E

CAJ/69/9
ORIGINAL: English
DATE: February 27, 2014

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

Geneva

ADMINISTRATIVE AND LEGAL COMMITTEE

Sixty-Ninth Session
Geneva, April 10, 2014

POSSIBLE DEVELOPMENT OF A UPOV SIMILARITY SEARCH TOOL
FOR VARIETY DENOMINATION PURPOSES

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 report on developments concerning the possible development of a UPOV similarity search tool for variety denomination purposes.

BACKGROUND

2. The Administrative and Legal Committee (CAJ), at its sixty-seventh session, held in Geneva on March 21, 2013, received a presentation from the Delegation of the European Union on the experience of the Community Plant Variety Office (CPVO) in the use of its denomination similarity search tool in the examination of proposed denominations. During the presentation, the CPVO proposed to explore the possibility to develop a UPOV similarity search tool for variety denomination purposes, which could be based on the CPVO search tool[1]. The CAJ welcomed the offer by the CPVO and agreed to include an item to consider that proposal at its sixty-eighth session, in October 2013 (see document CAJ/67/14 “Report on the Conclusions”, paragraphs 49 and 50).

3. The denomination search tab of the Plant Variety Database (PLUTO database) (https://www3.wipo.int/pluto/user/en/index.jsp) currently provides the following search types to find similar denominations:

(a) Similarity factor [CPVO search tool]

This will perform an analysis of the denomination you entered on a combination of factors including letters in common, relative lengths of the words and positions of the common letters. This is the most complex comparison method, and the search may take a few seconds to complete. The similarity factor has been developed by the French GEVES and the Community Plant Variety Office of the European Union (CPVO). However, please note that the results of the search by the similarity factor in the PLUTO database require interpretation and do not provide a guarantee as to the suitability of variety denominations which needs to be decided upon by the authority where plant variety rights is applied for.

A detailed explanation of the analysis is provided in the Annex to this document.

[1] The similarity factor was developed by the French *Group for Study and Control of Varieties and Seeds* (GEVES) and the Community Plant Variety Office of the European Union (CPVO)

(b) Fuzzy This will search for denominations that contain words spelled one or two characters differently from the terms you entered. This is similar to the Fuzzy match method in the Term Search tab.

(c) Phonetic This will search for denominations that contain words that sound similar to the terms you entered. This is similar to the Phonetic match method in the Term Search tab.

(d) Contains This will search for denominations that contain words that contain the same series of letters as the terms you entered. This is similar to the contains match method in the Term Search tab.

(e) Starts This will search for denominations that contain words that start with the same series of letters as the terms you entered. This is similar to the starts match method in the Term Search tab.

(f) Ends This will search for denominations that contain words that end with the same series of letters as the terms you entered. This is similar to the ends match method in the Term Search tab.

4. In exploratory discussions with the Office of the Union on how to develop a UPOV similarity search tool for variety denomination purposes, the CPVO clarified that all options should be considered and that, in the light of advances in information technology, the best tool might not necessarily use the CPVO search tool as a starting point. The main consideration would be to develop a tool that could be used by all UPOV members in order to minimize differences in the decisions on the suitability.

ESTABLISHMENT OF A WORKING GROUP

5. The CAJ, at its sixty-eighth session, held in Geneva, on October 21, 2013, considered document CAJ/68/9 “Possible development of a UPOV similarity search tool for variety denomination purposes” and approved the establishment of a working group to develop proposals for a UPOV similarity search tool for variety denomination purposes, as proposed in document CAJ/68/9, paragraphs 4 to 7, as follows (see document CAJ/68/10 “Report on the Conclusions”, paragraph 40):

