E

TWC/23/7-BMT/9/3
ORIGINAL: English
DATE: May 23, 2005

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Third Session
Ottawa, June 13 to 16, 2005

WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA PROFILING IN PARTICULAR

Ninth Session
Washington, D.C., June 21 to 23, 2005

CREATION OF DATABASES FOR MOLECULAR MARKERS: ONE APPROACH TAKING INTO ACCOUNT POSSIBLE COOPERATION BETWEEN AUTHORITIES/LABORATORIES

Document prepared by an expert from France

Summary

1. At the eighth session of the Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular (BMT) held in Tsukuba, Japan, from September 3 to 5, 2003, documents BMT/8/13 and BMT/8/14 considered how to select markers and how to score and store results. Storing data in databases was also introduced in document BMT/8/14. This document is based on the previous documents presented in UPOV and focuses on database modeling in the perspective of cooperation between laboratories/countries. It is based on the experience of GEVES (France) with various crops and techniques in databasing and its use of molecular data, for different purposes. At the 2005 sessions of the TWC meeting, to be held from June 13 to 16 in Ottawa, Canada, and the BMT, to be held in Washington D.C., United States of America, from June 21 to 23, simple uses of such databases will be shown.

Need for a Database

2. In common language, the existence of information in a computer file (e.g. a spreadsheet) is sometimes incorrectly referred to as "the data is kept in a database". The storage of information in a spreadsheet, for instance Excel©, where lines are different varieties (or lots, or samples) and columns are information in order to identify the variety (i.e. variety code) and information obtained (i.e. allele found) is such an example. This is very convenient when the number of lines and columns is small, as it allows a lot of functions (sort, average, …) on the data and nice outputs in order to provide descriptions or reports.

3. When the amount of information is large it becomes cumbersome, difficult and risky to enter, update and retrieve information in cells of spreadsheets. Only a database can allow the entering of the same information in different places (making updates less of a problem), or the entering of different types of information in a cell (making retrieval of information less problematic). When it is necessary to provide simultaneous access with different rights to many users, a database is needed. Access to the data is then controlled by software which allows different levels of access to the data (right to create, read, update, delete).

4. When a database is required, the data needs to be analyzed in order to provide a database model. A database model can comprise only a few tables or may involve more than one hundred tables. A table is the equivalent of a row x column spreadsheet, where rows are occurrences of the different objects, and the columns the different types of information needed to identify and describe each object. Database design and implementation is usually performed by IT (computer) experts in cooperation with the users of the information. Database design is not an easy task, and usually a design has a cost and an economical value (this is also true for data themselves).

5. Administrative information and technical descriptions of varieties already known, and of those under study, are commonly stored in the databases held by offices in charge of DUS testing and the granting of protection. The notes which are used to form the description, including electrophoresis when applicable, are often but not always in a database. The results from molecular studies are less commonly included in databases, and are often contained in special files, as they are not routinely use in the process of DUS studies.

Types of Exchange Using Databases

6. International organizations such as UPOV (UPOV-ROM Plant Variety Database) or the European Union, for instance, have developed databases in which information on known varieties and varieties under examination are available.

Exchanging information through files with an agreed format

7. In 1996, UPOV elaborated a format (Circular U 2462 09/11/1996) for providing variety information for inclusion in the UPOV-ROM Plant Variety Database. SGML (Standard Generalized Markup Language) has been used where recognition of the information is made by the use of tags. For instance, for the UPOV-ROM the tag <600> indicates the following information is the breeder's reference.

8. Many software packages provide assistance to export or input in XML (Extensible Markup Language). XML is able to manage not only identification of data fields, but also aspects of database design (i.e. description of tables). XML is used by some governments or organizations for exchange of information.

9. HTML (Hyper Text Markup Language) is used to describe the display of information on the computer screens through the internet, but it is not a way to exchange data.

