Plant Germplasm Conservation: Perspectives for the 2000s. 49-60 (1994). Taiwan Agr. Res. Institute

GERMPLASM RESOURCES INFORMATION NETWORK (GRIN) — A Case Study for Designing a National Plant Germplasm Information System.

Edward Bird
National Germplasm Resources Laboratory, USDA-ARS, BARC-West, Beltsville, Maryland 20705, USA.

ABSTRACT

The Germplasm Resources Information Network - Database Management Unit has been in operation for ten years. During this time, technology for managing information on genetic resources has improved and the demands for the information have increased. Many of the lessons learned by the Unit are relevant to germplasm programs in other countries and for other organisms. The most important lessons being: to maintain good communication between the system developers and the users; to keep the system flexible and adaptable to change; and to follow international standards whenever possible. The Database Management Unit is designing the third version of the software now and the system will continue to evolve as improvements in technology make it possible to manage more information, answer more questions, and communicate with other systems around the world.

INTRODUCTION

Ten years ago the Agricultural Research Service (ARS) created the Germplasm Resources Information Network-Database Management Unit (GRIN-DBMU) to complete and manage the information system for the United States National Plant Germplasm System (NPGS). Many changes have occurred within the NPGS since then and the challenge for the GRIN-DBMU has been to develop a flexible system and anticipate the needs of scientists and curators. The history of this system and the GRIN-

DBMU is detailed elsewhere (Mowder and Stoner, 1989). The purpose of this paper is to examine the influence of technology and other key factors on the decisions in building the system.

The goal of the GRIN-DBMU has always been to create a centralized information system to support a distributed repository system. Currently, there are four Regional Plant Introduction Stations, six special crop collections, eight clonal repositories, a base storage facility and five supporting sites distributed across the United States from Geneva, New York to Hilo, Hawaii, maintaining information on a wide diversity of genetic resources.

Information is accessible from the computer seven days a week and 365 days a year except when the database is off-line for maintenance and back-up.

TECHNOLOGY: Then and Now

At the time that the first version of GRIN was designed and developed (before 1984), information technology was very different from today. Very few of the repositories had a microcomputer because they were expensive and unproven.

Microsoft DOS was at version 2 and competing with CP/M and Apple DOS for marketshare. Telecommunication equipment for the repositories was being purchased with a speed of 120 characters per second (1200 bps) over the TELENET X.25 packet-switched network or synchronous dial-up modems, barely able to keep up with a good typist. Research scientists could reach the computer through a rotary of 300 and 1200 bps dial-up modems.

Curators and support staff had limited computer knowledge, usually confined to one particular machine and operating system which was different from site to site. The National Seed Storage Laboratory used a UNIVAC at Fort Collins Computer Center, the Southern Regional Plant Introduction Station at Griffin, Georgia used an IBM mainframe, and the Plant Introduction Office used a Datapoint computer. There was no electronic communication between these machines, all information was passed from site to site in printed form. Computer storage of information was expensive and the information was typically stored as short, cryptic codes decipherable only by the staff at that site.

Database management systems (DBMS) for production systems were either network or hierarchical, such as IDMS or IMS. Relational DBMS

were in their infancy and generally limited to small or experimental systems. The ARS had recently purchased a Prime model 750 minicomputer for the GRIN system which was rated at one million instructions per second (MIPS) and had 1200 megabytes (MB) of disk space. The GRIN software was being developed in FORTRAN with function calls to the Prime DBMS, a CODASYL standard network database. Each site was given a single terminal and printer to access the database.

Today, microcomputers have replaced terminals as the interface to the system. Nearly every scientist and staff person handling information in the NPGS has a microcomputer on their desk or will have one in the near future. Most scientists are computer literate, write manuscripts with their word processor, manage data sets with a file management system (e.g. dBASE, Foxpro, Paradox) or an electronic spreadsheet program (e.g. Lotus).

Microsoft is now shipping version 6 of its operating systems and dominates the operating system market. Computer storage of information and processing power are more affordable, evidenced by the introduction of gigabyte (1,000 MB) disk drives for a microcomputer and an Intel 80486 processor rated at 16-20 MIPS.

