UPOV
International Union for the Protection of New Varieties of Plants

E

Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular

Sixteenth Session
La Rochelle, France, November 7 to 10, 2017

BMT/16/5

Original: English
Date: October 25, 2017

STANDARDS FOR DATABASES CONTAINING MOLECULAR INFORMATION

Document prepared by the Office of the Union

Disclaimer: this document does not represent UPOV policies or guidance

1. The purpose of this document is to explore the possibility to use WIPO standard ST.26 for databases of molecular information.

2. The structure of this document is as follows:

INTRODUCTION 1
TYPES OF MOLECULAR DATABASES 2
WIPO ST.26 3
POSSIBLE USE OF WIPO ST.26 FOR DATABASES OF MOLECULAR INFORMATION 3
16.1. Feature Key gene 5
16.2. Feature Key source 5
16.3. Feature Key STS 6
16.4. Feature Key variation 6
17.1. Qualifier allele 7
17.2. Qualifier chromosome 7
17.3. Qualifier compare 7
17.4. Qualifier cultivar 7
17.5. Qualifier ecotype 8
17.6. Qualifier PCR_primers 8
17.7. Qualifier phenotype 8
17.8. Qualifier variety 8
17.9. Qualifier sub_species 9

INTRODUCTION

3. Document UPOV/INF/17/1: GUIDELINES FOR DNA-PROFILING: MOLECULAR MARKER SELECTION AND DATABASE CONSTRUCTION (“BMT GUIDELINES”) states as follows:

> 1.2 As improvements in technology and new equipment become available, it is important for the continued sustainability of databases that the interpretation of the data produced is independent of the equipment used to produce them. This is, for example, the case with DNA sequencing data. Initially, radioactively labeled primers and sequencing gels were used to produce such data, whereas this can now be done using fluorescent dyes followed by separation on high throughput, largely automated, capillary gel electrophoresis systems.
>
> 1.3 Despite these differences, the data produced with the various techniques are consistent with each other and independent of the techniques used to produce them. This can also apply to data produced using, e.g. DNA microsatellites (simple sequence repeats, SSR) or Single Nucleotide Polymorphisms (SNPs). This repeatability and reproducibility is important in the construction, operation

and longevity of databases and is very important in generating a centrally maintained database, populated with verified data from a range of sources.

1.4 The molecular techniques readily applicable for variety profiling are constrained by the requirement for the data to be repeatable, reproducible and consistent. Thus, while various multi-locus DNA profiling techniques have been successfully used for research, codominance cannot easily be recorded in many of them, and the reproducibility of complex banding patterns between laboratories using different equipment can be problematic."

4. Provided that the same methodologies are used (e.g. RAPD, AFLP, SSR, SNP) molecular marker data can be exchanged for a specific crop.

5. In order to facilitate data exchange, it is necessary to use databases. Databasing molecular data of varieties can be used to establish a collection of varieties of common knowledge for each species.

TYPES OF MOLECULAR DATABASES

6. Molecular databases are of two types (table 1):

- Primary database: used to archive experimentally-derived data submitted directly from researchers. This data is very large in terms of size. In the UPOV context, this type of database would contain genotypes of varieties of common knowledge.
- Secondary database: contains the results of analysis often of data in primary databases. In the UPOV context, it would allow selection of the set of varieties of common knowledge to be included in the growing trial.

	Primary database	**Secondary database**
Synonyms	Archival database	Curated database; knowledgebase
Source of data	Direct submission of experimentally-derived data from researchers	Results of analysis, literature research and interpretation, often of data in primary databases
Examples	• GenBank/EMBL/ DDBJ (nucleotide sequence) • Protein Data Bank (PDB, coordinates of three-dimensional macromolecular structures) • Medline (literature) • IMEx databases (protein interactions) • ArrayExpress Archive and GEO (functional genomics data)	• InterPro(protein families, motifs and domains) • UniProt Knowledgebase - SwissProt (sequence and functional information on proteins) • Ensembl (variation, function, regulation and more layered onto whole genome sequences)

Table 1. Primary database versus Secondary database

WIPO ST.26

7. WIPO ST.26 (http://www.wipo.int/export/sites/www/standards/en/pdf/03-26-01-rev.pdf) is the recommended standard for the presentation of nucleotide and amino acid sequence listings using XML. It defines the sequence disclosures in a patent application required to be included in a sequence listing.