Exchanging information through database agreed format

10. Rather than sending files to a place where someone is in charge of putting together all data, it is possible to update a common database online, or to define a common database model for the purpose of information exchange.

11. This approach is probably better if the aim is to reach harmonization and encourage cooperative work.

12. An example is a project with France, Germany, Spain and CPVO where information on *Zea mays* varieties is shared by the official crop experts of the three countries. A database model has been designed to fulfill the aims of the project, and each partner keeps in its own premises a full database, which is updated three times a year.

Steps to Define a Database

13. When information needs to be exchanged and/or grouped together through database format, it is important to:

- define the aims and general organization for the exchange of data, including ownership of data and confidentiality
- list participants and potential users
- list the expected uses of the data
- identify the logical objects to be used (a variety, a lot, a breeder….)
- select the type of information which is to be exchanged for each logical object, (the name of the variety, the reference of a lot, the address of a breeder…)
- define precisely each piece of information for the purpose of the exchange, as each partner may have different internal definitions. For instance, date of creation of a new

variety can be the date of entry in the database, the stamp date of the mail received, the date of signature of the official request, etc. Even with the same definition the coding might be different (species might be a national name, a Latin name, a numerical code, an alphabetic code…)

- define the database relationship between the logical objects. For this, the input of IT (computer) persons is essential as there are rules to follow in order to design an efficient and reliable database.
- estimate the volume of data and frequency of updates
- list the types of outputs, queries, sorting, and the type of treatment (computations) that are expected to be performed
- either define a common coding for the logical objects used (species, marker, …), which is the best solution; or define correspondences between the different codes used in countries, laboratories, organizations… this is more difficult to maintain.
- define how, and by whom, data will be regularly provided, checked, updated, and made available.

14. Usually, when there is a common agreed coding, a correspondence between the agreed coding and the coding used internally is kept in each laboratory or country database. This allows the use of specific internal coding which is often necessary to relate to previous data, to relate to legal or administrative coding, and to allow the identification of objects which are not yet internationally agreed, or which will not be internationally agreed but are useful in the laboratory or country.

15. NB: document BMT/8/13 give information on how to select a molecular marker system, the material to study, the standardization of protocols, intellectual property issues.

Logical Objects in Database Model

16. At least six logical objects are needed to define a logical model with the aim of storing molecular data obtained on varieties using different marker systems.

"store molecular data obtained on varieties using different marker systems"

17. Different marker systems refer, for example, to Micro-satellites (SSR and variants), Sequence data (SNP), etc.

18. In order to store data obtained via different techniques, we have to identify the techniques available and the techniques and/or types of markers used.

Technique/Marker

"store molecular data obtained on varieties using different marker systems"

19. Of course we need to keep the data themselves. Data represents presence/absence, or frequency of presence.

Data

20. Naming and coding of the results is variable (bands, migration distance, weight, …) depending on the technique, the laboratory, the software of the apparatus ...

21. We need to overcome this variability if we want to avoid having many database models. One possibility is to define the different results with Locus and Allele coding.

Locus

Allele

"store molecular data obtained on varieties using different marker systems"

22. Very often the object under study can be related to a variety. Nevertheless what is actually looked at is a sample from a given identified lot of a variety. In the model, we need to be able to trace results obtained on samples. From the results, an official or stabilized description is usually produced and can be stored with the same logical object.

Variety

23. Information on one or many species will be produced and kept; both plants and techniques are to be related to a given species (or sub-species, or group of species)

Species

24. To "store molecular data obtained on varieties using different marker systems", this information is probably a minimum core basis of any project for exchange on molecular data.

