



CAJ/68/9

ORIGINAL : anglais

DATE : 2 septembre 2013

UNION INTERNATIONALE POUR LA PROTECTION DES OBTENTIONS VÉGÉTALES

Genève

COMITÉ ADMINISTRATIF ET JURIDIQUE

**Soixante-huitième session
Genève, 21 octobre 2013**

POSSIBILITÉ D'ÉLABORATION D'UN OUTIL DE RECHERCHE DE L'UPOV DE SIMILARITÉ AUX FINS DE LA DÉNOMINATION VARIÉTALE

Document établi par le Bureau de l'Union

Avertissement : le présent document ne représente pas les principes ou orientations de l'UPOV

1. Le Comité administratif et juridique (CAJ), à sa soixante-septième session tenue à Genève le 21 mars 2013, a assisté à un exposé de la délégation de l'Union européenne sur l'expérience de l'Office communautaire des variétés végétales (OCVV) concernant l'utilisation de son outil de recherche de similarité aux fins de la dénomination lors de l'examen des dénominations proposées. Au cours de l'exposé, l'OCVV a proposé d'envisager la possibilité d'élaborer un outil de recherche de similarité pour l'UPOV aux fins de la dénomination variétale, qui pourrait s'inspirer du moteur de recherche de l'OCVV¹. Le CAJ s'est félicité de l'offre de l'OCVV et il est convenu d'inscrire un point à l'ordre du jour pour examiner cette proposition à sa soixante-huitième session (voir les paragraphes 49 et 50 du document CAJ/67/14 "Compte rendu des conclusions").

RAPPEL

2. L'onglet Recherche de dénomination de la base de données sur les variétés végétales (base de données PLUTO) (<https://www3.wipo.int/pluto/user/fr/index.jsp>) propose à l'heure actuelle les modes de recherche ci-après, pour trouver des dénominations similaires :

- a) Facteur de similarité [outil de recherche de l'OCVV] Ce mode permet d'analyser la dénomination que saisie en fonction d'une combinaison de facteurs tels que des lettres en commun, les longueurs relatives des mots et les positions des lettres communes. Il s'agit de la méthode de comparaison la plus complexe, et la recherche peut prendre quelques secondes. Le facteur de similarité a été élaboré par le Groupe d'étude et de contrôle des variétés et des semences (GEVES) (France) et l'Office communautaire des variétés végétales (OCVV). Veuillez toutefois noter que les résultats de la recherche à l'aide du facteur de similarité dans la base de données PLUTO nécessitent une interprétation et ne donnent pas de garantie quant à l'éligibilité des dénominations variétales, qui doit être décidée par l'administration à laquelle sont adressées les demandes de droits d'obtenteur.

Une explication détaillée de l'analyse figure dans l'annexe du présent document.

¹

Le facteur de similarité a été élaboré par le Groupe d'étude et de contrôle des variétés et des semences (GEVES) (France) et l'Office communautaire des variétés végétales (OCVV).

- b) Flou Ce mode permet de rechercher les dénominations contenant des mots qui s'écrivent comme les termes saisis à un ou deux caractères près. Il est semblable à la méthode de mise en correspondance Flou de l'onglet Recherche de termes.
- c) Phonétique Ce mode permet de rechercher les dénominations contenant des mots qui ressemblent phonétiquement aux termes que vous avez saisis. Il est semblable à la méthode de mise en correspondance Phonétique de l'onglet Recherche de termes.
- d) Contient Ce mode permet de rechercher les dénominations contenant des mots qui comprennent la même série de lettres que les termes saisis. Il est semblable à la méthode de mise en correspondance Contient de l'onglet Recherche de termes.
- e) Commence Ce mode permet de rechercher les dénominations contenant des mots qui commencent par la même série de lettres que les termes que vous avez saisis. Il est semblable à la méthode de mise en correspondance Commence de l'onglet Recherche de termes.
- f) Finit Ce mode permet de rechercher les dénominations contenant des mots qui finissent par la même série de lettres que les termes que vous avez saisis. Il est semblable à la méthode de mise en correspondance Finit de l'onglet Recherche de termes.

3. Lors de discussions préliminaires avec le Bureau de l'Union sur la façon d'élaborer un outil de recherche de l'UPOV de similarité aux fins de la dénomination variétale, l'OCVV a précisé que toutes les options devraient être envisagées et que, au vu des avancées dans le domaine des technologies de l'information, l'outil le plus efficace ne serait pas nécessairement basé sur le moteur de recherche de l'OCVV. La préoccupation première devrait être d'élaborer un outil que tous les membres de l'UPOV puissent utiliser, afin de limiter les divergences dans les décisions relatives à l'éligibilité.