Relational DBMSs have become the standard for large, multi-user production databases and object-oriented DBMSs are now the experimental systems. Third-generation coding languages such as FORTRAN are being replaced with fourth-generation forms packages and computer aided software engineering (CASE) tools. The GRIN-DBMU is currently developing a new version of GRIN in a relational DBMS to run under UNIX on a multi-processor minicomputer.

Telecommunication speeds and methods also have changed. Most locations have 9600 bps access over FTS2000 X.25 packet-switched service and we have just installed the first 56,000 bps access to a remote site. Dial -up access is now available at 19,200 bps using PEP technology or 9600 bps with MNP5/v. 42bis data compression. The computers at repositories are being connected togeter as local area networks (LAN) and these LANs connected into wide area networks. Currently, ten sites have installed LANs or are in the process of doing so.

The new and old GRIN central computers are also networked together and connected to the Internet which is a network of networks reaching to more than 100,000 computers around the world. A gopher server has been set up on the new computer to access the information of the National Genetics Resources Program of ARS. Eventually this

computer will have database of plant, animal, microbial germplasm maintained within the United States Department of Agriculture.

The following trends should continue for the foreseeable future:

1. Faster computers storing more information.
2. More rapid system development with more powerful tools.
3. Increasing connectivity (i.e. more networks and inter-networking) and faster communication speeds.

The first two trends will have their impact on the building and managing national information systems while the last trend well have its impact on sharing the information with the international community.

IMPACT OF CHANGING TECHNOLOGY ON GRIN

This changing technology has had several impacts on the GRIN. First, pre-GRIN databases started with simple models and data structures with the trend to increase complexity.

For example, examine how germplasm origin as a data structure has evolved in the GRIN system. In the GRIP Design Recommendation Report in 1981, only the country of origin was identified as a data structure for origin. In the first production version of the system, GRIN1, the origin state/province field was added to the origin record. This data is very important when a country is divided after a war or for other political reasons. Latitude and longitude fields were added to the accession record in GRIN2 so there is less question of the sample's origin when political boundaries change. Previously, this information was stored in narrative, and some still is, which is of little use except when manually searching the records.

Germplasm is increasingly being exchanged between gene banks and furthering the need for tracking more information about the origin and history of the material. Enhancements to GRIN2 included fields to the accession record for the history of the germplasm before its acquisition by the NPGS and who developed or collected the material.

The design of the new version of GRIN, GRIN3, calls for moving the whole concept of origin out of the accession record so we can note all of the places and people that may have been associated with an accession prior to acquisiton. This information is going to become increasingly important for determining the overlap between gene banks of the world as well as intellectual property rights and the Convention on Bio-diversity.

Another example of the increasing complexity of the data being

maintained can be found in the evaluation area. Initially, it was viewed as a limited set of descriptors, collected in a single environment and variability within the data was ignored or treated in a simplistic way. The curator could choose to store their data as a range (minimum and maximum) or as the mean of a set of observations. In the GRIN2 version, we introduced the single-observation record which allowed linking each observed value for an evaluation descriptor of an accession to a different environment and the number of descriptors for a crop did not need to be predefined. Instead, as additional descriptors were identified and evaluations conducted, the result could be added to the database without changing any software. The structure for a qualifier of the descriptor also was added. The qualifier field has been used for identifying the race of pathogen used in a particular evaluation. For example, wheat has been evaluated for Hessian fly (*Mayetiola destructor*) resistance of different biotypes or isolates (e.g. H3, H6, 'Great Plains') and it would not be desirable to treat each test on each strain as a separate descriptor. With the biotype stored in the qualifier field, it is possible to search the database for Hessian fly resistance or just those accessions resistant to the Great Plains isolate of Hessian fly.

GRIN3 will add even more information about an observation. The new database has fields for the minimum value, maximum value, standard deviation and sample size which will provide more information on the reliability of the observed value.

Future versions of the system will likely contain other kinds of data about the accession, including images of the plants, seeds, and fruiting structures. However, this has to wait for further developments in disk storage, telecommunications, and software. Disk storage needs to be less expensive and have higher transfer rates. A high quality image takes several megabytes of space and we have more than 400,000 accessions which may require several images. It now takes 3.5 minutes to send a megabyte image at 56,000 bps. Once the image is sent to a user, standards for displaying and manipulating images with common telecommunication software need to be implemented. Data compression techniques can reduce the size of the image files and shorten the transmission time but the software at the receiving and must also be compatible.