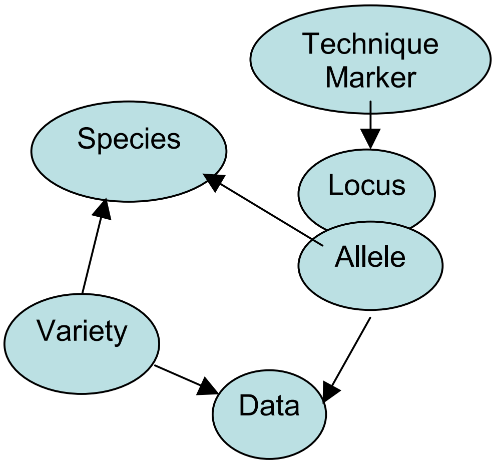


Figure 1: The 6 core logical objects and the links between them.

Data Dictionary

25. In a database, each of the objects becomes a table in which fields are defined. Some fields are needed to identify unambiguously each line (tuple), some fields are used to store the information, and other fields can be used for traceability.

26. For each of these tables an example of information field is given below:

Technique/Marker code: indicates the code or name of the technique or type of marker used
Examples: SSR, SNP, …

Locus code: indicates name or code of the locus for the species concerned
Examples: gwm 149, A2, …

Allele code: indicates name or code of the locus for the species concerned
Examples: 1, 123,…

Data value: indicates a data value for a given sample on a given locus-allele
Examples: 0 (absence), 1(presence), 0.25 (frequency)

Sample number: indicates the number given to the sample during the process analysis
Examples: 123, AP2124, …

27. In each table, the number of fields, their name and definition, the possible values, the rules to be followed, … have to be described. They are usually kept in a document called the "data dictionary".

Table Relationship

28. The links between the tables is also important for database design, a simple example is first given, followed by a proposal concerning molecular data.

Table	Link	Table	Description
Woman	0 or 1 to n	Child	0: A woman may have no child 1 to N: a woman may have 1 to n children (she is then a mother)
Child	1 to 1	Woman	A given child has only one biological mother

Table 1: Daily life example to illustrate the reading of the links between objects

Suggested relationship between the minimum core objects:

Table	Link	Table	description
Technique/marker	0 or 1 to n	Locus	0: A technique/marker can be present in Technique/marker, even if no locus/allele is yet used in the database 1 to n: a given type of marker can provide 1 to n useful loci
Locus	1 to 1	Technique/marker	A given locus is defined within the scope of a given technique/marker
Locus	1 to N	Allele	For each Locus 1, or more than 1, allele can be described
Allele	1 to 1	Locus	A given Allele is defined within the scope of a given Locus
Allele	0 or 1 to n	Data	0: a given Allele can be defined, but without data 1 to n: a given allele can be found in 1 to n data
Data	1 to 1	Allele	data corresponds to a given allele
Variety	0 or 1 to N	Data	0: the variety has no data 1 to N: the variety has data
Data	1 to 1	Variety	data corresponds to a given variety
Data	1 to 1	Species	data is obtained for a given variety, then for the species of the variety.
Species	0 or 1 to N	Data	0: a species can have no data. 1 to N: a species can have 1 to N data.

Table 2: database links between the 6 core objects

Physical Database Model

29. Given the objects, the information fields, the relationship between objects and adding other IT technical aspects that are not described here, a physical model can be established. This is commonly referred as “the database model”.

30. Below is a partial example of such a physical model, where the 6 core objects (technique/marker, locus, allele, species, variety, data) can be retrieved. Please note that in this physical model the object sample (“échantillon”) appears. The link from sample to official variety description, lot from a variety and its status, initial sample in case of result on a sub-sample, plant number from a sample, part of plant from a plant, etc cannot be seen in this screen shot. More explanations on the need to clarify the link to the variety are given in the paragraphs following the screen shot.