8. Intellectual property offices should accept any sequence listing compliant with this standard filed as part of a patent application or in relation to a patent application. PCT (Patent Cooperation Treaty) will use ST.26 as its Admin. Instructions Annex C.

9. The WIPO ST.26 XML structure is composed of:

- General information part:
  - ApplicationIdentification : Mandatory
    - `IPOfficeCode`
    - `ApplicationNumberText`
    - `FilingDate`
  - `ApplicantFileReference`: Optional
  - `EarliestPriorityApplicationIdentification` : Mandatory if Priority is claimed
  - `ApplicantName` : Mandatory
  - `ApplicantNameLatin` : Optional
  - `InventorName`: Optional
  - `InventorNameLatin`: Optional
  - `InventionTitle`: Mandatory in the language of filing
  - `SequenceTotalQuantity`: Mandatory

- Sequence data part: this is composed of one or more `SequenceData` elements. Each `SequenceData` has a mandatory attribute `sequenceIDNumber`.

Element	Description	Mandatory/Not Included	
		Sequences	Intentionally Skipped Sequences
`INSDSeq_length`	Length of the sequence	Mandatory	Mandatory with no value
`INSDSeq_moltype`	Molecule type	Mandatory	Mandatory with no value
`INSDSeq_division`	Indication that a sequence is related to a patent application	Mandatory with the value "PAT"	Mandatory with no value
`INSDSeq_feature-table`	List of annotations of the sequence	Mandatory	Must NOT be included
`INSDSeq_sequence`	Sequence	Mandatory	Mandatory with the value "000"

10. The structure of `INSDSeq_Sequence` is stipulated in the International Nucleotide Sequence Database Collaboration (INSDC). Only INSDC 49 feature keys and 80 qualifiers for nucleic acid sequences are retained because they are relevant for patent data.

POSSIBLE USE OF WIPO ST.26 FOR DATABASES OF MOLECULAR INFORMATION

11. In order to exchange data between different laboratories, in addition to using a common method and a common set of markers, it is important to agree on the common format to store and retrieve data. As XML is technology independent and is used widely to facilitate data search, retrieval and exchange, XML standards are proposed as a basis.

12. WIPO ST.26 is an XML standard used to describe sequence listing data in patent documents. The following XML code is an extract of the XML sample provided in ST.26 annex iii. It describes a DNA profile of a tomato.

<SequenceData sequenceIDNumber="5">
<INSDSeq>
<INSDSeq_length>133</INSDSeq_length>
<INSDSeq_moltype>DNA</INSDSeq_moltype>
<INSDSeq_division>PAT</INSDSeq_division>
<INSDSeq_feature-table>
<INSDFeature>
<INSDFeature_key>source</INSDFeature_key>
<INSDFeature_location>1..133</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>organism</INSDQualifier_name>
<INSDQualifier_value>Solanum lycopersicum</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>mol_type</INSDQualifier_name>
<INSDQualifier_value>genomic DNA</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>common name:
tomato</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>modified_base</INSDFeature_key>
<INSDFeature_location>15</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>mod_base</INSDQualifier_name>
<INSDQualifier_value>i</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>modified_base</INSDFeature_key>
<INSDFeature_location>22</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>mod_base</INSDQualifier_name>
<INSDQualifier_value>OTHER</INSDQualifier_value>
</INSDQualifier>
<INSDQualifier>
<INSDQualifier_name>note</INSDQualifier_name>
<INSDQualifier_value>xanthine</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
<INSDFeature>
<INSDFeature_key>variation</INSDFeature_key>
<INSDFeature_location>60</INSDFeature_location>
<INSDFeature_quals>
<INSDQualifier>
<INSDQualifier_name>replace</INSDQualifier_name>
<INSDQualifier_value>c</INSDQualifier_value>
</INSDQualifier>
</INSDFeature_quals>
</INSDFeature>
</INSDSeq_feature-table>
<INSDSeq_sequence>
atgaaattaaaacanaaaaggnatgataaaatgagatttgatataaaaaaggttttagagttagcagagaaggattttgaga
cggcatggagagagacaagggcattaataaaggataaacatattgacaata</INSDSeq_sequence>
</SequenceData>

```
        </INSDSeq>
    </SequenceData>
```

