Hunt Institute for Botanical Documentation
5th Floor, Hunt Library
Carnegie Mellon University
4909 Frew Street
Pittsburgh, PA 15213-3890
Telephone: 412-268-2434
Email: huntinst@andrew.cmu.edu
Web site: www.huntbotanical.org

The Hunt Institute is committed to making its collections accessible for research. We are pleased to offer this digitized item.

Usage guidelines
We have provided this low-resolution, digitized version for research purposes. To inquire about publishing any images from this item, please contact the Institute.

Statement on harmful and offensive content
The Hunt Institute Archives contains hundreds of thousands of pages of historical content, writing and images, created by thousands of individuals connected to the botanical sciences. Due to the wide range of time and social context in which these materials were created, some of the collections contain material that reflect outdated, biased, offensive and possibly violent views, opinions and actions. The Hunt Institute for Botanical Documentation does not endorse the views expressed in these materials, which are inconsistent with our dedication to creating an inclusive, accessible and anti-discriminatory research environment. Archival records are historical documents, and the Hunt Institute keeps such records unaltered to maintain their integrity and to foster accountability for the actions and views of the collections' creators.

Many of the historical collections in the Hunt Institute Archives contain personal correspondence, notes, recollections and opinions, which may contain language, ideas or stereotypes that are offensive or harmful to others. These collections are maintained as records of the individuals involved and do not reflect the views or values of the Hunt Institute for Botanical Documentation or those of Carnegie Mellon University.

About the Institute
The Hunt Institute for Botanical Documentation, a research division of Carnegie Mellon University, specializes in the history of botany and all aspects of plant science and serves the international scientific community through research and documentation. To this end, the Institute acquires and maintains authoritative collections of books, plant images, manuscripts, portraits and data files, and provides publications and other modes of information service. The Institute meets the reference needs of botanists, biologists, historians, conservationists, librarians, bibliographers and the public at large, especially those concerned with any aspect of the North American flora.

Hunt Institute was dedicated in 1961 as the Rachel McMasters Miller Hunt Botanical Library, an international center for bibliographical research and service in the interests of botany and horticulture, as well as a center for the study of all aspects of the history of the plant sciences. By 1971 the Library's activities had so diversified that the name was changed to Hunt Institute for Botanical Documentation. Growth in collections and research projects led to the establishment of four programmatic departments: Archives, Art, Bibliography and the Library.

MEMORANDUM

To: See distribution below

Date: 30 April 1974

From: R.J. Pichel
Chief, AGPE

Subject: Film on documentation

During the series of documentation seminars held by Dr. D.J. Rogers, Documentation Consultant, an IBM film "Man and Computer - A Perspective" was shown. There have been many requests for a reshowing of this excellent film which shows the basic functions of a computer machine. The film will, therefore, be shown again on

Wednesday, 8 May 1974
at 16.30 hours
in the Austria Room C.234

Distribution:

Dr. Albani, AGP
Mr. Wrigley, AFM
Mr. Willan, FOR
Dr. Peterson, AGPC
Dr. Furtick, AGPP
Dr. Chiarappa, AGPP
Mr. Pichel, AGPE
Miss Bennett, AGPE
Mr. Maukonen, AGP
Mr. Wignell, AGP

PL 2/8
RJP/alz
cc: AGP Reg (2)
Chrono: Rogers

DISTRIBUTION:

Mr. Pichel, AGPE
Dr. Peterson, AGPD
Dr. Furtick, AGPP
Dr. León, AGPC
Ir. Dudal, AGLS
Dr. Fuggelli, FOR
Dr. Allan, FOR
Mr. Wrigley, AFH

All AGPE Staff

chrono: Rogers

AGP Reg. (2)

DJR/amad

PL-2/8

19.2.74.

OFFICE MEMORANDUM

TO: See list below DATE: 19 February 1974

FROM: R.J. Pichel,
Chief,
Crop Ecology and Genetic Resources Unit.