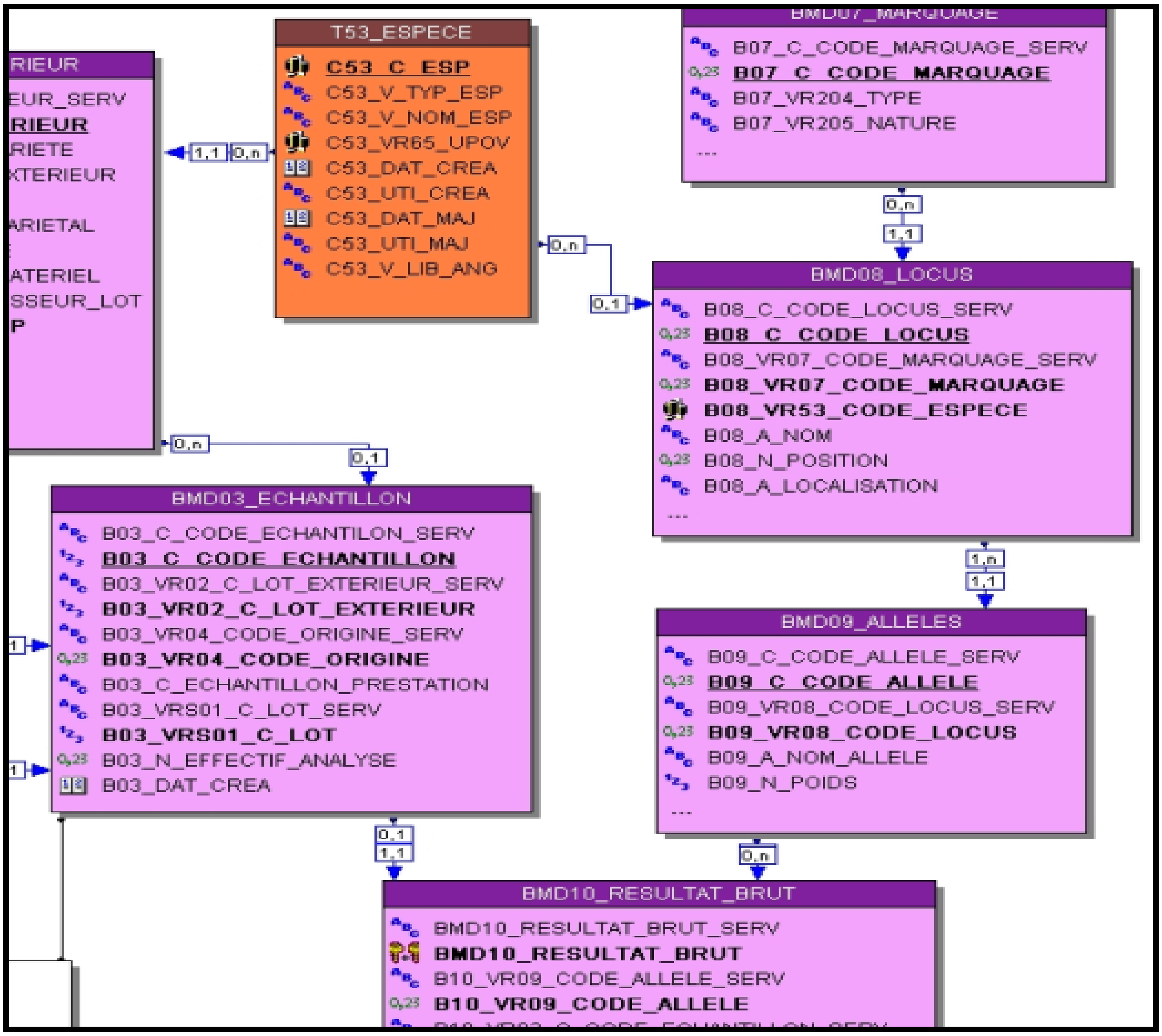


Figure 2: Partial screen copy of a physical database model where the 6 core objects are present. Special database tools can handle the different models (logical, physical, there are others as well which not described in this paper. Specialized database software can check for consistency between all models. Word© or Excel© can also be used to provide equivalent outputs, but consistency relies exclusively on the person who creates the documents.

