Account of the first ICTV virus data sub-committee workshop on database descriptions

Doorn, The Netherlands, 30th January – 4th February 1994

Participants: Dr. Lois Blaine (LB), USA
Dr. Cornelia Buchen-Osmond (CBO), Australia
Dr. Charles Calisher (CC), USA
Dr. Mike Dallwitz (MD), Australia
Dr. Tony Della-Porta (TD), Australia
Dr. Adrian Gibbs (AG), Australia
Dr. Marian Horzinek (MH), The Netherlands
Dr. Brian Mahy (BM), USA
Dr. Jack Maniloff (JM), USA
Dr. Mike Mayo (MM), UK
Dr. Fred Murphy (FM), USA

Guests: Dr. F. MacIntyre, ETI, Amsterdam
Dr. R. Sluys, ETI, Amsterdam
Dr. E. Gouda, Hortus botanicus, Utrecht

Location and sponsorship of the workshop

The workshop took place in Opleidingscentrum Bondsspaarbanken, Doorn. It was made possible because of sponsorship by (in alphabetical order): Bayer, Codata, EFFEM, European Union, IUMS, Merck, Roche and Virology Division.

Structure of the workshop

To illustrate the use of databases in other fields, LB described the Bergey's Bacterial Database, Dr. Sluys described the production by ETI of a range of biology databases, MD described the use of the DELTA language to generate databases of grasses and beetle larvae, and AG described the plant virus database developed using DELTA. These gave a background to discussion of how the universal virus database (UVD) should be structured. These discussions by the whole VDSC aired many issues and were followed by the division of the task of finalizing descriptor lists into small groups. The descriptors generated or selected by these groups were then presented to the whole VDSC for discussion.

The salient features of the various discussions and presentations are grouped under general topic headings. Actions decided on are grouped at the end of the Report.

1. Presentations

LB – The Bergey bacterial database

Bergey's is a private trust. Book sales generate enough income to finance the database. The purpose of the database is primarily to facilitate production of the hard copy descriptions. Text and Tables are supplied by volunteers. These are converted to RKC codes by fitting to standard descriptors. The contributors confirm the assignments and these are then entered using MICROIS. Williams and Wilkins fund Dr. M. Krichevsky to improve MICROIS and ATCC to do parsing and vocabulary assignment. The royalties from the Bergey's Manuals finance this. Williams and Wilkins own the format but Bergey's Trust own the data.

Dr. R. Sluys – ETI database

Funded by a combination of UNESCO, Dutch Government and University of Amsterdam money, ETI are constructing a 'biodiversity database'. ETI employ about 10 taxonomists and 10 programmers and work by using group editors to contact experts whom they supply with Linnaeus II software. The software did not seem as sophisticated as DELTA. The products made are CD ROM. It was suggested that this format might be more useful for 3rd World countries than book products. This method can be cheaper for supplying colour pictures than traditional hard copy.

MD – DELTA

DELTA has been agreed internationally as the standard format for taxonomic data input. (It had been agreed by previous committees, and it was repeated at the workshop, that DELTA will be used for UDV). DELTA input is connected by CONFOR to a key generator program or a text generator program for output purposes. For identification purposes INTKEY is used to examine files created for it by CONFOR. These programs, or others, can be used after data are entered in DELTA format. The core of DELTA format is characters assigned to taxa. A character is a feature line (descriptor, attribute) + a character state. Care is needed so that states are exclusive but able to accommodate an ambivalent state. Characters should also be constructed so as to facilitate a direct computer-based print-out of the characters of a taxon, or even a key. A WINDOWS version of INTKEY is being written and could well be done by 2 years from now. INTKEY is versatile. It can, for example, detect deficiencies in data sets by using DIAGNOSE.

AG – The plant virus database

This database has been used to generate book products. This is an immediate way of generating income. The database uses 570 characters, but an increase is projected, and 890 species have been input. About 200 characters are probably relevant to all viruses. This set of descriptors has been used as one source for the universal descriptor set.

Dr. E. Gouda – DELTA data processing

Dr. Gouda demonstrated software (TAXASOFT) designed to facilitate the input of data into DELTA format. The software impressed those familiar with the problem and was thought to hold promise for assisting with entering virus data. Some modifications might be needed because of the size of the character set.