SUBJECT: Seminars on Documentation

Dr. D.J. Rogers, Consultant in Genetic Resources Documentation, has offered to present a series of five one-hour seminars to explain the concepts of documentation for genetic resources. The seminars will cover general concepts of information, development of data structuring, basic introduction to computer functions in information management, and a brief, non-technical description of the information storage and retrieval system, TAXIR (Taxonomic Information Retrieval).

Because many of the problems of documentation and information retrieval in the Division are similar, it is hoped that this will be a meaningful presentation to demonstrate the means by which many of the problems can be handled.

To avoid too much conflict with busy schedules, the seminars will begin at 16.30 hours, and will be scheduled every other day, starting on a Monday. The presentation will be informal, with opportunity to discuss various points.

Please indicate on the detached portion below your, and your colleagues' interest in attending the seminars. We need to determine the appropriate room size for the presentations.

PLEASE RETURN AS SOON AS POSSIBLE. A SECOND NOTICE WILL BE SENT ANNOUNCING THE STARTING DATE AND LOCATION.

TO: Mr. R.J. Pichel, Chief,
Crop Ecology and Genetic Resources Unit.

FROM:

SUBJECT: Seminars on Documentation: Attendance

I expect to attend the seminars

I cannot attend

The following colleagues will attend: _____

_____

_____

M E M O R A N D U M

To: See list below

Date: 25 February 1974

From: R.J. Pichel
Chief, Crop Ecology and Genetic Resources Unit, AGP

Subject: Meeting room and time for Documentation Seminars

Following our initial memo of 19 February, Dr. D.J. Rogers will present a series of five 1-hour seminars on general principles of documentation for genetic resources, as follows:

1. Meeting room: German Room - C.263

2. Starting date: Monday, 4 March 1974

   Continuing dates: 6 March, 8 March, 11 March and 13 March

3. Time: 16:30

PL 2/8
DJR/alz

cc: AGP Reg (2)
Chrono: Rogers
Capitani, AFS

Distribution:
Pichel, AGPE
Frere, AGPE
Sykes, AGPE
Rijks, AGPE
Zaniboni, AGPE
Bastianelli, AGPE
Chiarappa, AGPP
Fugalli, FOR
Lanly, FOR
Buongiorno, FOR
Willan, FOR
Palmberg, FOR
Marcharik, FOR
Riquier, AGLS
Marsch, FOR

+ Brandolini, AGP
Bruyckx, AGPP
Shuyler, "
Mrs. Zammarano
Singh

SEMINARS ON DOCUMENTATION

NAME	UNIT	DIVISION	PHONE	ROOM
Pichel, R.J.	CEGR	AGP	3598	C769
Sykes, T.	"	"	3585	C769
Frere, J.	"	"	4040	C854
Rijks, J.O.	"	"	4040	C854
Zaziboni, A.	"	"	4353	C790
Fugali,O.	FORM	FOR	3213	B503
Lanly	" Forest Resources Survey	"	4341	"
Marsch	" "	"	"	"
Buongiorno	"	"		B504
Willan, R.L.	" Afforestation	"	4143	B503
Palmberg, Miss C.	" "	"	4750	B503
Harcharik,	" "	"	"	"
Riquier, J.	Soils Resources Group Soil Survey	AGLS		B7L3
Chiarappa	AGPP		3441	C746

SEMINAR ON DOCUMENTATION

D.J. Rogers
Crop Ecology and Genetic Resources Unit

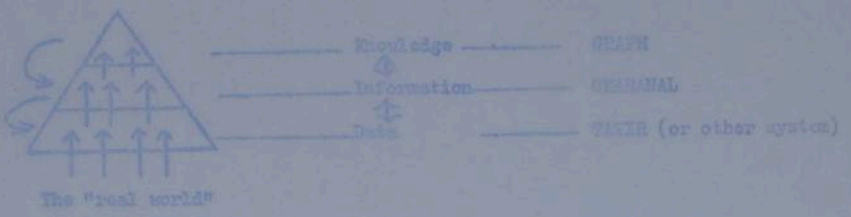
SEMINAR No. 1
4 March 1974

A. Introduction to the Seminar
   1. Background philosophy for "data management".
   2. Some understanding of computers — hardware and software.
   3. User's description of TAXIR, and some related programmes.