31. When different options are possible in the logical model, one option must be defined to implement the database.

32. Example: Data can be obtained on a sub-sample (one or many plants) of a given lot in a given test, but it can also be the result that is kept as the official description of a variety after a number of tests.

33. In order to identify the type of data, three options (among many) are listed below:

1. The database is a cooperative database in which only the unique official description accepted by the country is stored.

2. The database can contain different types of data (raw, preliminary, checked, averaged, official, …).

3. The database can contain any type of data obtained in a given test, on any type of plant material.

34. In option 1, by definition, the data description is unique for a given variety, a direct link from data to variety can be chosen.

35. In option 2, each data can have a field "status of data" in which the different possibilities will be stored. In that case each data has one and only one status. If more than one status is needed, different descriptions (one per status) can be kept.

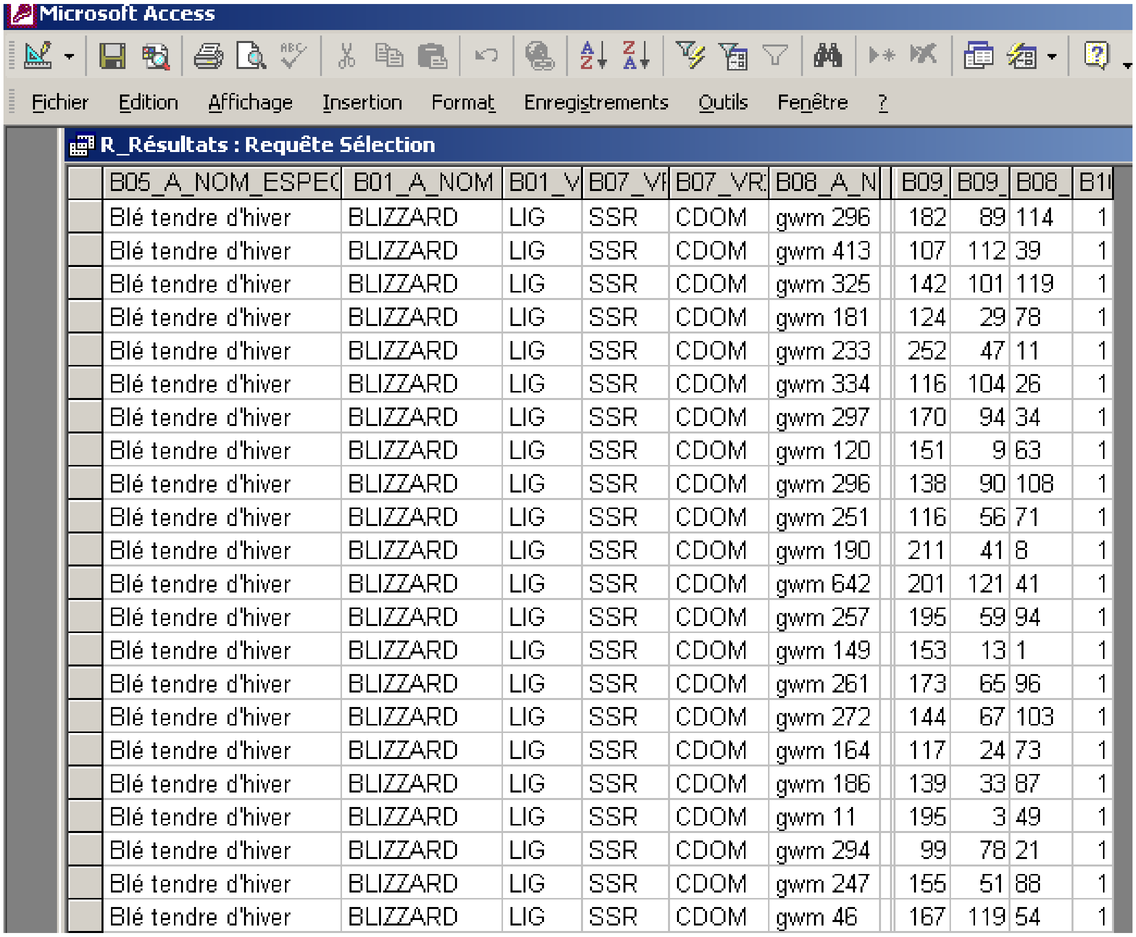
36. In option 3, each data is linked with a unique material identifier which refers to the material (cultivar, lot, sample, sub sample…) and with a unique test identifier which refers to the test (date, place, responsible, protocol, …).