> Composition of the working group:
>
> (a) Denomination examiners from members of the Union (3 to 6 experts);
>
> (b) WIPO Global Databases Service (responsible for the PLUTO database);
>
> (c) Community Plant Variety Office of the European Union (CPVO); and
>
> (d) Office of the Union.
>
> The work plan of the working group will be established by the working group itself; however, it is anticipated that the first step will be to review the search types currently available in the denomination search tab of the PLUTO database, particularly the Similarity factor (CPVO search tool), and to review search types in use in other situations (e.g. in relation to trademarks) that might provide an alternative basis for a UPOV similarity search tool.
>
> The review of the suitability of search types will, in particular, take into account document UPOV/INF/12 “Explanatory notes on variety denominations under the UPOV Convention”. In that regard, the working group will need to refer to the CAJ for further guidance if its work indicates that a review of document UPOV/INF/12 would be necessary for the development of an effective UPOV similarity search tool.
>
> The meetings of the working group will be hosted by the Office of the Union in Geneva and will be chaired by the Office of the Union. The meetings will not be arranged to coincide with UPOV sessions and electronic participation by denomination examiners and the CPVO will be anticipated.

Proposals developed by the working group will be presented to the CAJ and to the Technical Committee (TC), and the CAJ and TC will receive a brief report of the meetings of the working group.

6. The CAJ, at its sixty-eighth session, noted the suggestion by the Delegation of the European Union for the inclusion in the working group of denomination examiners from the Netherlands and Spain and the importance of ensuring that there was sufficient coverage by the experts of the linguistic aspects of variety denominations (see document CAJ/68/10 “Report on the Conclusions”, paragraph 41).

7. The CAJ, at its sixty-eighth session, agreed that members and observers should be encouraged to provide suggestions on matters concerning the tasks of the working group (see document CAJ/68/10 “Report on the Conclusions”, paragraph 42).

8. The TC, at its fiftieth session, to be held in Geneva, from April 7 to 9, 2014, will receive a report on the developments concerning the possible development of a UPOV similarity search tool for variety denomination purposes (see document TC/50/14 “Variety denominations”). The comments of the TC at its fiftieth session will be reported to the CAJ at its sixty-ninth session. The first meeting of the working group will be arranged for June/July, 2014, and a report will be made to the CAJ at its seventieth session, to be held in Geneva, in October 2014.

9. The CAJ is invited to note that:

(a) the comments of the TC concerning the possible development of a UPOV similarity search tool for variety denomination purposes at its fiftieth session will be reported to the CAJ at its sixty-ninth session; and

(b) the first meeting of the working group will be arranged for June/July, 2014, and a report will be made to the CAJ at its seventieth session, to be held in Geneva, on October 13 and 14, 2014.

[Annex follows]

CPVO
Community Plant Variety Office

SEARCHING PROCEDURE

1. GENERAL

As a conclusion of the study phase of the project presented to its AC in November 2003, the Office proposed in a first instance, to take over National software, to adapt them to the centralised database and to run them all for each test. Possibility to miss one close denomination would that way be very limited

On a longer run, the development of a CPVO software was foreseen, with the possibility of development of linguistic features.

In practice, the specifications of the French software have been taken as a basis for implementation of the searching procedure in the CPVO database.

2. RULES ESTABLISHING SUFFICIENT DISTINCTNESS BETWEEN 2 VARIETY DENOMINATIONS

According to the Basic regulation on Community plant variety rights, one of the rules a variety denomination must fulfill is that it should not be identical to/may be confused with a variety denomination under which another variety of the same/a closely related species has been registered.

This rule has been interpreted in the guidelines of the Administrative Council of the CPVO on variety denominations:

♣ A difference of only one letter or number, or of an accent on a letter, should generally be regarded as confusing.

♣ Differences of two or more letters should not generally be regarded as confusing except where the same letters are simply juxtaposed.

♣ Moreover, a variety denomination should not convey the false impression that the variety is related to, or derived from, another variety;

The purpose of the searching program will be to discover denominations of the same class in the database that could be in conflict with a proposed denomination

3. SEARCHING PROCEDURE

The tests are carried out by an internal program of the ORACLE database on the CPVO server (better performance).

To carry out a test, the interface program (web site, …) executes a procedure called:

TESTDENOMINATION

As parameters, the interface program transmits the denomination to be tested and the Species code to which the variety belongs.

