|  |  |
| --- | --- |
|  | E |
| International Union for the Protection of New Varieties of Plants |  |

|  |  |
| --- | --- |
| 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

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

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

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.”

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

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

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

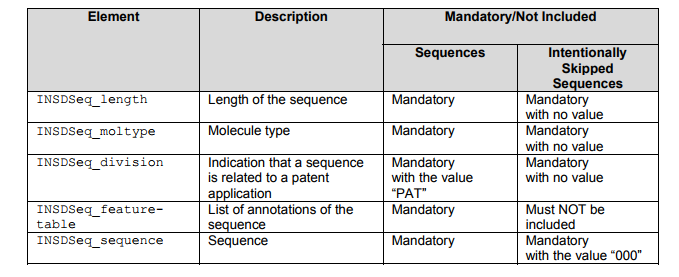
# WIPO ST.26

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.

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.

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.



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

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.

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>

</INSDSeq>

</SequenceData>

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) |

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.

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.

The following four features are likely candidates:

12. 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.

* 1. 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

* 1. 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

* 1. 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

The following 9 qualifiers are likely candidates:

12. 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.

* 1. 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>

* 1. 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.

* 1. 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.

* 1. 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.

* 1. 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: ttgattttttggtcayccwgaagt,rev\_name: CO1R4, rev\_seq: ccwvytardcctarraartgttg</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: hoge1, fwd\_seq: cgkgtgtatcttact, rev\_name: hoge2, rev\_seq: cg&lt;i&gt;gtgtatcttact</INSDQualifier\_value>

<INSDQualifier\_value>fwd\_name: CO1P1, fwd\_seq: ttgattttttggtcayccwgaagt, 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 &lt; and &gt; since they are reserved characters in XML.

* 1. 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>

* 1. 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>

* 1. 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>

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

[End of document]