37. A number of IT features (keys, indexes, naming conventions, type of data storage…) are also to be defined in addition to logical features in order to create a database in a computer. These IT features are not described in the present document.

Storing and Retrieving Results

38. When data are stored in a database, there are many ways to select/retrieve them, and to output or to compute. Access can be controlled in order that only designated persons can input data, the same or others can validate the data, whilst others only have access to read data, …

39. Some information can also be masked to some users, so that, for example, a crop expert may only have access only to their crops of interest, non validated results are only known to the laboratory, etc…



B05_A_NOM_ESPE	B01_A_NOM	B01_V	B07_VI	B07_VR	B08_A_N	B09_	B09_	B08_	B1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 296	182	89	114	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 413	107	112	39	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 325	142	101	119	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 181	124	29	78	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 233	252	47	11	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 334	116	104	26	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 297	170	94	34	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 120	151	9	63	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 296	138	90	108	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 251	116	56	71	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 190	211	41	8	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 642	201	121	41	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 257	195	59	94	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 149	153	13	1	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 261	173	65	96	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 272	144	67	103	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 164	117	24	73	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 186	139	33	87	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 11	195	3	49	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 294	99	78	21	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 247	155	51	88	1
Blé tendre d'hiver	BLIZZARD	LIG	SSR	CDOM	gwm 46	167	119	54	1

Figure 3: Example of "instantaneously made" crop expert request to retrieve SSR markers on Blizzard trough Access© as query tool, via ODBC from Oracle© or other database systems. More user-friendly and powerful tools such as Business Objects© for instance can also be used.

Cooperative Databases

40. The creation and maintenance of databases requires the necessary software and IT support. Defining a database to fulfill its intended purpose needs time and explanations, for both providers of data and users of the database. Some countries have already implemented databases containing molecular data. When molecular data are in a database; if efficient links are provided with administrative data, DUS test data, agronomic evaluation data, when applicable, post-registration variety control data, … then it is a very easy and powerful tool.

41. Usually cooperative databases do not contain all data available in each country, but only a subset. Agreed codification of objects is essential to retrieve information from other countries and to identify common data (when information is available in more than one country). Rules or recommendations on how to use, or not to use, the data provided is often the more difficult aspect of the work.

What Kind of Database Software do we Need?

42. In practice nobody is free to choose a given database software which will be acceptable by any partner/user. Most institutions/organizations select one (or a few) database software(s)

depending on technical needs and software abilities, budget constraints or cost, the recommendation from authorities or co-operators, etc. For a large amount of data there are several database systems like ‘Oracle©’ in France or ‘IBM-Informix©’ in Germany. For a small data set and for data exchange ‘MS-Access©’ might be the database system of choice. It is also common to work with different database systems in parallel when the institution/organization is not too small. For exchanging data the type of database system and the version number are also important in order to harmonize the format of the data. Therefore, no general recommendation can be made on which software (or operating system) to use; that is a matter for partners to discuss and to define in their co-operative project.

Examples of Data That Can be Stored in Such a Database

43. In order to make a link with the results as they are obtained in the laboratory, three illustrations have been included in this document. They are only examples, as such a model can be applied to Agricultural, Vegetable, Ornamentals, Fruit crops … . It can also be applied for autonomous or allogamous crops, and to different types of varieties.

