UPOV

International Union for the Protection of New Varieties of Plants

E

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

BMT/16/5 Add.

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

Original: English
Date: November 3, 2017

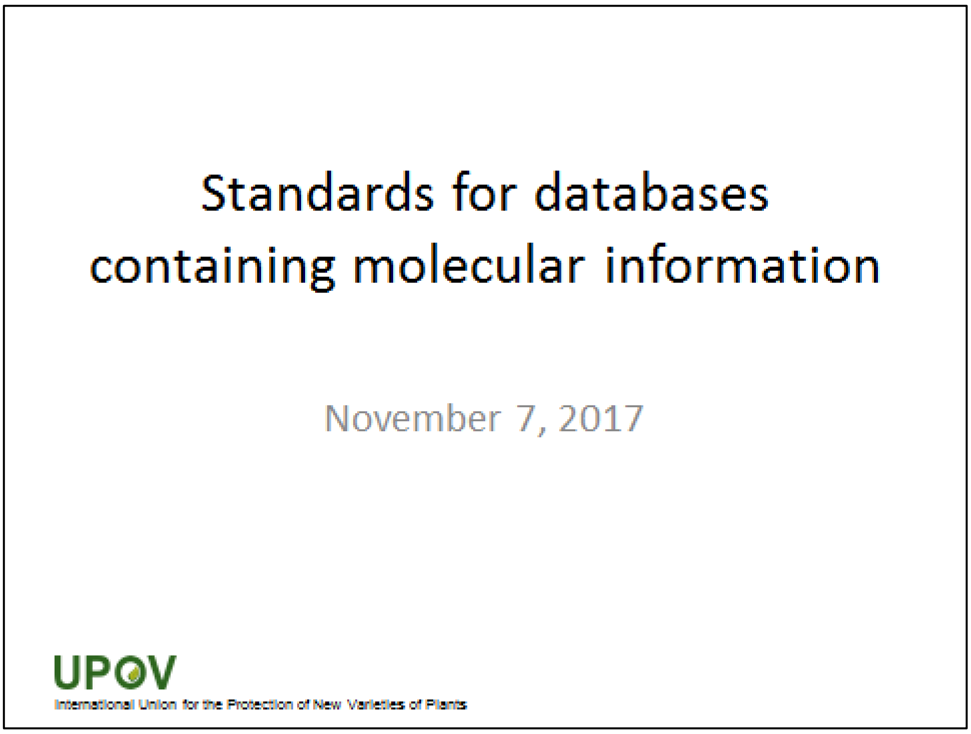
ADDENDUM TO
STANDARDS FOR DATABASES CONTAINING MOLECULAR INFORMATION

prepared by the Office of the Union

Disclaimer: this document does not represent UPOV policies or guidance

The Annex of this document contains a copy of a presentation “Standards for databases containing molecular information” to be made by the Office of the Union at the sixteenth session of the Working Group on Biochemical and Molecular Techniques, and DNA-Profiling in Particular (BMT).

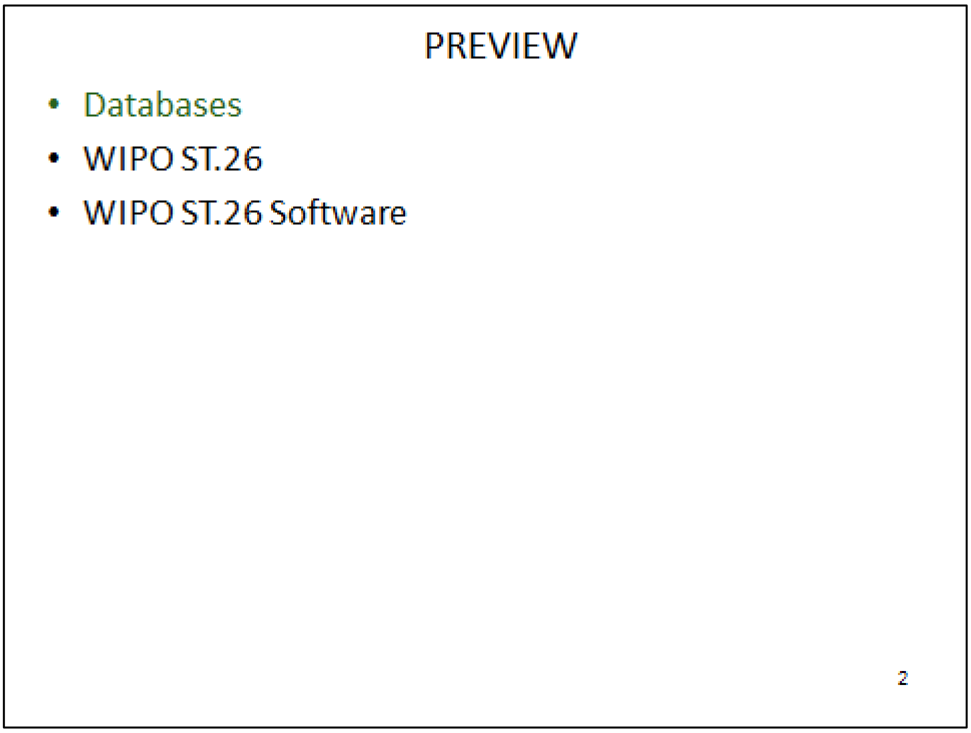
[Annex follows]



Standards for databases containing molecular information

November 7, 2017

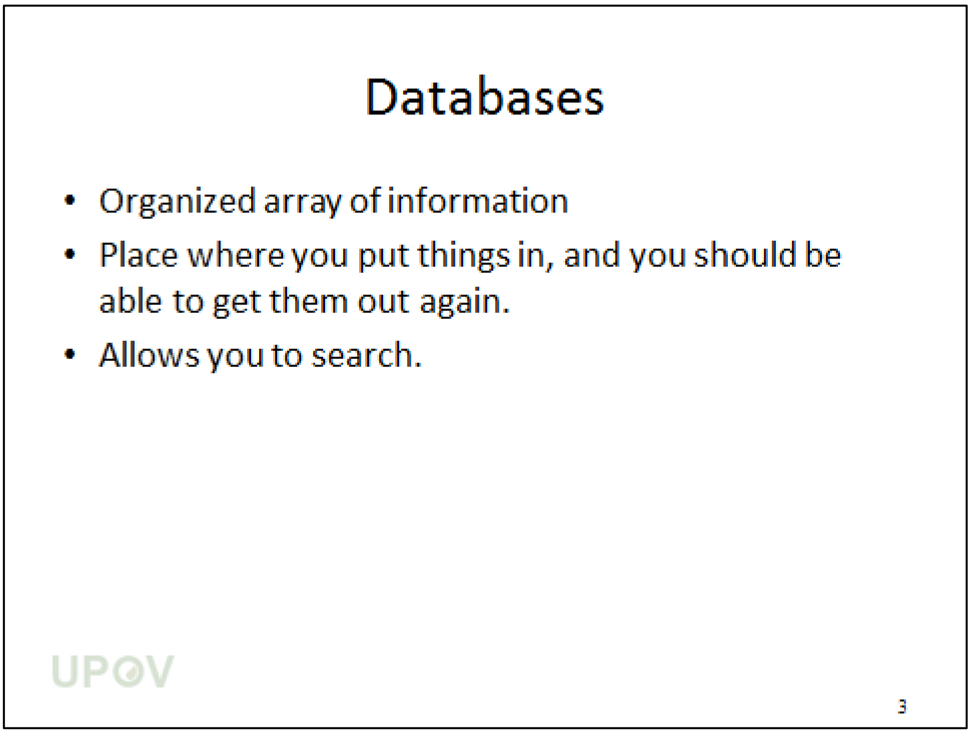
UPOV
International Union for the Protection of New Varieties of Plants



