6E12) 为该值大于 4.6×1012。

③字符型数据的检索条件式：由项目名或 RKC 码、BLANK、大小关系符、数值及用引号括起来的文字组成。如：

(GENUS = "LACTOBACILLUS")

(DATE-ISOLATED > 101572)

(BLANK SPECIFIC-EPITHET)

其中 BLANK SPECIFIC-EPTHEI 指种名为空的菌株。

④复合检索条件式：在数据项和检索条件式中使用 AND，OR，ANY 及括号构成的逻辑表达式。如：

(29003 AND NOT 29006)

[(BLANK SPECIFIC-EPITHET AND GENUS = "PROTEUS") AND (29001 OR 29002)]

[ANY (2 29001 NOT 29002 29003)]

最后一例为在29001、NOT 29002 和29003三个条件中满足任意两个条件。

(2) 检索命令的组成: QUERYN 程序可以解释从终端接受的检索命令，如果没有语法错误的话，将按照检索命令的要求从数据文件中检索所要求的数据，其结果从终端上输出或作成数据传输用的子文件。如果检索命令有语法错误，QUERYN 程序在终端输出错误信息，改正后重新进行检索。

检索命令由命令语、待检索文件名、待输出信息项以及检索条件式组成。命令语由GIVE、EXTRACT、SAVE、SHOW、REPORT、TABULATE 及 END 等7种形式组成。待检索文件名在 QUERYN 开始已时指定为 MAIN1，使用命令 EXTRACT 时，可以从 MAIN1 中取出数据，所作成的子文件为 SUBnn (nn 为自然数)。检索项目指定几乎都使用 RKC 代码。由于 DATA. DES 程序已经在格式文件里对字段数据 (FIELD DATA) 中字符型数据的项目名称进行了登录，所以在指定这些项目时，使用那些已经登录了的项目名称即可。检索条件式由 RKC 代码以及 AND、OR、NOT、BLANK、ANY 和 ALL 等关键字组成。检索命令可以在每个词或记号结束后换行，并且可以跨越数行。

(3) 检索命令的功能:

① GIVE: 在进行检索之前，希望确认自己所要的数据 (RKC 码)是否在数据文件中存在时使用本命令。例如

GIVE MAIN1 25000 TO 25999: 表示输出数据文件 MAIN 1 中所包含的第25节的 RKC 码清单及代码总数。

GIVE MAIN1 FIELDS: 表示输出数据文件 MAIN1 中所包含的字符型数据的项目名、RKC 码及数据长度的清单。

② EXTRACT: 这是一条将菌株数据作成子文件的命令，并且这个子文件同用 CREATE 程序生成的数据文件的结构是一致的。同时系统将按照生成顺序自动给子文件付上 SUB1、SUB2··· 的文件名。目前执行一次 QUERYN 可最多生成21个子文件。

EXTRACT MAIN1 (25193 AND NOT 24450) 表示把所有 RKC 码25193为 Yes 且24450为 No 的菌株数据从 MAIN1 中取出，生成 SUB1 子文件。EXTRACT 不但可以从 MAIN1 中取数据生成子文件，还可以从子文件再生成子文件。

③ SAVE: 用 EXTRACT 生成的子文件，当 QUERYN 结束时，也自动被清除。如果需要保存的话就使用本命令。

SAVE SUBnn 所保存的文件可以用于以后的数据检索以及在 QUERYN 以外的鉴定、分析等处理。

④ SHOW: 按指定的信息项输出满足检索条件的菌株的内容。例如:

SHOU MAIN1(STRAIN 2004 25181 TO 25197) (25197)

表示从 MAIN1 中取出 RKC 码25197为 Yes 的全部菌株，并只显示它们的属名 (RKC 码2004) 及 RKC 码025181到025197范围内的性状内容。

如果在检索条件式中使用 ALL，则按指定的信息项输出全部菌株的内容。如果在信息项目表中使用 NOTHING，则只显示满足该条件的菌株数目。

数据的输出以20个菌株为一个单位进行，所以在输出过程中，可以终止其输出。比如当满足某检索条件的菌株很多时，可以终止其输出，打入必要的命令，使其在打印机上输出。

⑤ REPORT: 本命令生成一个输出报告格式文件，并记录在系统中。在输出的时候，如果该文件已经存在的话，系统将按照该文件所规定的格式进行显示或打印输出；若该文件不存在，系统将自动生成一个标准的格式文件，并依此输出。

⑥ TABULATE: 这是一条对二值或数值数据进行统计的命令。例如:

TABULATE MAIN1 (24009 TO 25184) (ANY (2 NOT 2518125184 25190)): 表示

输出满足检索条件的菌株总数及 RKC 码 024009 至 025184 范围内分别为 Yes、No、BLANK 的菌株数。

数值型数据时，输出最大值、最小值以及标准差。此时不包括那些数据为 BLANK 的菌株。

⑦ END：这是一条结束 QUERYN 命令。这条命令将使控制权从 QUERYN 转为计算机系统。

4. 其它功能：MICRO-IS 系统除了上述程序以外，还有对数据进行修改、删除、追加的 FORMEDIT 程序；为了某些使用目的，而将某一种格式的数据文件传送给另一种格式数据文件的 MOVESTRN 程序；进行数据文件拷贝的 COPY 程序。

应用程序 IDDNEW 是用来进行鉴定用的。将微生物的性状按一定格式进行整理后，可使用该程序进行鉴定。此外，使用 QUER YN 的 REPORT 功能，可以不去特别编写文件变换程序，就可以容易的将子文件与研究者自己的各种分析程序连接起来。

目前 WDC 的计算机系统，已经可以通过日本电信电话株式会社（NTT）和国际电信电话株式会社（KDD）的信息交换网与国内外进行联机检索。同时还可以使用该网络进行各种信息交换。

（四）MICRO-IS 的软硬件环境

该系统最初运行在 IBM 370/168 上，由美国国立卫生研究院（NIH）的 Krichevshy 开发，软件语言主要为 PL/I。目前已出现微机版本。

参 考 文 献

[1] 赵玉峰：微生物学数据库，分析微生物学专集，p. 269—289，科学出版社，1988。

[2] 菅原秀明、饭野义男：情报管理，26(8)：640—648，1983。

[3] 坂本直人、菅原秀明：化学と生命，18(6)：352—363，1980。