AFLP - Rosa

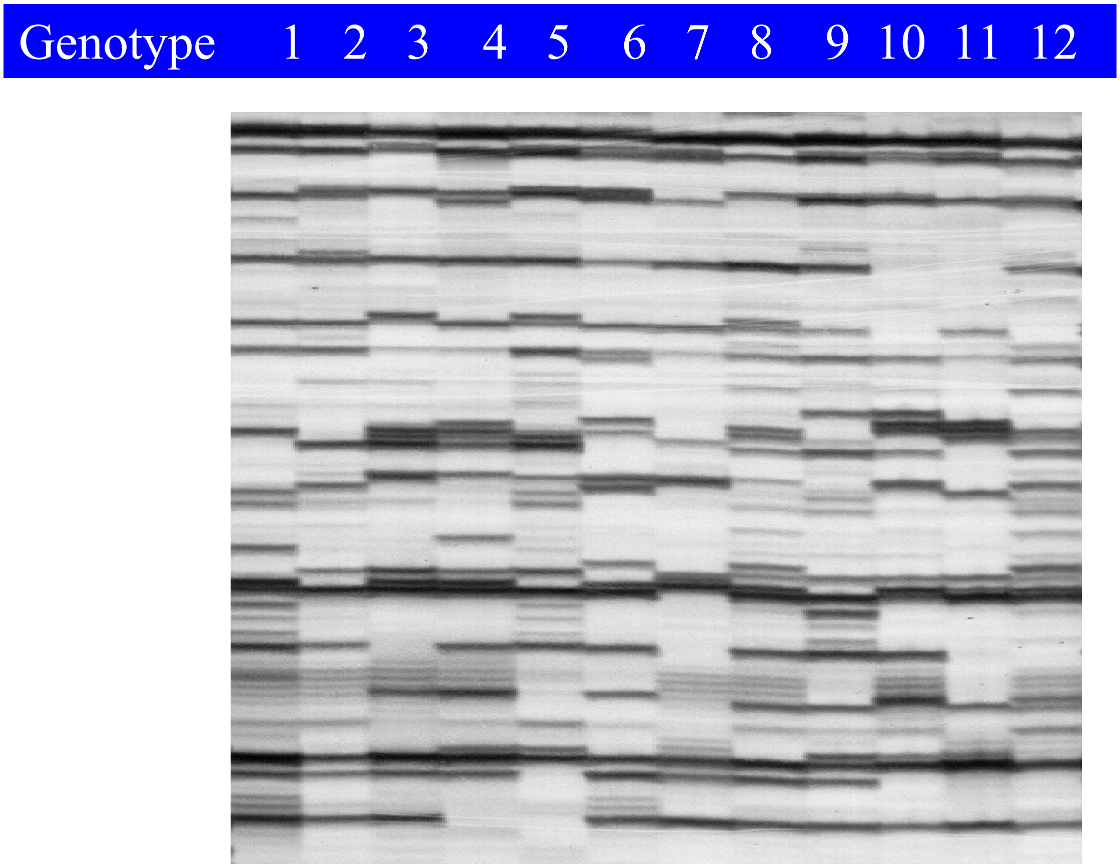


Figure 4: AFLP data from Rosa sp.