PROPOSITION

4. Compte tenu du besoin de connaissance en matière d'examen des dénominations variétales et d'outils informatiques, il est proposé de créer un petit groupe de travail chargé d'élaborer des propositions relatives à un outil de recherche de l'UPOV de similarité aux fins de la dénomination variétale. Ce groupe pourrait être composé comme suit :

- a) Examineurs de dénominations en provenance de membres de l'Union (trois à six experts);
- b) Service des bases de données mondiales de l'OMPI (responsable de la base de données PLUTO);
- c) Office communautaire des variétés végétales (OCVV) de l'Union européenne; et
- d) Bureau de l'Union.

5. Le programme de travail du groupe de travail sera établi par le groupe de travail lui-même; cependant, il est prévu que la première étape consiste à examiner les modes de recherche déjà disponibles dans l'onglet Recherche de dénomination de la base de données PLUTO, notamment le facteur de similarité (outil de recherche de l'OCVV), ainsi qu'à examiner les modes de recherche utilisés dans d'autres situations (p. ex. en lien avec les marques) susceptibles de fournir une autre base à un outil de recherche de l'UPOV de similarité.

6. L'examen de la pertinence des modes de recherche tiendra compte, en particulier, du document UPOV/INF/12 "Notes explicatives concernant les dénominations variétales en vertu de la Convention UPOV". À cet égard, le groupe de travail devra se référer au CAJ pour obtenir des indications supplémentaires si ses travaux indiquent qu'un examen du document UPOV/INF/12 est nécessaire à l'élaboration d'un outil de recherche de l'UPOV de similarité efficace.

7. Il est proposé que les réunions du groupe de travail soient accueillies et présidées par le Bureau de l'Union à Genève. Les réunions ne seront pas organisées de sorte à coïncider avec les sessions de l'UPOV et la participation électronique des examinateurs de dénominations et de l'OCVV sera prévue. Les propositions élaborées par le groupe de travail seront présentées au CAJ ainsi qu'au Comité technique (TC), et ces comités recevront un bref compte rendu des réunions du groupe de travail.

8. *Le CAJ est invité :*

a) à approuver la création d'un groupe de travail chargé d'élaborer des propositions relatives à un outil de recherche de l'UPOV de similarité aux fins de la dénomination variétale, proposé dans les paragraphes 4 à 7 du présent document; et

b) à suggérer que les membres de l'Union proposent à des examinateurs de dénominations dotés de l'expérience adéquate d'intégrer le groupe de travail.

[L'annexe suit]

UNION EUROPÉENNE
Office communautaire des variétés végétales

PROCÉDURE DE RECHERCHE

1. GÉNÉRALITÉS

En conclusion de la phase d'étude du projet présenté au Conseil d'administration en novembre 2003, l'Office a proposé, dans un premier temps, d'acquérir les logiciels nationaux, de les adapter à la base de données centralisée et de les exécuter pour chaque test. Les risques de rater une dénomination similaire seraient ainsi très limités.

À plus long terme, l'élaboration d'un logiciel de l'OCVV était envisagée, avec la possibilité de mettre au point des fonctionnalités linguistiques.

En pratique, les spécifications du logiciel français ont été choisies pour servir de base à la mise en œuvre de la procédure de recherche dans la base de données de l'OCVV.

2. RÈGLES ÉTABLISSANT UNE DISTINCTION SUFFISANTE ENTRE DEUX DÉNOMINATIONS VARIÉTALES

L'une des dispositions du Règlement de base instituant un régime de protection communautaire des obtentions végétales prévoit qu'une dénomination variétale ne doit pas être identique à une dénomination variétale ou pouvoir être confondue avec une dénomination variétale sous laquelle une autre variété de la même espèce ou d'une espèce voisine est enregistrée.

Cette règle a été interprétée dans les principes directeurs du Conseil d'administration de l'OCVV sur les dénominations variétales :

- ♣ Une différence d'une seule lettre ou d'un seul chiffre, ou une différence d'accentuation, devrait généralement être considérée comme pouvant porter à confusion.
- ♣ Des différences de deux lettres ou plus ne devraient généralement pas être considérées comme pouvant porter à confusion, sauf lorsque les mêmes lettres sont simplement juxtaposées.
- ♣ En outre, une dénomination variétale ne devrait pas donner l'impression erronée que la variété est liée à une autre variété ou dérivée d'une autre variété.

Le programme de recherche visera à identifier des dénominations de la même classe dans la base de données qui pourraient être incompatibles avec une dénomination proposée.

3. PROCÉDURE DE RECHERCHE

Les tests sont réalisés par un programme interne de la base de données ORACLE sur le serveur de l'OCVV (meilleurs résultats).