PREVIEW

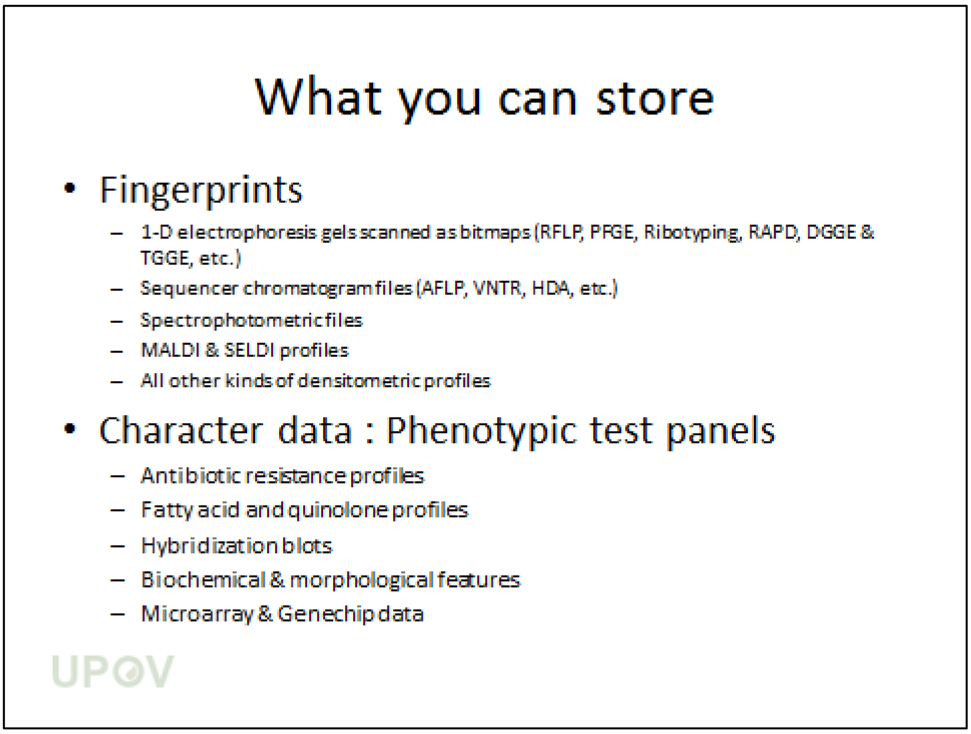
- Databases
- WIPO ST.26
- WIPO ST.26 Software

2



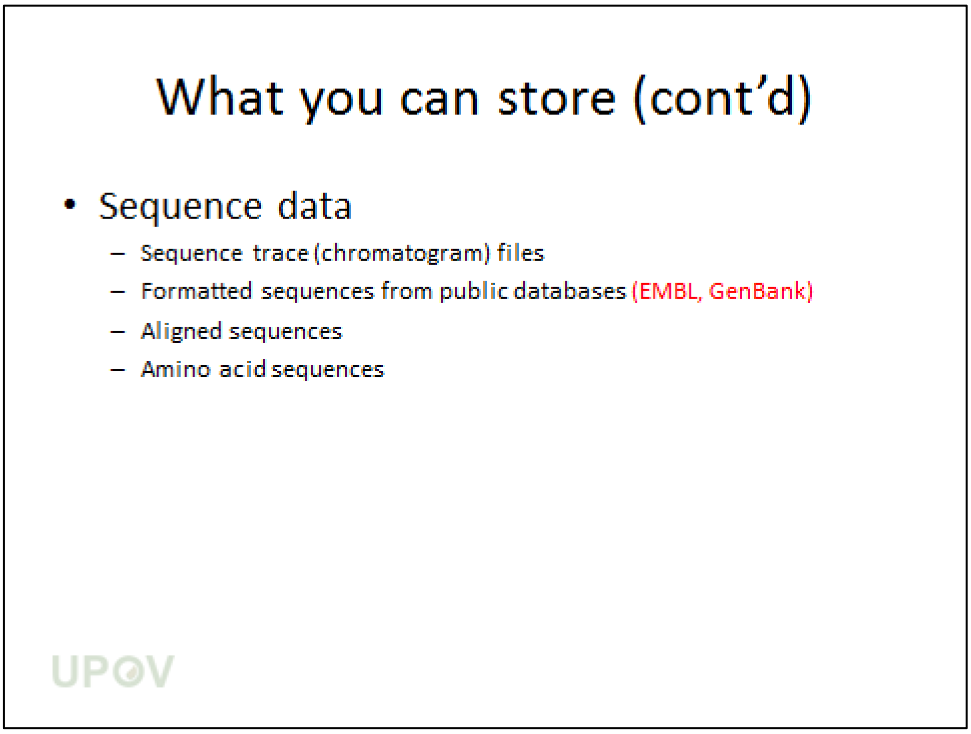
Databases

- Organized array of information
- Place where you put things in, and you should be able to get them out again.
- Allows you to search.



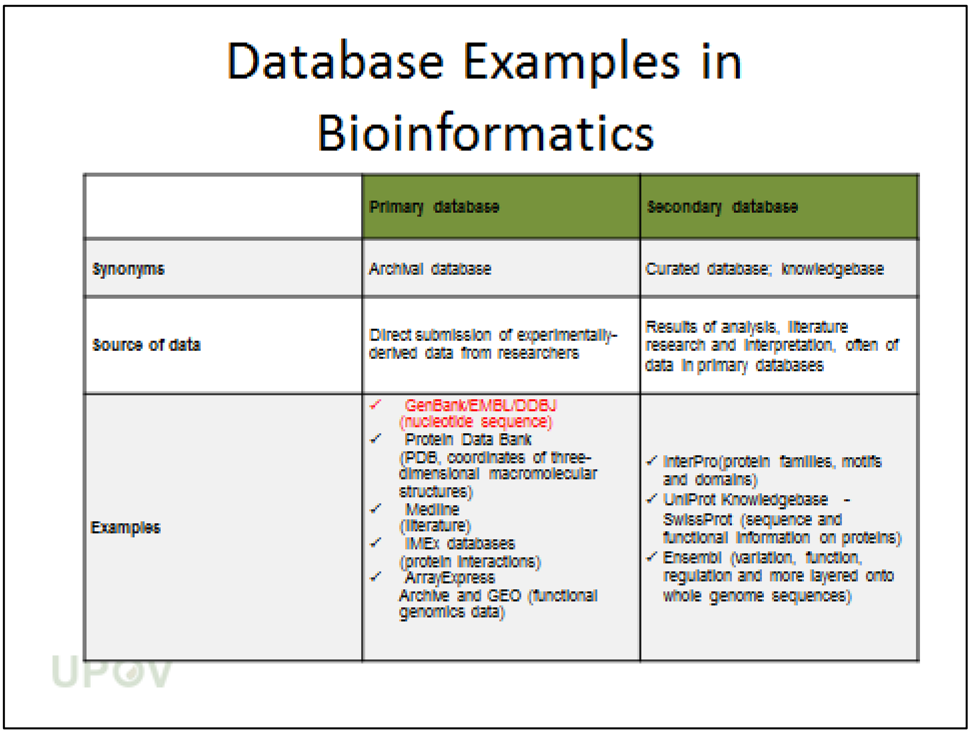
What you can store

- Fingerprints
  - 1-D electrophoresis gels scanned as bitmaps (RFLP, PFGE, Ribotyping, RAPD, DGGE & TGGE, etc.)
  - Sequencer chromatogram files (AFLP, VNTR, HDA, etc.)
  - Spectrophotometric files
  - MALDI & SELDI profiles
  - All other kinds of densitometric profiles
- Character data : Phenotypic test panels
  - Antibiotic resistance profiles
  - Fatty acid and quinolone profiles
  - Hybridization blots
  - Biochemical & morphological features
  - Microarray & Genechip data