The procedure returns the identifier of the test carried out to the interface (column Testid). With this identifier we can read the 2 tables constituting the result of the test: tables TESTS and TESTRESULTS.

The table TESTS contains general data on the test: date of the test, identifier of the person requesting the test, denomination, species code, class or genus code, computer running time, excluded words, error message, ….

The table TESTRESULTS contains the lists of the denominations which have been found as similar by the procedure TESTDENOMINATION. Each similar denomination is coupled with a similarity index.

4. Description of the procedure TESTDENOMINATION

Input control

The species code must exist in the denominations database.
The maximal length of the denomination is 100 characters (blank characters included).
Excluded characters: punctuation characters and /-_ ‘. Stressed characters are not allowed.

Denominations composed of several words are allowed: 4 words maximum. They must be separated by a blank space.

a) First operation : split of the entry denomination into individual words

The denomination tested is cut in elementary words. Blank characters are considered as separators and are deleted.

Non Latin standard characters (Stressed characters,…) are replaced by Latin standard characters. All letters are converted in capital letters. See the translation table.
Example: Déjà becomes DEJA

Double letters are reduced to a single letter.
Example : HELLO becomes HELO.

Each elementary word is compared to the list of the words excluded for testing (i.e.: color YELOW, RED, PURPLE,…)

Elementary words found in this list are excluded from the similarity test.

In the following description, the elementary words obtained are named the Words Tested (WT)

Example: Denomination ‘Tänau TARI YELLOW’

The word yellow is excluded from the tests. Words taken into account (WT) are :

- TANAU,
- TARI,
- TANAUTARI,
- TARIYELOW,
- TANAUTARIYELOW.

b) Second operation : building up a list of words of reference in the class

The program searches the class or genus attached to the species to determine the scope of the similarity search.

In case of denominations registered as codes, one single string of characters is considered, without blank spaces.

The software builds up a list of all the elementary words belonging to denominations of the varieties of the class.

All of the words with more than 3 characters as the longest WT and less than 3 characters of the shortest WT are not taken into account.

All the words of reference that belong to the list of the words excluded for testing mentioned above (i.e.: color YELLOW, RED, PURPLE, …) are excluded from the similarity test.

The Same way as above, double letters of the words of reference are reduced to a single letter.

In the following description, the elementary words included in this list are named the words of reference (WR)

c) The principles of the similarity test

For each WT, the procedure SIMILARITYTEST calculates an index of similarity against each WR included into the list of denominations built up above.

The list of WR is sorted out according to the value of the similarity index. All of the WR with a similarity index superior to a predefined threshold are excluded from the results.

Example:

Denomination tested 'Tänau TARI YELLOW'

++

The list of words of reference contains 10.000 elementary words.

TANAU, TARI, TANAUTARI, TARIYELOW and TANAUTARIYELOW are tested against all 10.000 WR. 50.000 tests are carried out (5 words x 10 000 WR).

5. Detailed description of the similarity test.

The steps below are implemented for each couple (WT,WR).

Preliminary filter: in the list of WR, all words with more than 3 characters as the WT, or less than 3 characters as the WT are not considered for the following calculations.

First step : Calculation of Ki2

Formula : Ki2 = sum(di)^{2}/(Length(WT)-1)(length(WR)-1)

Where di = difference of number of letters between the word tested and the word of reference. All letters of both words are taken into account.

Example :

WT : **ALADIN** length 6 characters.

If we compare this WT to the existing character string: **DYLAN** (5 characters)

	A	L	D	I	N	Y
ALADIN	2	1	1	1	1	0
DYLAN	1	1	1	0	1	1

Chi2 = $((2\text{-}1)^2 + (1\text{-}1)^2 + (1\text{-}1)^2 + (1\text{-}0)^2 + (1\text{-}1)^2 + (0\text{-}1)^2) / (6\text{-}1)(5\text{-}1)$

Chi2 = (1 + 0 + 0 + 1 + 0 + 1)(5*4)

Chi2 = 3/20

Chi2 = 0,15

Then we keep for the following calculation all words where:

- Ki2<=0,3 and length of WT >=5 characters
- Ki2<=0,4 and length of WT = 4 characters
- Ki2<=0,5 and length of WT < 4 characters

Second step : Four calculations based on the selection of first step.