ROLE OF THE USERS

The key to the success of the GRIN system has been user involve-

ment. Continual dialogue between the DBMU and the users at the repositories during the development and enhancement of the system through telephone conversations, electronic mail and regular meetings have been used to gather input from the sites, resolve conflicting requests between sites, and educate the staff at the sites on how to use the software. Involving these people has also been important in making them aware that it is their system and their data. It is easy for the personnel to give a remote system lower priority and find work to do locally. The system has to have clear benefits to each and every user either by making them more productive or the work less tedious.

ROLE OF STANDARDS

Standards are important to genetic resources information systems for several reasons. First, to convey information from one person to another. If the Plant Introduction Office and repositories use different sets of codes for names for countries then the data must be converted to be useful. If both the Plant Introduction Office and the repositories agree to an international standard like ISO country names and alpha-2 or alpha-3 codes, not only should they be able to understand each other but the information should also be acceptable to other gene banks around the world.

The use of standards also reduces the amount of overhead that must be passed with the data. If two locations agree on a standard list of country names and codes then the lists do not have to be exchanged every time the accession data is passed between the users.

Finally, the standards help with the maintenance of the data. When a new introduction record is added, the standard list of names can be used to verify the spelling and thus reduce errors at the time of data entry.

The GRIN-DBMU continues to monitor several standards setting organizations which may impact the data and the date structures of GRIN. The following are some examples:

- Interantional Standards Organization. (ISO)-currently working on a standard list of states and provenances.
- International Organinzation for Plant Information (IOPI)-working on a standard list of scientific names for name and data structures to hold taxonomic, economic, and geograpical information.
- International Board of Plant Genetic Resources. (IBPGR)-developing crop-specific descriptors, standard systems for handling genetic

resources, and coordination of international base collections.

- American National Standards Institute (ANSI)-continuing to develop standards for computer languages, particularly C, C++, and SQL.
- National Center for Biotechnology Information (NCBI)-Promoting the use of Abstract Syntax Notation (ASN.1) for the exchange of date electronically.

The users of the system are surveyed regularly to determine what technologies have been implemented and become the de facto standards for such things as hardware including printers and modems; telecommunication and database software; and LAN configurations and operating system.

THE PAINFUL LESSONS.

During the ten years that the GRIN system has been operational the GRIN-DBMU has learned many lessons, some more painfully than others. The first lesson was to load good, accurate data and take the time to thoroughly check it for consistency. The initial loading of the GRIN database took data from a wide variety of systems operating at the repositories and support sites. There were problems both with the consistency of the data and the software used to load the data into the database. Since the data was being maintained independently, transposition errors in the accession identifiers were not detected until two sites claimed the same accession. Even worse problems were errors created by the loading programs because a large number of records were effected.

The second lesson was to take the time to test and check the software. The second version of the software, GRIN2 had over 250,000 lines of FORTRAN code and reached the users without much quality control. The users became very frustrated because of the problems encountered and the software aborting their work.

Another lesson was that once the users accept the system and rely on its use, it is extremely important to keep the system reliable and functional. Disaster recover and contingency plans need to be reviewed and updated regularly. Twice the GRIN system has been down for an extended period of time. Once for the electrical re-wiring of the computer room due to an uninformed building electrician's knowledge of what is required to ground the high frequency feedback from a computer. The second instance was due to internal inconsistencies in the database that went undetected

for several months and eventually required the unloading and reloading of the entire system, which itself took more than a month. In both cases, repositories that had been using the system for their day-to-day operations were forced to revert back to manual record keeping and then update the database when it came back on-line. The extra work and the delays caused by the manual processing of the data reduced the productivity of the staff, increased their frustration, and gave them a sense of not being in control of their operation.

SOME OTHER VALUABLE LESSONS LEARNED

Keep communication lines open with the staff at the repositories through as many means as possible. The more informed and educated the users are about the system, the better they respond to questions about what they would like changed and improved.