SSR - Zea mays

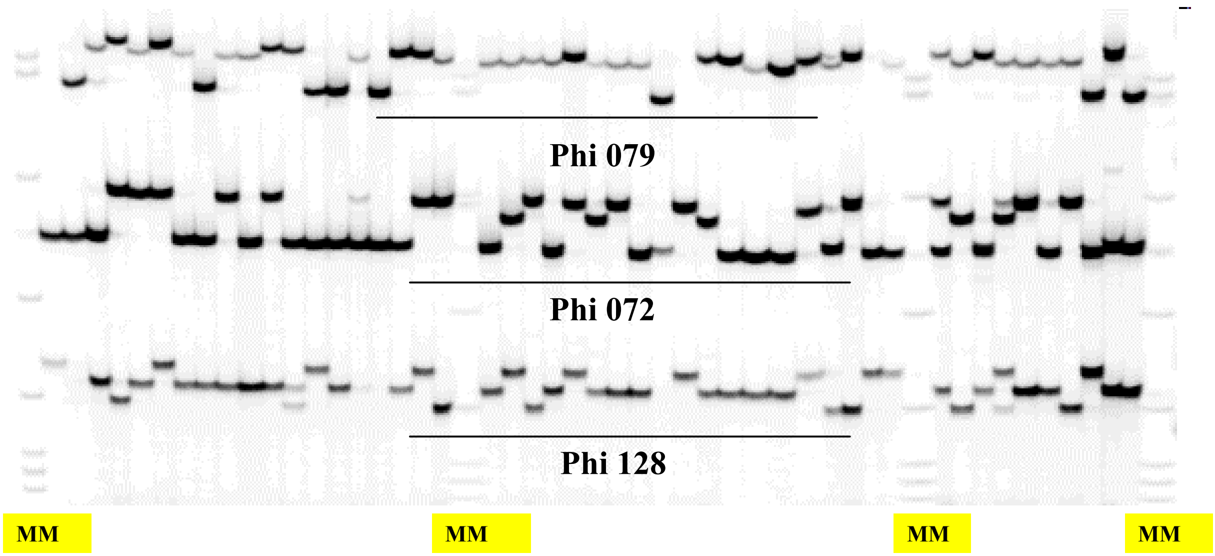
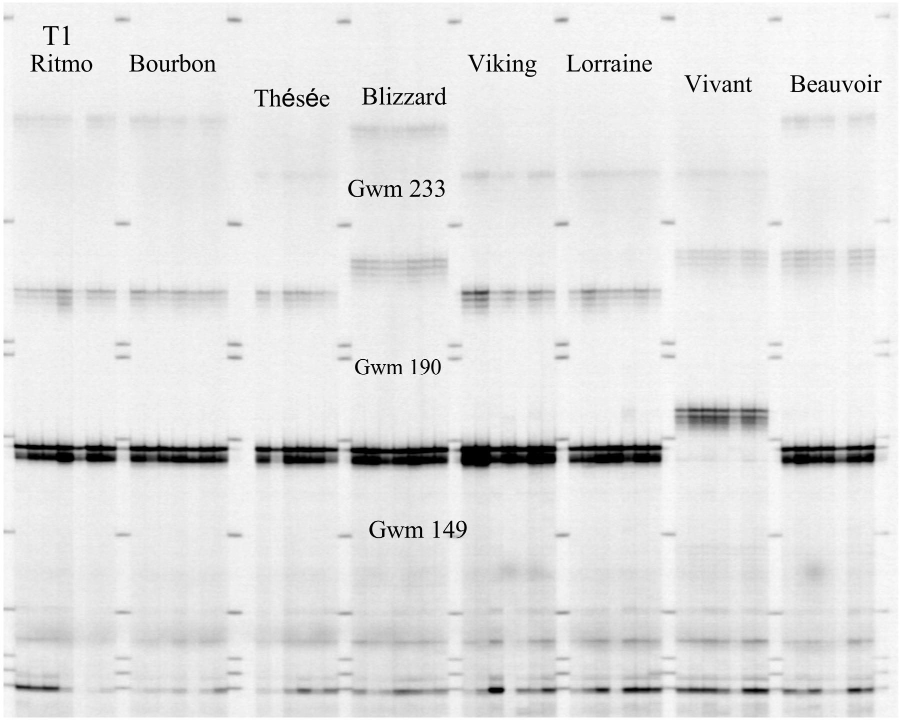


Figure 5: SSR data from Zea mays.

SSR - Wheat



Migration Triplexe T1 : gwm149 + gwm190 + gwm233 8 varieties Blé tendre (7 seeds/variety)

Figure 6: SSR data from Triticum aestivum

[End of document]