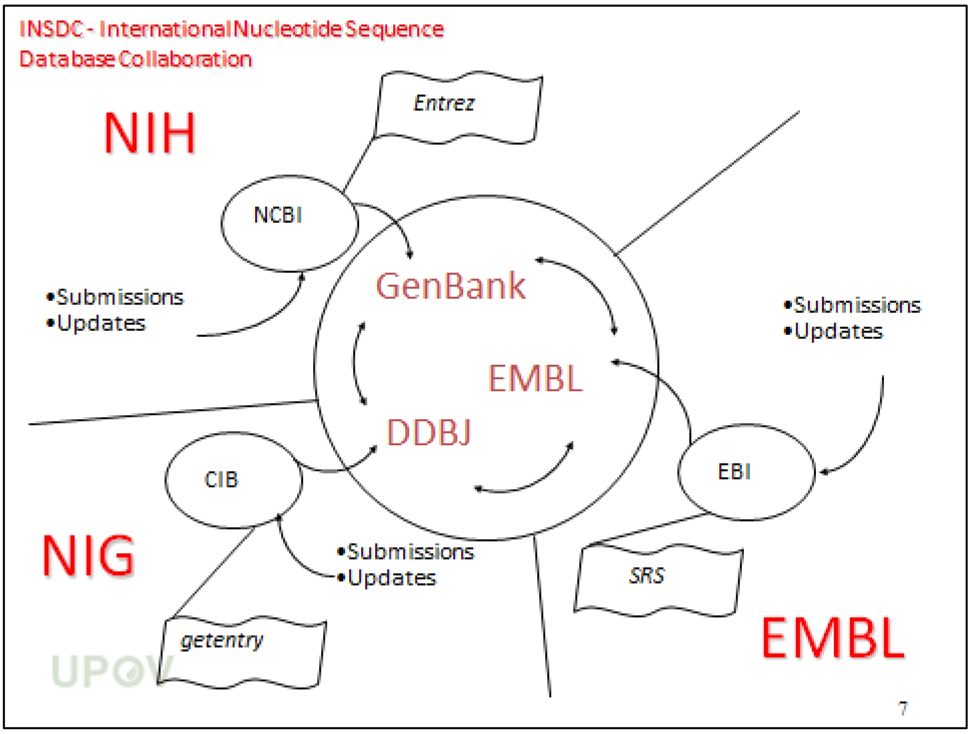
What you can store (cont'd)

- Sequence data
  - Sequence trace (chromatogram) files
  - Formatted sequences from public databases (EMBL, GenBank)
  - Aligned sequences
  - Amino acid sequences



Database Examples in Bioinformatics

	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)