Try to anticipate changes in the way the users want to use the system and the information they wish to store. As users added more data to the system and used it for maintenance of their inventories, the GRIN-DBMU received requests to add additional fields and procedures to manage the information. For example, when seed is sent to NSSL for long-term storage, they wished to record the site inventory sample identifier as well as the accession identifier of the material. Other fields were added for recording when an accession was placed in quarantine, where it was held in quarantine, and when it was removed from quarantine.

Another valuable lesson that was learned is to spend time understanding how the software works and how the data is managed at a very technical level for the purpose of tuning and correcting problems. Several times in the last ten years the GRIN-DBMU has had to use a binary editor on the files in the DBMS to correct pointers, patch b-trees in indexes, and reconstruct data. On a three gigabyte database it takes a large amount of the resources of the computer to monitor and protect the integrity of the data. A program to validate all of the links for all the inventory records in the current system, about 700,000 records, takes a month to complete.

Finally, as a system grows in size, it takes people with many different talents and specialties to manage the system. The GRIN-DBMU currently employs 14 people including:

- Database manager to supervise the group.
- Secretary.

- System administrator to oversee the hardware maintenance and computer room operations.
- Database administrator to monitor the database and lead the development of new software.
- LAN administrator to manage the Laboratory LAN and act as back -up to the system administrator and database administrator.
- Three software developers for writing new software and maintaining existing code.
- Data administrator to manage the loading of data and developing rules for how the data should interrelate.
- Crop Advisory Committee facilitator to keep good communications with the users of the system, both managers of the information and their industrial counterparts who guide our efforts.
- Three other people that work with data, either entering data or managing critical parts of the system, such as the standards area or intellectual property rights, such as patents and cultivar registrations.
- Technical document specialist for producing manuals and other documentation for the system.

The GRIN system would not be where it is today, if it wasn't for the contributions by all of the individuals in the GRIN-DBMU working as a team. Members of the unit also perform other roles including software testing, technical evaluation and equipment purchase, generating reports for management, and formatting data for distribution.

As the computer industry continues to change, the GRIN-DBMU needs to develop additional expertise in other areas. Most of the emphasis and expertise is on plant germplasm and a central, multi-user database. With the 1990 Farm Bill, the role of the GRIN-DBMU was expanded to include other agriculturally important germplasm (microbial, livestock, poultry, fish, forest trees, etc). Microcomputers software continues to change and expertise is needed to update existing software (eg. pcGRIN) and write new software to take advantage of the features such as windowing environments, database front-end tools, and multi-user LAN applications. Finally, expertise is needed in the GRIN-DBMU to link our databases with those in other countries and other kinds of computers. This requires knowledge of other scientific disciplines, international telecommu-

nications and, possibly foreign languages.

Disclaimer: Brand names of computer hardware and software are necessary to report factually on available data; however, the USDA neither guarantees nor warrants the standard of the product, and the use of the name by USDA implies no approval of the product to the exclusion of others that may also be suitable.

LITERATURE CITED

Germplasm Resources Information Project. 1981. GRIP Design Recommendations. Working Report. College of Agricultural Sciences, Laboratory of Information Science in Agriculture, Colorado State University, Fort Collins, Colorado.

Mowder, J. D. and A. K. Stoner. 1989. Information systems. p. 57-65. In: J. Janick (ed.) 1989. Plant Breeding Reviews. Volume 7. The National Plant Germplasm System of the United States. Timber Press, Inc., Portland, Oregon.

Appendix I

GRIN System

Growth in the Database

1983	--	300 MB allocdted
1985	--	900 MB allocdted
1988	--	1,800 MB allocdted
1992	--	3,400 MB allocdted

Source History Example

PI 511511 *Zea mays* (field corn)

- Collected in Brazil
- Donated by CIMMYT in Mexico
- Received through Pioneer Hi-Bred

GRIN System

Resistance to Hessian Fly, Biotype' Great Plains'

Accession ID	Resistance
PI 68284	2
PI 243787	1
PI 320193	1
PI 324928	1

(select WHEAT Hessian-Fly $<$ 4 and DSQUAL = 'BIOTYPE-GP')

Appendix II Organizational Structure of the DBMU