B. Why ? — a common basis to begin
   1. Many different systems for very specific purposes.
   2. Very few systems designed for all information needs.
   3. Very few systems are user-oriented.
   4. Need for a system that is general, to be used by people not expert in computers.

C. What is a "system" ?
   1. Generally, an organisational structure to accomplish a series of tasks.
   2. In CRU, all the functions associated with genetic resources.
   3. As a part of documentation, the computer programmes necessary to take care of the data.

D. Embedding documentation into a larger system
   1. The "triangle of knowledge".



E. Basic concepts for our documentation systems

1. Reflection (or model) of methods already known and understood, that is, the biological discipline of taxonomy.
2. The methods employed consider the structure of information, not the substantive content of information.

Example :

Words consist of letters (including consonants and vowels), and this is the structure of a word.

3. Example of the taxonomic structure useful in building an IR system.

a) The binomial scientific terms

Genus	- a noun	Triticum	- wheat
Species	- an adjective	aestivum	- a particular kind of wheat.

The above combination designates very precisely one kind of wheat.

b) Using the above structural format, in the IR milieu,

Descriptor	- a noun	colour
Descriptor state	- an adjective	red, green, blue, etc.

c) The Latin binomial is an exclusive system — the two terms allow no overlap with anything else.
Likewise, a descriptor and associated descriptor state are non-overlapping, or mutually exclusive

Example :

Descriptor - colour of stem
Descriptor states - red
yellow
red-yellow (this state used only for those plants exhibiting both red and yellow).

SEMINAR ON DOCUMENTATION

D. J. Rogers
Crop Ecology and Genetic Resources Unit

SEMINAR 2
6 March 1974

A. Review of last Seminar

1. The triangle of knowledge — our interest in the basal portion.
2. Comparative methods — from the model of taxonomy.
3. Generalities on methods of description.

B. Purpose of this Seminar

1. Further description of methods of data description.
2. To guarantee that our methods accomplish the necessary tasks.

C. Continuing discussion on the methods of description

1. The item : the object which forms the basis for description. Ex. : in genetic resources, the accession may be the item ; in other applications, a person specialising in some aspect of genetic resources function, etc.
2. Data may be extrinsic to the item (ex. : collection number), or may be intrinsic to the item (ex. : shape of some structure of the item)

D. The classification of types of descriptors (please see attached example)

1. Purpose of classification of types

a) To insure coverage of required information for any purpose.
b) To provide a framework to compare one set of data with any other similar set of data.

E. Structure of Descriptors, in terms of their function.

1. NAME descriptor : usually words — any combination up to 90 letters.
2. CODE descriptor : a form of the NAME descriptor, but shortened to facilitate recording. (ex. : the months of the year, represented by numbers, consecutively, from 1 to 12.
3. ORDER descriptor : Numerical or alphabetical, defined in a "from-to" format, giving maximum range of included data, from smallest to largest.

F. The concepts of a data bank

1. For sake of efficiency, necessary to organise data into logical sets.
2. For sake of costs in data gathering, need to keep within bounds of time, personnel capacity, and budget.
3. For sake of efficient computer use, need to keep sets of data within some limit of size.
4. When any one data bank becomes too large, may be subdivided, again using some logic.