CBO – Development work on the descriptors

Descriptors have been assembled from the list compiled for the ATCC Workshops, 1991 and by extracting descriptive statements from the Vth ICTV Report. The list (of 1000 descriptors) is not yet complete. The problem involved in assembling descriptors became increasingly apparent during the workshop. A decimal numbering system has been devised which will be adopted when data are entered. The principles outlined were:

1. Use the DELTA system
2. Select properties from a character list composed of
   - multiple choice statements
   - numerical statements
   - free text statements for comments
3. Specify each character unambiguously

CC – Experiences with DELTA entry for some 'arboviruses'

CC showed a scheme of 294 characters which commenced with biological properties. This scheme was used as one source for developing a list of descriptors for biological characters.

2. Work on developing a descriptor list

The characters being described were divided into 7 types. Introduction and history, classification and taxonomy, virion properties, replication, host and disease, ecology and control (i.e. epidemiology and transmission) and diagnosis and control. Classification and taxonomy was settled by a general discussion. Other parts were attempted by small working groups. Based largely on the descriptor list assembled by CBO, descriptors were selected for the virion morphology and its chemical composition (virion properties). Some progress was made in other areas. These will be further developed by more work by individuals from VDSC (see Task List).

3. Data collection notes

The full database will be assembled from information collected about individual virus strains. An attempt will be made to avoid asking for input of higher taxon characteristics by sending out questionnaires partially completed. Experts would be asked to check these properties. Questionnaires would be sent together with floppy disks such that only relevant questions would appear. It was pointed out that the level of detail in the resulting database would exceed that of the 6th Report and therefore generation of future ICTV Reports would involve selection from the database, not its output complete. Detailed character notes will be needed. The questionnaire should, if possible, be the descriptor list asking for character state to be selected, but this will not always be possible. Answers can be selections of more than one state and could include ranking.

4. Dissemination of progress/promotion of activities

It was thought important to publicise the UVD exercise in order to improve changes of funding. It was decided to use the plant virus database as a 'flagship' for this as books have been and will again be published from the database. Also CD ROM output should be considered, possibly timed for release concurrently with the 6th Report. A provisional goal was set of producing output to coincide with the 7th Report (approx. mid 1997) as a 'ghost'

7th Report prepared from DELTA-formatted data. A working group was delegated to organise this. Publicity about the Doorn workshop would be in Virology Division News and if possible ASM News and SGM Quarterly.

Individuals were delegated to do this.

5. Delta training

The desirability of more VDSC members becoming proficient in DELTA was discussed. Some training was thought valuable although an organised training workshop should be left until more near the time when the bulk of the data are to be entered.

Task List

Finance and other VDSC business

MH, LB, AG, FM, BM: explore how to achieve long-term funding of the UVD project.
MH: consult Virology Division about the use of 'trust status' to assist in fund raising.
LB: continue liaison with Williams and Wilkins.
AG, LB: seek funding for a meeting in ca. 12 months of (for example) study group chairmen to promote enthusiasm for the next phase.
MH: contact Dr. Naik (Poona) about possible involvement with UVD.

Dissemination/promotion

MH, MM: prepare a note about the Doorn workshop for Virology Division News.
JM: to explore possible publication of such a note in ASM News.
MM: to explore possible publication of such a note in SGM Quarterly.
MH, AG, CC: explore the possibility of promoting the UVD by distribution of CD ROM or floppy disks of some data from the 6th Report.

Descriptor List

MAM: produce a draft list of descriptors for 'Replications' by consultation with appropriate colleagues.
CC, TP: produce a descriptor list for 'Transmission, control and diagnosis' (including antiserum detail).
AG, CBO: collect together existing and to be developed lists of descriptiors and circulate to VDSC for comment.
AG, CBO, TD, MD: after VDSC input, convert the character list to DELTA format, send to VDSC members for input of test data and do a test run. Report back to VDSC through MH.
MM, JM, TD: develop guidelines for use by those supplying data on receipt of input of VDSC members after seeing the complete descriptor list.