# Rechercher et remplacer à l'aide de «RegEx» (1)

Hans Häsler, Lausanne

**Avec InDesign CS3, on peut enfin utiliser des «regular expressions» lors d'une action de rechercher/remplacer. Cela nécessitait l'intégration du logiciel utilitaire «grep». Il est à présent possible d'écrire des chaînes de textes composées de signes spéciaux, qui permettent de rechercher et de remplacer des bouts de textes d'une façon très souple.**

Les développeurs de scripts sont déjà familiers avec cette nouveauté. Mais même un scripteur très doué doit investir pas mal de temps pour apprendre les subtilités qui sont propres à chaque implémentation.

Ainsi les références aux groupes définies. Celles-ci doivent être formatées \\1 en utilisant *Satimage.osax*, tandis qu'avec *Adobe InDesign* il faut insérer **$1** dans le champ «Remplacer par». Mais parlons-en plus tard.

### Au fait, il n'y a rien de nouveau...

L'utilisation de caractères joker (ou *wildcards*) dans une chaîne de recherche est intégrée dans QuarkXPress depuis de nombreuses années (fig. 1).

Adobe a aussi introduit cette option, tout en l'élargissant. Avec InDesign CS2, on peut insérer des métacaractères. Cela signifie que ces signes ne sont pas utilisés littéralement mais qu'ils ont une signification spéciale.

Ces *wildcards* se distinguent par un caractère caret (= **^**). Le caractère «nouvelle ligne» est représenté par **^n**, une ellipse par **^e**, des caractères quelconques par **^$** et des chiffres quelconques par **^9** (fig. 2).

### ... mais c'est beaucoup plus puissant

Le fait que le nombre des caractères joker doit correspondre exactement à celui du texte à trouver rend leur utilisation moins souple. Lorsque les valeurs des jours et des mois n'ont qu'une seule position, il faut établir plusieurs chaînes de rechercher et de remplacer.

### Une nouveauté bienvenue

Il est donc très pratique que les expressions rationnelles soient intégrées dans InDesign CS3. Ainsi, on peut utiliser le dialogue natif. De plus, les formatages locaux ne sont pas perdus, comme c'était le cas lors de l'application directe d'un script. Avec InDesign CS2, on était obligé d'exporter le texte dans un fichier au format balisé. Mais cela comportait pas mal de risques.

**«grep»? Qu'est-ce que c'est?**

Un programme créé dans les années septante, pour les systèmes d'exploitation Unix. Utilisé pour la recherche et le filtrage de chaînes de caractères.

Selon certaines sources, ce terme est un raccourci de *global regular expression print*.

La base historique, dit-on, serait une commande de l'éditeur Unix «ed»: g/RE/p.

**Qu'est-ce qu'une «regular expression»?**

Une chaîne de caractères, formée de signes génériques (aux significations spécifiques) et de caractères littéraux. En bref, un motif utilisé pour rechercher dans des textes des chaînes de lettres distinctes, afin de les modifier.

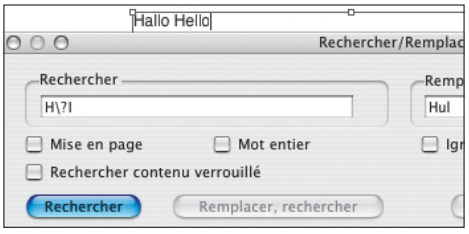


**Fig. 1 – Une vieille option QuarkXPress. Le caractère joker «\?» représente une lettre quelconque. Les deux mots «Hallo» et «Hello» sont changés à «Hullo».**

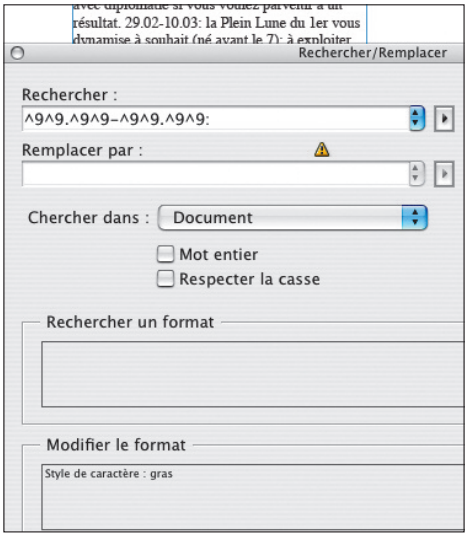


**Fig. 2 – InDesign CS2. Le joker «^9» représente un chiffre quelconque. Combiné avec un point, un trait d'union et un deux-points, des chaînes de dates peuvent être trouvées et formatées.**

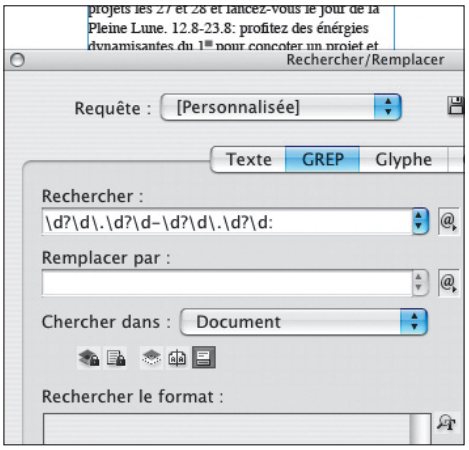


**Fig. 3 – InDesign CS3. Le joker «\d» représente un chiffre quelconque. Le point d'interrogation signifie que la position à sa gauche ne doit pas être présente obligatoirement. Le point est un signe très spécial (il trouve un caractère quelconque). Il faut donc annuler ce statut en plaçant devant un *backslash* (une barre oblique inverse), quand il s'agit de trouver un point «réel».**