INSDC - International Nucleotide Sequence Database Collaboration

NIH

NCBI

Entrez

- Submissions
- Updates

GenBank

EMBL

DDBJ

- Submissions
- Updates

EBI

CIB

NIG

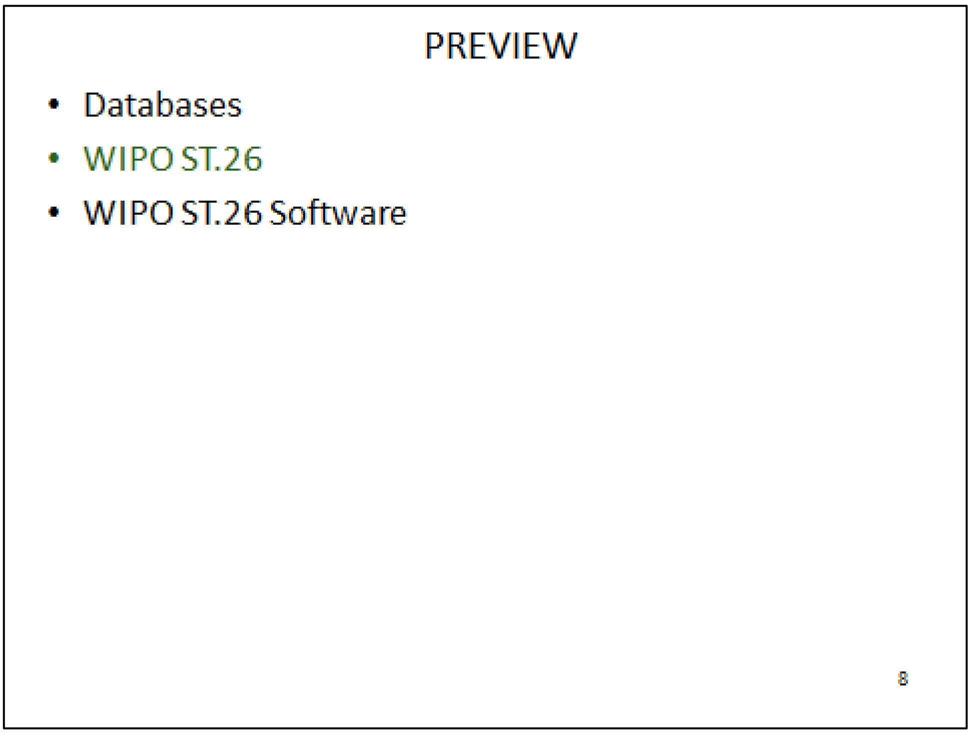
- Submissions
- Updates

SRS

getentry

EMBL

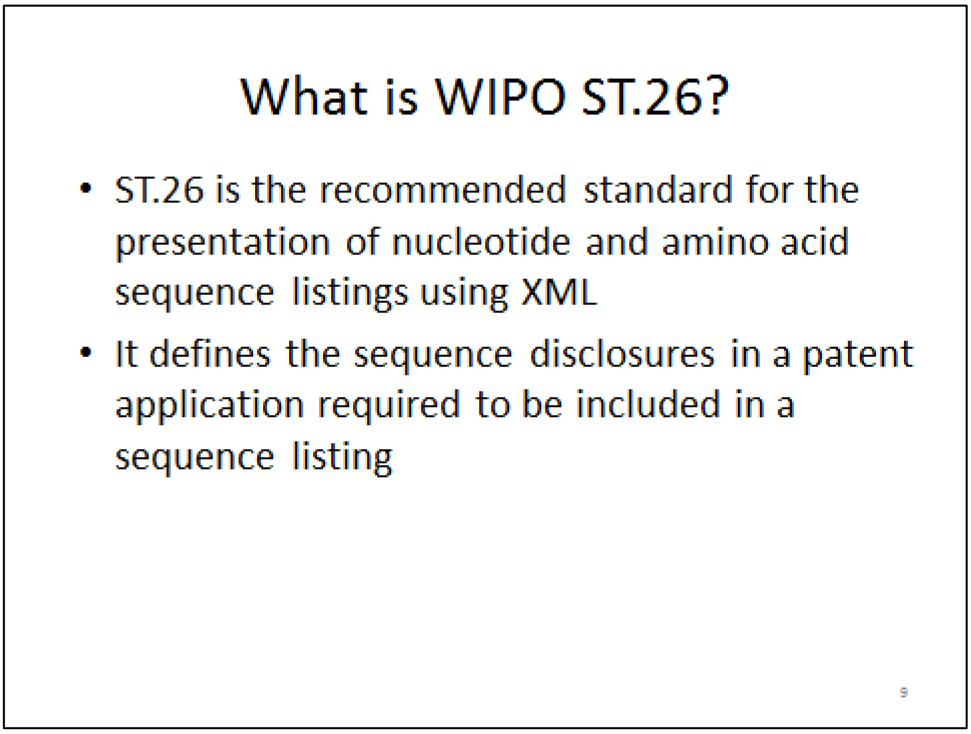
7



PREVIEW

- Databases
- WIPO ST.26
- WIPO ST.26 Software

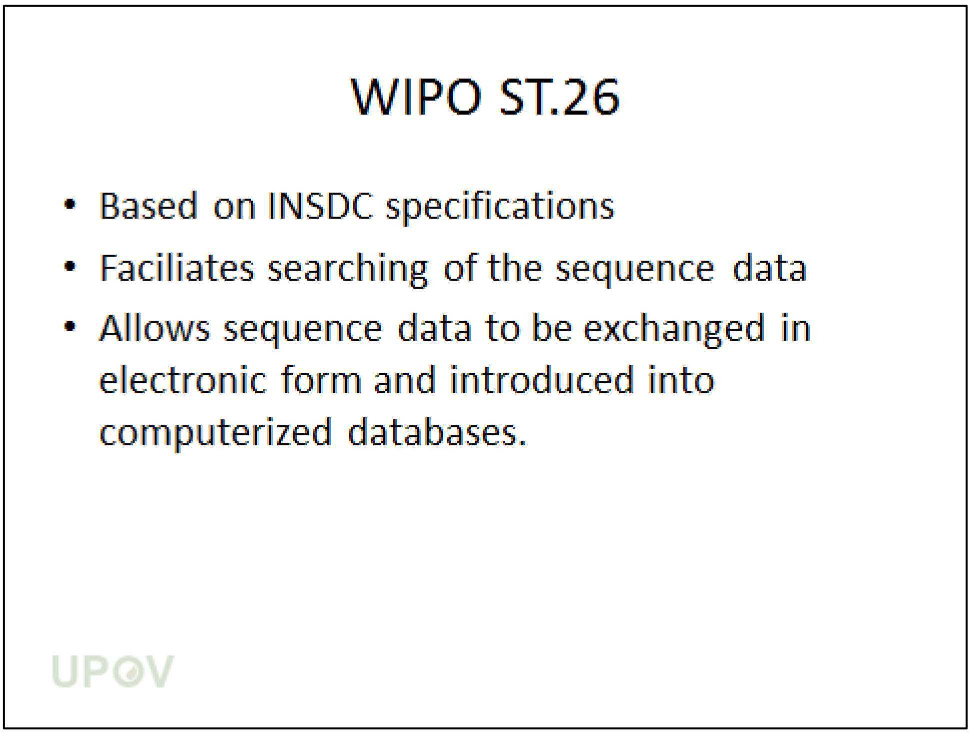
8



What is WIPO ST.26?

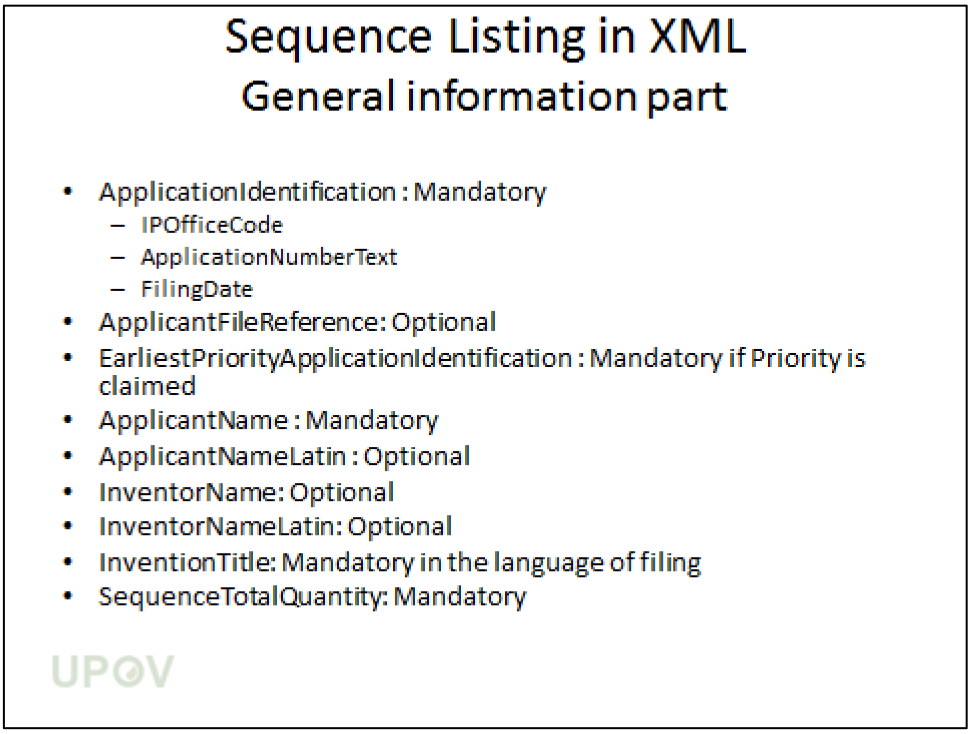
- ST.26 is the recommended standard for the presentation of nucleotide and amino acid sequence listings using XML
- It defines the sequence disclosures in a patent application required to be included in a sequence listing

9



WIPO ST.26

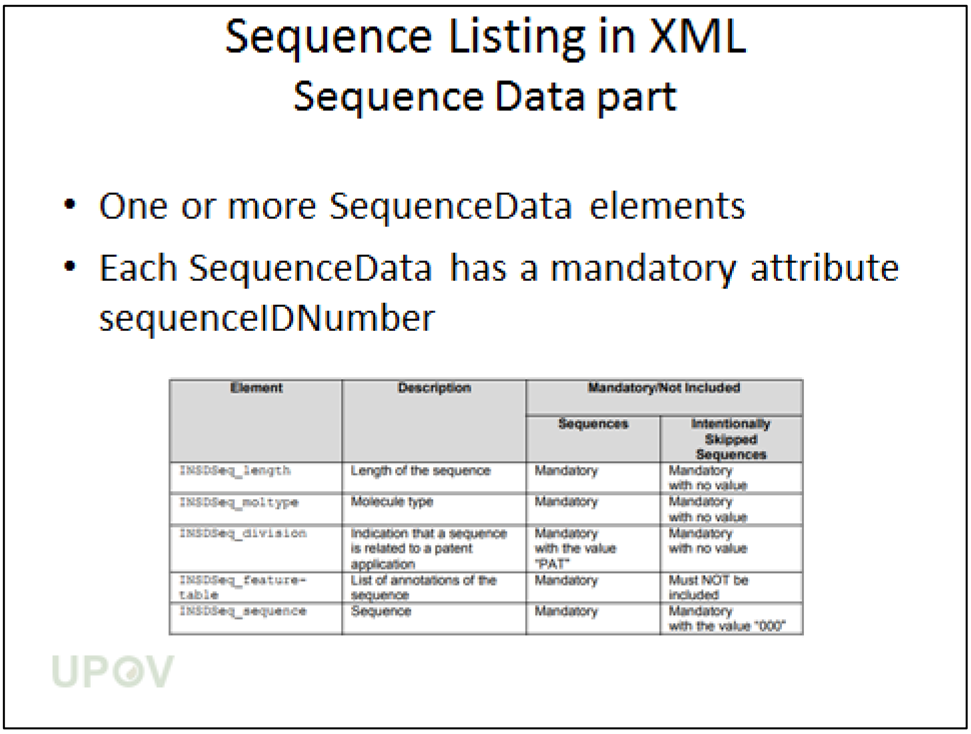
- Based on INSDC specifications
- Faciliates searching of the sequence data
- Allows sequence data to be exchanged in electronic form and introduced into computerized databases.



Sequence Listing in XML

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

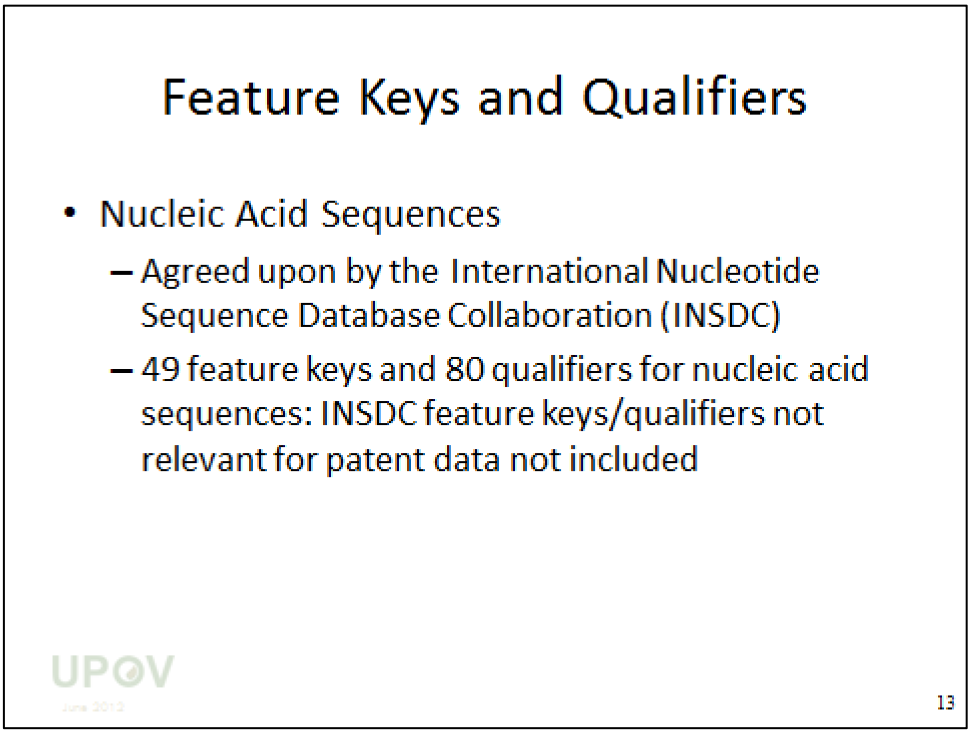


Sequence Listing in XML

Sequence Data part

- 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"