Provisional, not fully developed,
CLASSIFICATION OF DESCRIPTOR TYPES IN GRU WORK

1. "Finding information"

Accession information
- accession numbers
- accession dates
- accession types

Collection information
- names of collectors
- collector's numbers
- collection dates

Nomenclature information
- scientific names
- common names
- numeric designations of individuals

Origin
- country of origin
- state or province
- precise locality
- latitude/longitude
- donor of accession

Storage information
- physical location of accession in store
- conditions of storage
- bag, envelope, box, can, etc.
- storage temperature, etc.

2. Organismic information

Biochemical
physiological
genetic information
- chromosome number
- crosses

Morphological
- plant dimensions
- plant habit
- root characteristics
- stem (or tuber) characteristics
- leaf characteristics
- flower characteristics
- fruit characteristics
- seed characteristics.

3. Pest and disease information

(Resistance to, types of organisms, etc.)

Bacterial
Fungal
Viral
Insect
Nematode
Other pest or disease

4. Environmental information

Ecological
Environmental damage
wind
water
drought
other

5. Rejuvenation or regeneration

Germinability tests
History
dates of tests
where work done
when returned to storage
etc.

6. Use information

Breeder's data
Specific application
food quality
advantageous or detrimental quality
etc.

7. Other categories, as required

SEMINAR ON DOCUMENTATION

D.J. Rogers
Crop Ecology and Genetic Resources Unit
Seminar 3.

8 March 1974

A. Before proceeding to discussion of TAXIR, need a short discussion on computing machines.

1. What they are and what they do - movie from IBM

2. The physical equipment referred to as "hardware"

3. The sets of written instructions which direct the functions of hardware - the programmes - referred to as "software"

B. Basic Concepts of hardware

input - compute - output

1. Input:

Any of several devices by which data and software are presented to the machine, such as keypunch, paper tape, or magnetic tape. Also refers to "that which is put in", data, for example.

2. Computer:

Consists of a Central Processing Unit (CPU) and associated memory units, such as discs or drums.

(a) In the CPU all instructions (programmes) are executed

(b) Size of CPU is a measure of the capacity of the machine - measured usually in two terms: size of internal memory, and speed of computation:

(i) internal memory of CPU referred to as "core" memory;

(ii) memory is measured in "bits", "bytes", and "words";

(iii) a "bit" is a single on-off signal, and one to several bits required to store a single alphabetic or numeric character;

(iv) a "byte" is a term peculiar to IBM machinery, consisting of 8 bits;

(v) a "word" is a set of bits, determined as one unit. Design of hardware determines the length of a word. IBM, for example, has standardized on a 32 bit word, consisting of four 8-bit "bytes"; UNIVAC has a 36 bit word, and Control Data Corporation (CDC) has either 48 or 64 bit words. (This latter is by far the best configuration for our efforts);

(vi) Size of CPU memory varies, depending on design and cost. A small size machine may have 8 000 words, a medium size, 16 - 32 000 words, and the largest computers may well have over 200 000 words. In the jargon of the computing field, 1 000 = "K", so a small machine will have "8K" memory, medium "32K" and large one "200K".

(c) Speed of the operations can approach the speed of light, measured in extremely small fractions of a second, the smallest fraction being the "nanosecond", or 1/1 000 000 of a second;

(d) Associated memory to CPU may be of two general types:

(i) magnetic tape - the tape stores electronic signals much the same as in a tape recorder. While very large quantities of data may be stored on "mag" tape, this is the slowest type from which to recover the data for processing in the CPU;

(ii) disc or drum memory - frequently referred to as "random access" memory, stores incredibly large quantities of data, frequently over 1 000 000 bits. Much faster recovery of data than from tapes.

3. Output:

Even computer takes several forms, usually printed, but may be another mag tape, another set of punched cards, or microfilm, or sometimes a display device like a TV tube, referred to as a cathode ray tube (CRT).

Printing machines vary in speed, increasing as the size of the machine increases, up to about 2 000 lines of 128 characters each line, per minute.

Output in the form of graphs or maps is produced on a separate piece of hardware, either a "flat-bed" plotter, or moving paper plotter.