13. In the above example, the following features are used:

Feature key: source	identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence
Feature key: modified_base	the indicated nucleotide is a modified nucleotide and should be substituted for by the indicated molecule (given in the mod_base qualifier value)
Feature key : variation	a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)

14. Based on the specific constraints described above, some adaptations would need to be made to WIPO ST.26 in order to fit plant DNA-profiling needs.

15. Regarding the INSDC sequence part, as a first stage, it would be necessary to check if all retained 49 feature keys and 80 qualifiers would be relevant for the UPOV context.

16. The following four features are likely candidates:

16.1. Feature Key	gene
Definition	region of biological interest identified as a gene and for which a name has been assigned
Optional qualifiers	allele function gene gene_synonym map note operon product pseudo pseudogene phenotype standard_name trans_splicing
Comment	the gene feature describes the interval of DNA that corresponds to a genetic trait or phenotype; the feature is, by definition, not strictly bound to its positions at the ends; it is meant to represent a region where the gene is located.

16.2. Feature Key	source
Definition	identifies the source of the sequence; this key is mandatory; every sequence will have a single source key spanning the entire sequence
Mandatory qualifiers	organism mol_type
Optional qualifiers	cell_line cell_type chromosome clone clone_lib

	collected_by collection_date cultivar dev_stage ecotype environmental_sample germline haplogroup haplotype host identified_by isolate isolation_source lab_host lat_lon macronuclear map mating_type note organelle PCR_primers plasmid pop_variant proviral rearranged segment serotype serovar sex strain sub_clone sub_species sub_strain tissue_lib tissue_type variety
Molecule scope	any

16.3. Feature Key	STS
Definition	sequence tagged site; short, single-copy DNA sequence that characterizes a mapping landmark on the genome and can be detected by PCR; a region of the genome can be mapped by determining the order of a series of STSs
Optional qualifiers	allele gene gene_synonym map note standard_name
Molecule scope	DNA
Comment	STS location to include primer(s) in primer_bind key or primers

16.4. Feature Key	variation
Definition	a related strain contains stable mutations from the same gene (e.g., RFLPs, polymorphisms, etc.) which differ from the presented sequence at this location (and possibly others)
Optional qualifiers	allele compare frequency gene gene_synonym map

		note phenotype product replace standard_name
	Comment	used to describe alleles, RFLP's, and other naturally occurring mutations and polymorphisms; use the replace qualifier to annotate a deletion, insertion, or substitution; variability arising as a result of genetic manipulation (e.g. site directed mutagenesis) must be described with the misc_difference feature

17. The following 9 qualifiers are likely candidates:

17.1.	Qualifier	allele
	Definition	name of the allele for the given gene
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>adh1-1</INSDQualifier_value>
	Comment	all gene-related features (exon, CDS etc) for a given gene should share the same allele qualifier value; the allele qualifier value must, by definition, be different from the gene qualifier value; when used with the variation feature key, the allele qualifier value should be that of the variant.