Pour réaliser un test, le programme d'interface (site Web, etc.) exécute une procédure appelée : TESTDENOMINATION

La dénomination visée par le test et le code de l'espèce à laquelle appartient la variété sont entrés comme paramètres par le programme d'interface.

La procédure renvoie l'identifiant du test réalisé vers l'interface (colonne Testid). Cet identifiant permet de visualiser les deux tableaux qui constituent les résultats du test : TESTS et TESTRESULTS.

Le tableau TESTS contient les données générales relatives au test : date du test, identifiant de la personne qui a demandé le test, dénomination, code de l'espèce, code de la classe ou du genre, temps d'exécution de l'ordinateur, mots exclus, message d'erreur, etc.

Le tableau TESTRESULTS contient les listes des dénominations qui ont été considérées comme similaires par la procédure TESTDENOMINATION. Chaque dénomination similaire est associée à un indice de similarité.

4. Description de la procédure TESTDENOMINATION

Contrôle d'entrée

Le code de l'espèce doit figurer dans la base de données des dénominations.

La dénomination ne doit pas compter plus de 100 caractères (espaces inclus).

Caractères exclus : signes de ponctuation et /-_ '. Les caractères accentués ne sont pas autorisés.

Les dénominations composées de plusieurs mots sont autorisées, dans la limite de quatre mots. Les mots doivent être séparés par un espace.

a) Première opération : division de la dénomination de l'entrée en mots distincts

La dénomination visée par le test est divisée en mots élémentaires. Les espaces, considérés comme des séparateurs, sont supprimés.

Les caractères standards non latins (caractères accentués, etc.) sont remplacés par des caractères standards latins. Toutes les lettres sont transformées en lettres capitales. Voir le tableau de conversion.

Exemple : Déjà devient DEJA

Les lettres doubles deviennent des lettres simples.

Exemple : HELLO devient HELO.

Chaque mot élémentaire est confronté à la liste de mots exclus du test (p. ex. color YELLOW, RED, PURPLE, etc.)

Les mots élémentaires figurant dans cette liste sont exclus du test de similarité.

Dans la description ci-après, les mots élémentaires obtenus sont appelés "Mots testés" (*Words Tested* – WT)

Exemple : dénomination "Tānau TARI YELLOW"

Le mot *yellow* est exclu des tests. Les mots pris en compte (WT) sont les suivants :

- TANAU,
- TARI,
- TANAUTARI,
- TARIYELLOW,
- TANAUTARIYELLOW.

b) Seconde opération : établissement d'une liste de mots de référence dans la classe

Le programme recherche la classe ou le genre associé à l'espèce pour déterminer la portée de la recherche de similitude.

Pour les dénominations enregistrées en tant que codes, une chaîne unique de caractères est prise en compte, sans espace.

Le logiciel établit une liste de tous les mots élémentaires appartenant aux dénominations des variétés de la classe.

Tous les mots comptant plus de 3 caractères que le WT le plus long et moins de 3 caractères que le WT le plus court ne sont pas pris en compte.

Tous les mots de référence appartenant à la liste susmentionnée de mots exclus des tests (= color YELLOW, RED, PURPLE, etc.) sont exclus du test de similitude.

Comme précédemment, les lettres doubles des mots de référence sont transformées en lettres simples.

Dans la description ci-après, les mots élémentaires inclus dans cette liste sont appelés "Mots de référence" (*Words of reference* – WR).

c) Les principes du test de similitude

Pour chaque WT, la procédure SIMILARITYTEST calcule un indice de similitude par rapport à chaque WR inclus dans la liste de dénominations précédemment établie.

Les WR sont classés en fonction de la valeur de l'indice de similitude. Tous les WR dont l'indice de similitude est supérieur à un seuil prédéfini sont exclus des résultats.

Exemple :

Dénomination testée "Tānau TARI YELLOW"

La liste de mots de référence contient 10 000 mots élémentaires.

TANAU, TARI, TANAUTARI, TARIYELLOW et TANAUTARIYELLOW sont testés par rapport à ces 10 000 mots de référence. Cinquante mille tests sont réalisés (5 mots x 10 000 WR).

5. Description détaillée du test de similitude

Les étapes ci-après sont mises en œuvre pour chaque paire (WT, WR).

Filtre préliminaire : dans la liste de WR, tous les mots comptant plus de 3 caractères que le WT et moins de 3 caractères que le WR ne sont pas pris en compte pour les calculs suivants.

Première étape : calcul de Ki2

Formule : $Ki2 = \text{somme}(di)^2 / (\text{longueur}(WT)-1)(\text{longueur}(WR)-1)$

Où di = différence de nombre de lettres entre le mot testé et le mot de référence. Toutes les lettres des deux mots sont prises en compte.

Exemple :

WT : **ALADIN** – 6 caractères.