C. Basic Concepts of software

1. The idea of a predetermined method of directions:

(a) In our mind, we have a set of instructions to add, subtract, multiply, order, and follow through some procedure to accomplish some task;

(b) The computing machine has no such "built-in" set of instructions, and must be directed to accomplish each step in the correct order to complete any task. The task can be any type, from solving equations, retrieving data from storage, or printing a computer-drawn art work;

(c) The above indicates that we must use a "stored" programme, to direct the computer;

(d) All functions in the computer are accomplished by answering a series of yes-no type questions, or, in terms of electronic signals, instead of yes-no, on-off. This means of signals to direct the computer is called "machine" language, and the programme looks like a series of random zeros and ones. This is very confusing, and boring, to human beings, and machine language programming is extremely prone to human error.

2. The concept of a "higher" language for software.

(a) Since we understand, and feel more comfortable with, instructions which look more like natural language, computer specialists have developed several programme languages which can be "interpreted" by the machine. These languages have become universally accepted by all computer manufacturers, so that any machine can use these "higher" languages. One such language, of greatest general acceptance, is called by the acronym FORTRAN, for "formula translation";

(b) FORTRAN is the language used for TAXIR, which we will describe in the next seminar.

SEMINAR ON DOCUMENTATION

D.J. Rogers

Crop Ecology and Genetic Resources Unit

Seminar 4
11.3.74

Description of TAXIR:

A. Description of the Program:

1. 2 000 FORTRAN statements (more or less) instruct the computer to perform all necessary functions required in information storage and retrieval.
2. The FORTRAN used is one that is functional on medium to large size computing machines.
3. Each FORTRAN statement causes the machine to perform some necessary activity.
4. The first such statement, for example, essentially turns on the machine - START.
5. From this first statement onward, the TAXIR program places the data bank in storage, and then makes the data available to retrieve (to ask questions).
6. Other FORTRAN statements cause the machine to produce (print) a number of useful outputs.

B. Commands or statements, in certain FORMAT, allows the user to perform functions which are required;

1. The first command made by the user tells the computer to DEFINE ITEMS from cards or tapes; with each item, one defines the descriptors (and their type) which tells the machine which set of descriptors will be included in that data bank - separate data banks can be built from one set of input data cards. In the Introduction Unit, a data bank for accessions for the herbarium, for the storage functions (1) Long term storage; (2) Distribution; (3) Working collection.
2. After the command BUILD DATA BANK, the program automatically lists (prints) all items and the full set of descriptors, and descriptor states, with a list of errors in the data. The list of errors tells the user all the mistakes made by the keypunch operator; i.e. an item with too few descriptors, or an illegally prepared descriptor state.

3. Another command, executed after all corrections are made, causes the computer to provide a CONTROL VOCABULARY (CV). The CV is a list of each descriptor, and under each descriptor, each descriptor state. The CV tells the user all the words he has used in defining each item. It tells the user whether one word has been spelled several different ways (such as country of origin may have been spelled more than one way "TERKEY", "TORKEY" and "TURKEY"; where obviously only "TURKEY" is correct).
4. With the Control Vocabulary, one may also discover that several descriptor states are synonymous, and unnecessary.
5. The next command available is "CORRECTION", to correct the mistakes made in any descriptor, or descriptor state.
6. Still other commands are "ADD MORE DESCRIPTORS/DESCRIPTOR STATES (where more data are required); "DELETE DESCRIPTOR/DELETE ITEM" (where duplication has occured, or some other mistake in the data has been made).