4 calculations are carried out in this second step:

- Calculation of the percentage of common letters
- Calculation of the percentage of NON common letters
- Calculation of the percentage of difference of length.
- Calculation of rank Kendall correlation

Calculation of the percentage of common letters (CL)

CL = 1-(Nb of common letters)/ (length (WT))

Example : ALADIN and DYLAN
CL = 1-4/6 = 0.33

CL is equal to 0 when all letters are found in the WR.

Second example: BANANAS and BANS

All letters of BANANAS can be found in the word BANS.
CL= 1- 7/7 => CL=0

Calculation of the percentage of NON common letters (NCL)

NCL = (Nb of letters in WR not in WT)/ (length (WT))

Example : ALADIN and DYLAN
NCL = 1/6 = 0.16

NCL is equal to 0 when all letters in the WR are in the WT.

Second example: BANANAS and BANS

All letters of BANS can be found in the word BANANAS.
NCL= 0/7 => NCL=0

Calculation of the percentage of difference of length.

DL = (Difference of length between the 2 strings)/ (length of WT)

Example : BANANAS and ANANAS
DL = 1/7
DL is equal to 0 when the lengths of the 2 words are equal.

Calculation of rank Kendall correlation

Formula : KC = 6 *sum(Di)^2/N*(N^2-1)

Where :

- Di is equal to the difference of position of the common letters (Li) minus the difference of position of the previous letters . If the WT has several occurrences of the same letter, we use the first letters as a reference for the position.
  If the WR has several occurrences of the same letter, we use the closest letters as the same letter in the WT.
- N is equal to the number of common letters between WT an WR

Example :

WT : **ALADIN** and **DYLAN** .

	A	**L**	**D**	**N**
ALADIN	1	2	4	6
DYLAN	4	3	1	5
Difference	1-4=-3	2-3=-1	4-1=3	6-5=1
Di	-3-0 = -3	1-(-3) = 4	3-(-1) = 4	1-3 = -2

4 letters in common A, L, D and N

KC = 6* ($(-3)^2 + (4)^2 + (4)^2 + (-2)^2$) / (4*($4^2$ -1))

KC = 6* (9 + 16 + 16 + 4) / 4*15

KC = 6* 45 / 60

KC = 4.5

WT : **ALADIN** and **BALADIN** .

	A	**L**	**D**	**I**	**N**
ALADIN	1	2	4	5	6
BALADIN	2	3	5	6	7
Difference	1-2=-1	2-3=-1	4-5=-1	5-6=-1	6-7=-1
Li	-1-0 = -1	-1-(-1) = 0	-1-(-1) = 0	-1-(-1) = 0	-1-(-1)=0

5 letters in common A, L, D, I and N

KC = 6* ($(-1)^2 + (0)^2 + (0)^2 + (0)^2 + (0)^2$) / (5*($5^2$ -1))

KC = 6* (1) / 5*24

KC = 1 / 20

KC = 0,05

In this example we can see that if the same sequence of letters are present in the 2 words (LADIN in our example), the gap between the 2 sequences is taken into account only one time.

If the 2 words WT and WR are identical then the Kendall rank is equal to 0.

Third step : Calculation of the similarity index

Similarity index = KC + CL + NCL + DL

A reference word is **selected** if the scores are inferior or equal to :

	KC	**CL**	**DL**	**Similarity index**
Length Searched word > 4 characters	<=1,5	<=0,22	<=1,5	<1,2
Length Searched word = 4 characters	<=1,5	<=0,25	<=1,26	<1,2
Length Searched word < 4 characters	<=1	<=0,34	<=1,0	<1,2

Reference denominations are sorted out for displaying result by similarity index then by alphabetical order.

[End of Annex and of document]