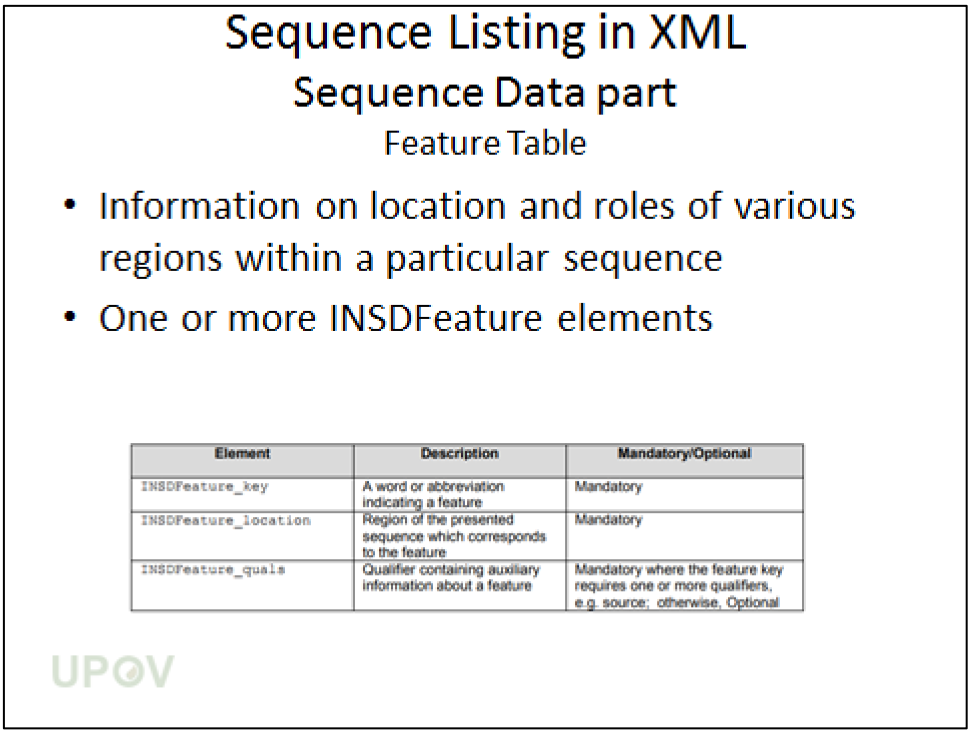


Feature Keys and Qualifiers

- Nucleic Acid Sequences
  - Agreed upon by the International Nucleotide Sequence Database Collaboration (INSDC)
  - 49 feature keys and 80 qualifiers for nucleic acid sequences: INSDC feature keys/qualifiers not relevant for patent data not included

June 2012

13



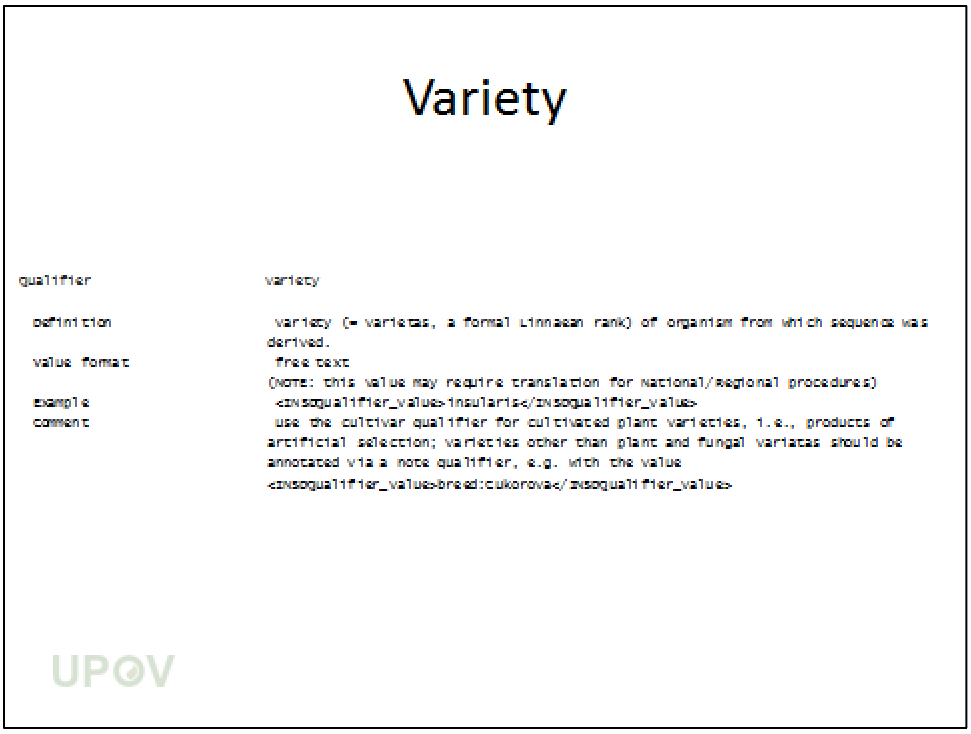
Sequence Listing in XML

Sequence Data part

Feature Table

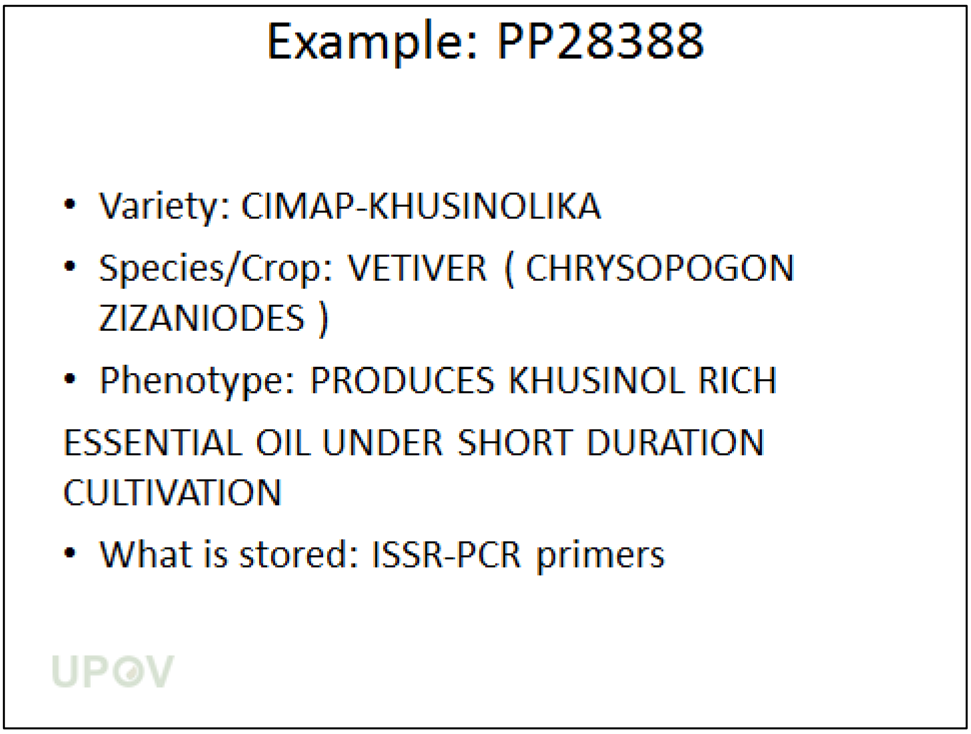
- Information on location and roles of various regions within a particular sequence
- One or more INSDFeature elements