C. Questioning the data banks

1. After all the problems in "B" have been solved, one is ready to use the data bank in any of many different ways:

   a) Complete lists of all items in the bank, with certain descriptors;
      Example: PRINT: items with country of collection, TURKEY"
      The results will be a collection of all types of accessions made in Turkey, including everything the data bank contains from this country.
   b) Precisely defined answers to specific requirements:
      Example: PRINT: PRECISE LOCALITY, ASPECT, LATITUDE and LONGITUDE for items with GENUS SECALE and SPECIES montanum but not SPECIES CEREALE.
   c) Accession No.
      Example: PRINT: ROOM, SHELF no., BOX NUMBER for items with GENUS TRITICUM and SPECIES AESTIVUM and VARIETY ..... and LOCALITY AFGHANISTAN or IRAN.

Note use of certain words: and, or, not. These permit formulation of many types of precise questions, giving the user a limitless number of ways to derive the most benefit from his data.

Because of the freedom in formatting data, and because of the flexibility of the TAXIR program, the user is provided with a long-sought tool.

The ability to change, modify, add or delete information is part of the system.

TAXIR has a number of allied programs which can be used, for example: Statistical analysis of data
Plotting and contouring on maps
Character analysis and clustering program
(TAXIMETRIC PROGRAMS)

TAXIR is, therefore, not just a computer program, but rather a number of programs integrated into a system. Learning to use the system takes some time, but the results and benefits far outweigh the costs.

SEMINAR ON DOCUMENTATION

D.J. Rogers

CROP ECOLOGY AND GENETIC RESOURCES UNIT

Seminar 5

13.3.74

A. Review of Seminars

1. The parts of the documentation function
   a) Data-oriented, precomputer
   b) Computer-oriented
   c) Management functions required to integrate these

B. Flow of Work in a) above

1. Determine objectives of data bank(s)
2. Determine the item for each data bank
3. Determine types (classification) of descriptors needed in each data bank
4. Choose the descriptors, and the order needed to collect the data
5. Determine the descriptor types (NAME, CODED, ORDER)
6. Choose means of preparation for computer input
   a) Key punch most common
   b) Paper tape
   c) Mark sensing
   d) Mag tape

C. Flow of Work in A.1.b) above

1. Choose the software package needed (determined by objectives)

2. Choose appropriate computing facilities

   a) Closed-shop computer center operation
   b) Open-shop computer center operations

3. Determine whether appropriately trained personnel in computing center

D. For Both B. and C. above, Determine Budgetary Requirements

1. Factors in budget

   a) Is the data bank central to functions of your unit (organization)
   b) Personnel trained in data gathering
   c) Is there a line item in your budget for computing functions
      i) Includes trained programmers, software costs, and computer time
   d) If no line item, determine possibility of outside contract
   e) Is there a documentation group for your unit (organization) to take care of the whole documentation function. What are their costs?

E. Discussion of Seminar Participant Papers

1. Instructions given not clear - assignment was to:

   a) List a set of data banks of interest to the individual or unit
   b) Prepare for one data bank a list of descriptors

E. 1. c) Flow for this would have been to follow Steps in D. above

2. Most problems in design of descriptors

   a) Definition of descriptors: a single basis for comparison, defined over the full set of the data bank. Further, the descriptor is (by rule) mutually exclusive, and non-overlapping

b) Examples to illustrate a)

i) Dates

The day of the month is a single basis for comparison
The month of the year " " " " " "
The year " " " " " "

Why important? To make a precise query, for example, on flowering times, days more important than months

ii) Names of persons

In almost all languages there is a family designation (in English, the surname) and a designation of the individual (in English, the given name(s). Therefore:
Family name or surname is a single basis for comparison
Given name " " " " " "

Why important? If we are to pinpoint a single individual in a query, must have the capacity to ask for Smith, John, as different from Smith, Robert. This can happen in a long data bank with many names in it

iii) Names of plants

Genus name is a single basis for comparison
Species name is a single basis for comparison
Authority for genus name is a single basis for comparison
Authority for species name is a single basis for comparison

Why?

3. Overlap in requirements between different disciplines

a) Several data banks included requirements for
climatic data
soil data
ecological data

b) Documentation function, if performed correctly, would allow sharing of the expertise among all the various functions of FAO