Si l'on compare ce WT avec la chaîne de caractères existante : **DYLAN** (5 caractères)

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

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

$$\text{Chi}2 = (1 + 0 + 0 + 1 + 0 + 1)(5*4)$$

$$\text{Chi}2 = 3/20$$

$$\text{Chi}2 = 0,15$$

On conserve donc pour le calcul suivant tous les mots pour lesquels :

- $\text{Ki}2 \leq 0,3$ et longueur du WT ≥ 5 caractères
- $\text{Ki}2 \leq 0,4$ et longueur du WT = 4 caractères
- $\text{Ki}2 \leq 0,5$ et longueur du WT < 4 caractères

Deuxième étape : quatre calculs basés sur la sélection de la première étape

Au cours de cette deuxième étape, quatre calculs sont réalisés :

- Calcul du pourcentage de lettres communes
- Calcul du pourcentage de lettres NON communes
- Calcul du pourcentage de différence de longueur
- Calcul du coefficient de corrélation des rangs de Kendall

Calcul du pourcentage de lettres communes (*common letter* = CL)

$$\text{CL} = 1 - (\text{nombre de lettres communes}) / (\text{longueur (WT)})$$

Exemple : ALADIN et DYLAN

$$\text{CL} = 1 - 4/6 = 0,33$$

CL est égal à 0 lorsque toutes les lettres figurent dans le WR.

Second exemple : BANANAS et BANS

Toutes les lettres de BANANAS figurent dans le mot BANS.

$$\text{CL} = 1 - 7/7 \Rightarrow \text{CL} = 0$$

Calcul du pourcentage de lettres NON communes (*non-common letter* = NCL)

$$\text{NCL} = (\text{nombre de lettres du WR absentes dans le WT}) / (\text{longueur (WT)})$$

Exemple : ALADIN et DYLAN

$$\text{NCL} = 1/6 = 0,16$$

NCL est égal à 0 lorsque toutes les lettres du WR figurent dans le WT.

Second exemple : BANANAS et BANS

Toutes les lettres de BANS figurent dans le mot BANANAS.

$$\text{NCL} = 0/7 \Rightarrow \text{NCL} = 0$$

Calcul du pourcentage de différence de longueur (*difference of length = DL*)

$DL = (\text{différence de longueur entre les 2 chaînes}) / (\text{longueur du WT})$

Exemple : BANANAS et ANANAS

$DL = 1/7$

DL est égal à 0 lorsque les deux mots sont de la même longueur.

Calcul du coefficient de corrélation des rangs de Kendall (*Kendall correlation = KC*)

Formule : $KC = 6 \cdot \text{somme}(Di)^2 / N \cdot (N^2 - 1)$

Où :

- Di est égal à la différence de position des lettres communes (Li) moins la différence de position des lettres précédentes. Si le WT contient plusieurs occurrences de la même lettre, les premières lettres sont utilisées comme référence pour la position. Si le WR contient plusieurs occurrences de la même lettre, on utilise les lettres les plus proches de la même lettre dans le WT.
- N est égal au nombre de lettres communes entre le WT et le WR

Exemple :

WT : **ALADIN** et **DYLAN**

	A	L	D	N
ALADIN	1	2	4	6
DYLAN	4	3	1	5
Différence	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 lettres en commun A, L, D et N

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

$$KC = 6 \cdot (9 + 16 + 16 + 4) / 4 \cdot 15$$

$$KC = 6 \cdot 45 / 60$$

$$KC = 4,5$$

WT : **ALADIN** et **BALADIN**

	A	L	D	I	N
ALADIN	1	2	4	5	6
BALADIN	2	3	5	6	7
Différence	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 lettres en commun A, L, D, I et N

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

$$KC = 6 \cdot (1) / 5 \cdot 24$$

$$KC = 1 / 20$$

$$KC = 0,05$$

Dans cet exemple, on peut voir que si la même séquence de lettres figure dans les deux mots (ici LADIN), l'écart entre les deux séquences n'est pris en compte qu'une seule fois.

Si les deux mots WT et WR sont identiques, le coefficient de Kendall est égal à 0.

Troisième étape : calcul de l'indice de similitude

Indice de similitude = KC + CL + NCL + DL

Un mot de référence est **sélectionné** si les scores sont inférieurs ou égaux à :

	KC	CL	DL	Indice de similitude
Longueur du mot recherché > 4 caractères	<= 1,5	<= 0,22	<= 1,5	< 1,2
Longueur du mot recherché = 4 caractères	<= 1,5	<= 0,25	<= 1,26	< 1,2
Longueur du mot recherché < 4 caractères	<= 1	<= 0,34	<= 1,0	< 1,2

Les dénominations de référence sont triées en fonction de l'indice de similitude, puis dans l'ordre alphabétique.

[Fin de l'annexe et du document]