Element	Description	Mandatory/Optional
INSDFeature_key	A word or abbreviation indicating a feature	Mandatory
INSDFeature_location	Region of the presented sequence which corresponds to the feature	Mandatory
INSDFeature_quals	Qualifier containing auxiliary information about a feature	Mandatory where the feature key requires one or more qualifiers, e.g. source; otherwise, Optional



Variety

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>

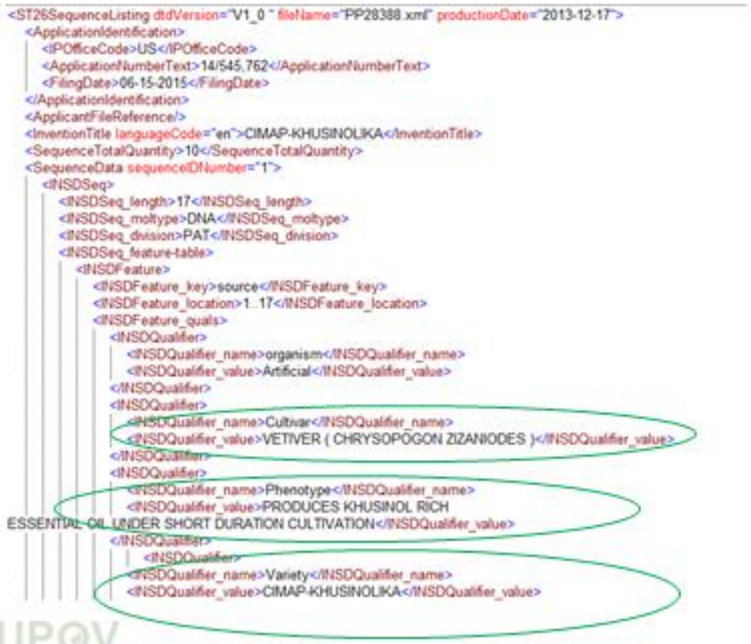


Example: PP28388

- Variety: CIMAP-KHUSINOLIKA
- Species/Crop: VETIVER (CHRYSOPOGON ZIZANIODES)
- Phenotype: PRODUCES KHUSINOL RICH

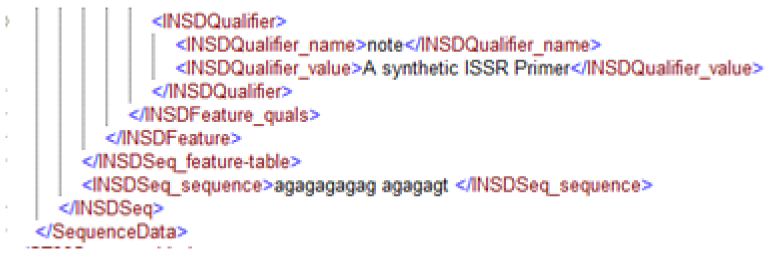
ESSENTIAL OIL UNDER SHORT DURATION CULTIVATION

- What is stored: ISSR-PCR primers



```
<ST26SequenceListing dtdVersion="V1_0 " fileName="PP28388.xml" productionDate="2013-12-17">
  <ApplicationIdentification>
    <IPOfficeCode>US</IPOfficeCode>
    <ApplicationNumberText>14/545,762</ApplicationNumberText>
    <FilingDate>06-15-2015</FilingDate>
  </ApplicationIdentification>
  <ApplicantFileReference/>
  <InventionTitle languageCode="en">CIMAP-KHUSINOLIKA</InventionTitle>
  <SequenceTotalQuantity>10</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">
    <INSDSeq>
      <INSDSeq_length>17</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..17</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier>
              <INSDQualifier_name>organism</INSDQualifier_name>
              <INSDQualifier_value>Artificial</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
              <INSDQualifier_name>Cultivar</INSDQualifier_name>
              <INSDQualifier_value>VETIVER ( CHRYSOPOGON ZIZANIODES )</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
              <INSDQualifier_name>Phenotype</INSDQualifier_name>
              <INSDQualifier_value>PRODUCES KHUSINOL RICH
ESSENTIAL OIL UNDER SHORT DURATION CULTIVATION</INSDQualifier_value>
            </INSDQualifier>
              <INSDQualifier>
              <INSDQualifier_name>Variety</INSDQualifier_name>
              <INSDQualifier_value>CIMAP-KHUSINOLIKA</INSDQualifier_value>
```

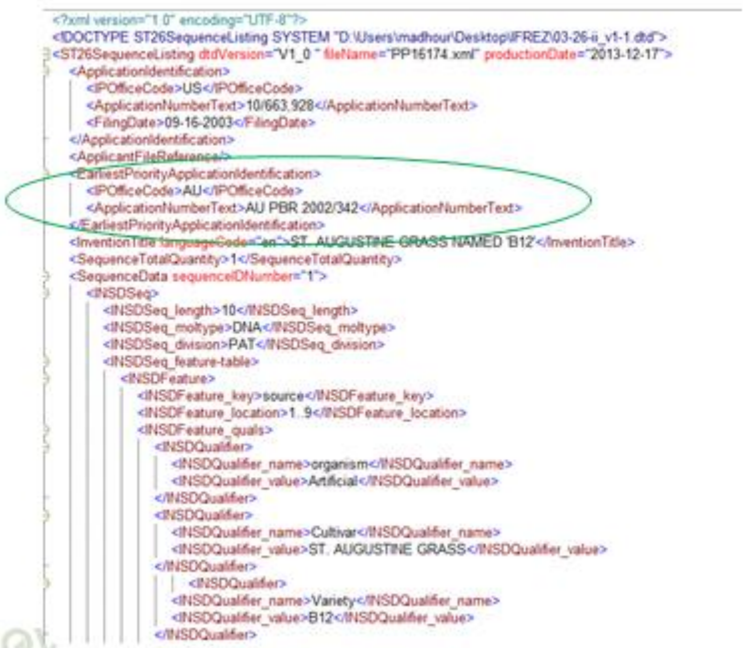
ISSR Primer



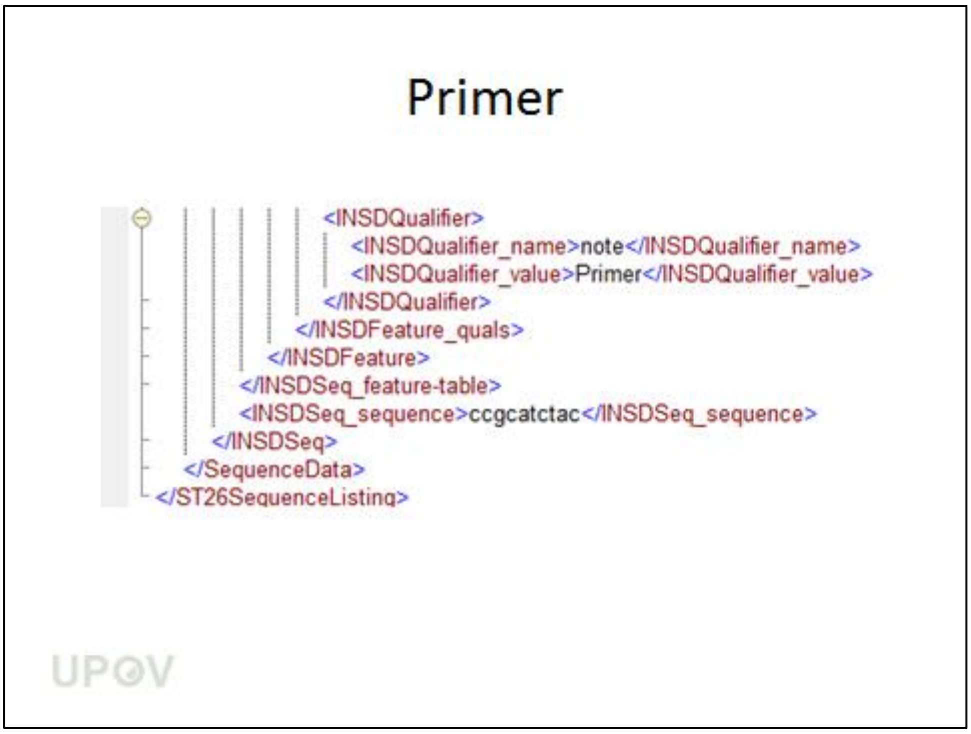
```
            <INSDQualifier>
              <INSDQualifier_name>note</INSDQualifier_name>
              <INSDQualifier_value>A synthetic ISSR Primer</INSDQualifier_value>
            </INSDQualifier>
          </INSDFeature_quals>
        </INSDFeature>
      </INSDSeq_feature-table>
      <INSDSeq_sequence>agagagagag agagagt </INSDSeq_sequence>
    </INSDSeq>
  </SequenceData>
```