17.2.	Qualifier	chromosome
	Definition	chromosome (e.g. Chromosome number) from which the sequence was obtained
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>1</INSDQualifier_value> <INSDQualifier_value>X</INSDQualifier_value>

17.3.	Qualifier	compare
	Definition	Reference details of an existing public INSD entry to which a comparison is made
	Value format	[accession-number.sequence-version]
	Example	<INSDQualifier_value>AJ634337.1</INSDQualifier_value>
	Comment	This qualifier may be used on the following features: misc_difference, unsure, and variation. Multiple compare qualifiers with different contents are allowed within a single feature. This qualifier is not intended for large-scale annotation of variations, such as SNPs.

17.4.	Qualifier	cultivar
	Definition	cultivar (cultivated variety) of plant from which sequence was obtained
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>Nipponbare</INSDQualifier_value> <INSDQualifier_value>Tenuifolius</INSDQualifier_value> <INSDQualifier_value>Candy Cane</INSDQualifier_value> <INSDQualifier_value>IR36</INSDQualifier_value>
	Comment	'cultivar' is applied solely to products of artificial selection; use the variety qualifier for natural, named plant and fungal varieties.

17.5.	Qualifier	ecotype
	Definition	a population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat
	Value Format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>Columbia</INSDQualifier_value>
	Comment	an example of such a population is one that has adapted hairier than normal leaves as a response to an especially sunny habitat. 'Ecotype' is often applied to standard genetic stocks of Arabidopsis thaliana, but it can be applied to any sessile organism.

17.6.	Qualifier	PCR_primers
	Definition	PCR primers that were used to amplify the sequence. A single PCR_primers qualifier should contain all the primers used for a single PCR reaction. If multiple forward or reverse primers are present in a single PCR reaction, multiple sets of fwd_name/fwd_seq or rev_name/rev_seq values will be present
	Value format	[fwd_name: XXX1,]fwd_seq: xxxxx1,[fwd_name: XXX2,]fwd_seq: xxxxx2, [rev_name: YYY1,]rev_seq: yyyyy1,[rev_name: YYY2,]rev_seq: yyyyy2
	Example	<INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatttttggtcayccwgaagt,rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value> <INSDQualifier_value>fwd_name: hoge1, fwd_seq: cgkgtgtatcttact, rev_name: hoge2, rev_seq: cg<i>gtgtatcttact</INSDQualifier_value> <INSDQualifier_value>fwd_name: CO1P1, fwd_seq: ttgatttttggtcayccwgaagt, fwd_name: CO1P2, fwd_seq: gatacacaggtcayccwgaagt, rev_name: CO1R4, rev_seq: ccwvytardcctarraartgttg</INSDQualifier_value>
	Comment	fwd_seq and rev_seq are both mandatory; fwd_name and rev_name are both optional. Both sequences must be presented in 5'>3' order. The sequences must be given in the symbols from Section 1 of this Annex, except for the modified bases, which must be enclosed within angle brackets < >. In XML, the angle brackets < and > must be substituted with < and > since they are reserved characters in XML.

17.7.	Qualifier	phenotype
	Definition	phenotype conferred by the feature, where phenotype is defined as a physical, biochemical or behavioural characteristic or set of characteristics
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>erythromycin resistance</INSDQualifier_value>

17.8.	Qualifier	variety
	Definition	variety (= varietas, a formal Linnaean rank) of organism from which sequence was derived.
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>insularis</INSDQualifier_value>
	Comment	use the cultivar qualifier for cultivated plant varieties, i.e., products of artificial selection; varieties other than plant and fungal variatas should be annotated via a note qualifier, e.g. with the value <INSDQualifier_value>breed:Cukorova</INSDQualifier_value>

17.9.	Qualifier	sub_species
	Definition	name of sub-species of organism from which sequence was obtained
	Value format	free text (NOTE: this value may require translation for National/Regional procedures)
	Example	<INSDQualifier_value>lactis</INSDQualifier_value>

18. WIPO ST.26 can be used to provide a description of a specific variety in the form of Sequence data.

[End of document]