Example: PP16174

- Variety: B12
- Sepecies/Crop: ST. AUGUSTINE GRASS
- Prior application number: AU PBR 2002/342
- What is stored: Primer
  - ccgcatctac

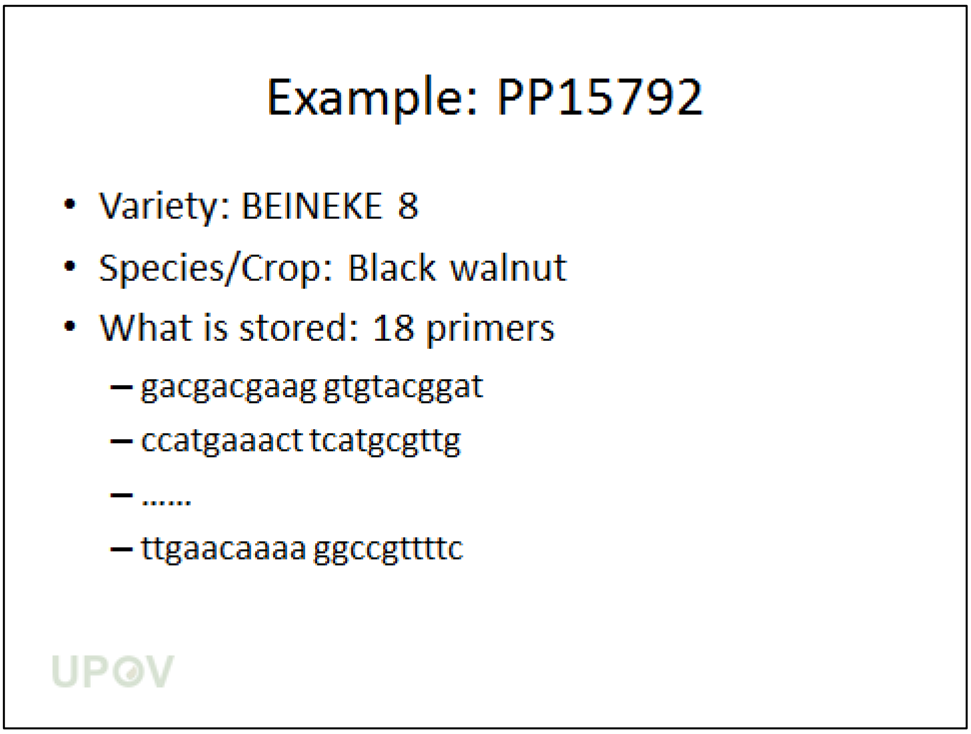


```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE ST26SequenceListing SYSTEM "D:\Users\madhour\Desktop\FREZ\03-26-ii_v1-1.dtd">
<ST26SequenceListing dtdVersion="V1_0 " fileName="PP16174.xml" productionDate="2013-12-17">
  <ApplicationIdentification>
    <IPOfficeCode>US</IPOfficeCode>
    <ApplicationNumberText>10/663,928</ApplicationNumberText>
    <FilingDate>09-16-2003</FilingDate>
  </ApplicationIdentification>
  <ApplicantFileReference/>
  <EarliestPriorityApplicationIdentification>
    <IPOfficeCode>AU</IPOfficeCode>
    <ApplicationNumberText>AU PBR 2002/342</ApplicationNumberText>
  </EarliestPriorityApplicationIdentification>
  <InventionTitle languageCode="en">ST. AUGUSTINE GRASS NAMED 'B12'</InventionTitle>
  <SequenceTotalQuantity>1</SequenceTotalQuantity>
  <SequenceData sequenceIDNumber="1">
    <INSDSeq>
      <INSDSeq_length>10</INSDSeq_length>
      <INSDSeq_moltype>DNA</INSDSeq_moltype>
      <INSDSeq_division>PAT</INSDSeq_division>
      <INSDSeq_feature-table>
        <INSDFeature>
          <INSDFeature_key>source</INSDFeature_key>
          <INSDFeature_location>1..9</INSDFeature_location>
          <INSDFeature_quals>
            <INSDQualifier>
              <INSDQualifier_name>organism</INSDQualifier_name>
              <INSDQualifier_value>Artificial</INSDQualifier_value>
            </INSDQualifier>
            <INSDQualifier>
              <INSDQualifier_name>Cultivar</INSDQualifier_name>
              <INSDQualifier_value>ST. AUGUSTINE GRASS</INSDQualifier_value>
            </INSDQualifier>
                <INSDQualifier>
              <INSDQualifier_name>Variety</INSDQualifier_name>
              <INSDQualifier_value>B12</INSDQualifier_value>
            </INSDQualifier>
```



Primer

```
          <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>Primer</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>ccgcatctac</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
</ST26SequenceListing>
```

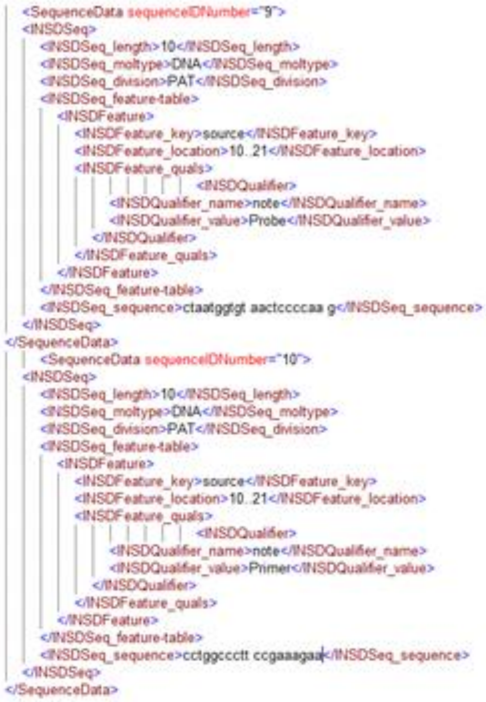


Example: PP15792

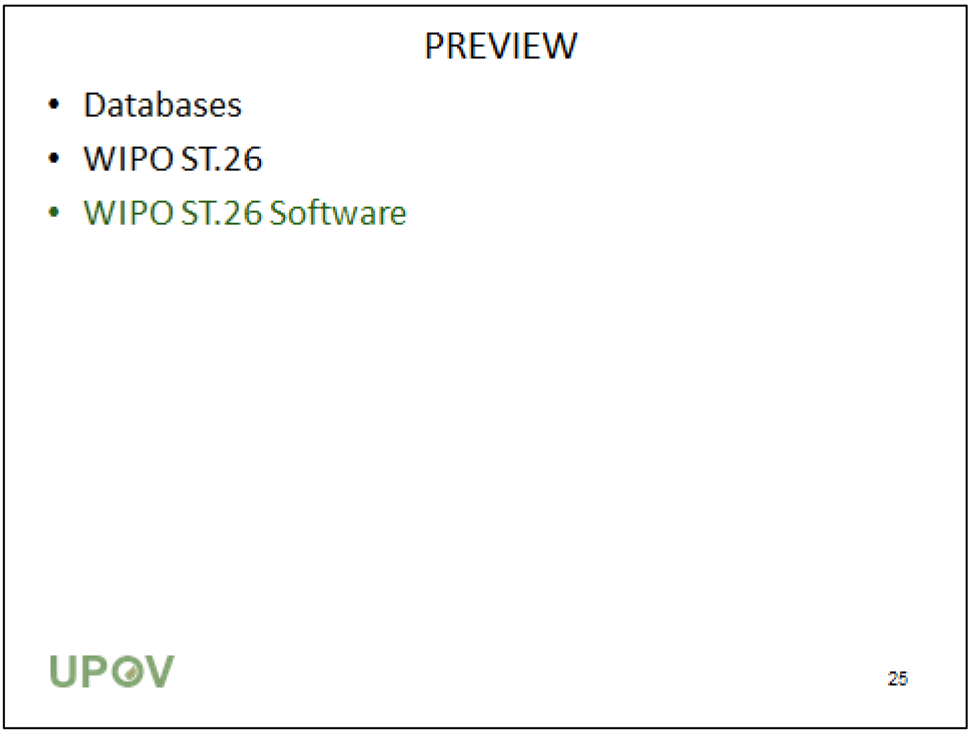
- Variety: BEINEKE 8
- Species/Crop: Black walnut
- What is stored: 18 primers
  - gacgacgaag gtgtacggat
  - ccatgaaact tcatgcgttg
  - ……
  - ttgaacaaaa ggccgttttc

Example: PCT/US2015/055339

- TOMATO PLANTS WITH IMPROVED DISEASE RESISTANCE
- UPOV TG/44/11 Char 57: resistance to Tomato yellow leaf curl virus
- What is stored: probes and primers



```
  <SequenceData sequenceIDNumber="9">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>10..21</INSDFeature_location>
        <INSDFeature_quals>
                      <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>Probe</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>ctaatggtgt aactccccaa g</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
    <SequenceData sequenceIDNumber="10">
  <INSDSeq>
    <INSDSeq_length>10</INSDSeq_length>
    <INSDSeq_moltype>DNA</INSDSeq_moltype>
    <INSDSeq_division>PAT</INSDSeq_division>
    <INSDSeq_feature-table>
      <INSDFeature>
        <INSDFeature_key>source</INSDFeature_key>
        <INSDFeature_location>10..21</INSDFeature_location>
        <INSDFeature_quals>
                      <INSDQualifier>
            <INSDQualifier_name>note</INSDQualifier_name>
            <INSDQualifier_value>Primer</INSDQualifier_value>
          </INSDQualifier>
        </INSDFeature_quals>
      </INSDFeature>
    </INSDSeq_feature-table>
    <INSDSeq_sequence>cctggccctt ccgaaagaa</INSDSeq_sequence>
  </INSDSeq>
</SequenceData>
```

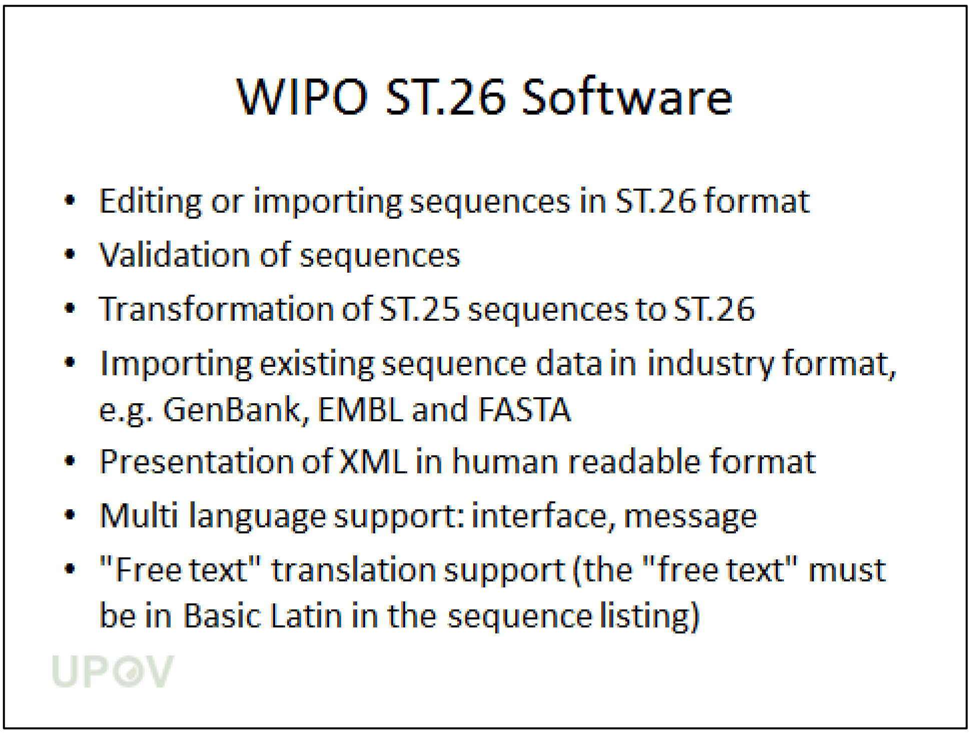


PREVIEW

- Databases
- WIPO ST.26
- WIPO ST.26 Software

UPOV

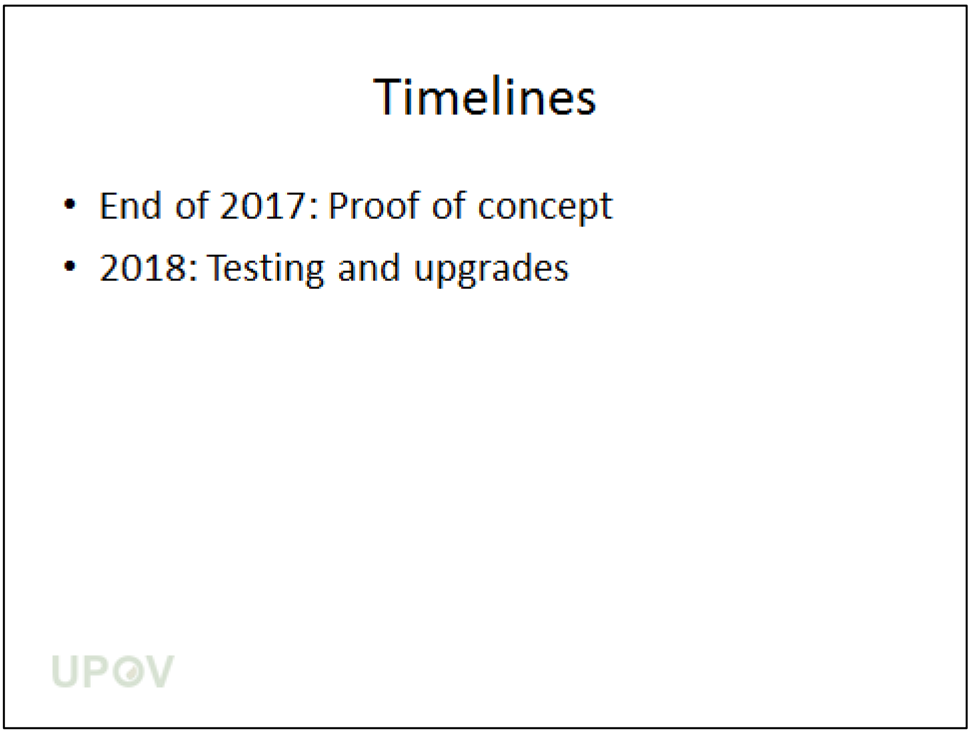
25



WIPO ST.26 Software

- Editing or importing sequences in ST.26 format
- Validation of sequences
- Transformation of ST.25 sequences to ST.26
- Importing existing sequence data in industry format, e.g. GenBank, EMBL and FASTA
- Presentation of XML in human readable format
- Multi language support: interface, message
- "Free text" translation support (the "free text" must be in Basic Latin in the sequence listing)

UPOV



Timelines

- End of 2017: Proof of concept
- 2018: Testing and upgrades

[End of Annex and of document]