### Le dialogue rechercher/remplacer étendu

Il ne faut pas avoir peur: les gens qui n'ont pas envie de toucher à cette nouveauté peuvent utiliser le dialogue comme d'habitude. Le menu local est réglé à l'article «Texte». Il faut sélectionner «GREP» quand on veut profiter des possibilités élargies.

Afin de pouvoir comprendre ces chaînes de caractères cryptiques et d'être en mesure de les taper sans trop d'hésitations, il est conseillé d'acquérir une base solide. Il faudrait aller visiter quelques sites internet qui relatent cette matière complexe.

Avant de détailler quelques particularités des expressions rationnelles, expliquons comment on peut mettre en évidence les chaînes formant les décans dans un texte d'un horoscope (fig. 3).

### Un rechercher/remplacer beaucoup plus performant

Le format abrégé des dates comprend des valeurs à une et à deux positions pour les quantièmes et les numéros des mois.

L'extrait **\d?\d\.** de la chaîne de recherche doit être lu de droite à gauche. D'abord un point (= \.) précédé d'un chiffre quelconque (= **\d**). Puis un point d'interrogation, signifiant que le caractère à sa gauche (encore un chiffre) peut être présent, mais ne doit pas obligatoirement exister. En multipliant ces signes et en insérant un trait d'union et un deux-points à la fin, on obtient une chaîne de recherche polyvalente.

Afin de mieux comprendre les significations, le motif de recherche est expliqué ci-après, signe par signe.

| | |
|---|---|
| **\d?\d** | un seul ou deux chiffres |
| **\.** | un point |
| **\d?\d** | un seul ou deux chiffres |
| **-** | un trait d'union |
| **\d?\d** | un seul ou deux chiffres |
| **\.** | un point |
| **\d?\d** | un seul ou deux chiffres |
| **:** | un deux-points |

Pour les premiers pas, il est conseillé d'utiliser les pièces détachées offertes par InDesign dans son dialogue rechercher/remplacer. Un signe «@» muni d'une flèche se trouve à droite du champ «Rechercher».

En dirigeant le pointeur de la souris sur ce bouton, le commentaire «Caractères spéciaux pour la recherche» devient visible. Un clic de souris ouvre le menu.

Les trois premiers articles n'ont rien d'extraordinaire, mais ils sont faciles à mémoriser (tabulation **\t**, saut de ligne forcé **\n**, fin de paragraphe **\r**). Mais les habitués de l'ancienne méthode (**^t**, **^n**, **^p**) seront obligés de s'adapter.

La section suivante, les «Symboles», a été élargie fortement (fig. 5). Lors des premiers essais, on remarque que le caractère caret habituel a été remplacé par un tilde. La raison: le premier a une signification spéciale dans un environnement RegEx. Il marque le début d'une ligne. D'autres signes sont munis d'un *backslash* (= \ = une barre oblique inverse).

Les articles suivants (comme «Marques», «Césures et tirets», «Espace») contiennent tous les métacaractères familiers, mais également avec un tilde à la place du caret.

Mais ce n'est qu'à partir de la section «Caractères génériques» que cela devient intéressant (fig. 6). Ces signes permettent d'écrire des motifs très souples.

Le «Chiffre quelconque» (= **\d**, comme *digit*, angl. pour chiffre) a déjà été utilisé dans la figure 3. Avant de nous occuper de ce chapitre, il faut expliquer quelques signes spéciaux. Trois d'entre eux (= **+ ? ***) se trouvent également sous «Répétition».

## Encore des signes spéciaux

Le changement mentionné du caractère de commande indique qu'avec RegEx il existe bien d'autres caractères dont la signification est différente de leur valeur littérale.

```
[ ] ( ) { } + ? * . | \ ^ $
```

**Fig. 4 – Des caractères avec un statut spécial.**

### *Les crochets*

Ceux-ci sont utilisées pour contenir des chaînes de caractères: **[a-z]**, par exemple, trouve un seul signe contenu dans la chaîne de «a» à «z». Donc une lettre minuscule quelconque, y compris tous les caractères accentués.

D'autres environnements «grep» sont plus stricts. Ils renvoient uniquement un caractère de la chaîne ininterrompue des 26 signes (de ASCII 97 à 122).

Essayons-le: taper dans un bloc texte les signes «123é456» (sans les guillemets). Entrer **\d[a-z]\d** dans le champ «Rechercher», la chaîne qui devrait trouver un chiffre, un caractère minuscule et encore un chiffre. Cliquer sur le bouton «Rechercher» et les signes «3é4» sont sélectionnés.

Mais que faire quand on aimerait trouver plusieurs caractères? Dans ce cas, il faut insérer un signe «+» derrière la parenthèse fermante: **\d[a-z]+\d**. Ainsi, un chiffre, au moins un caractère minuscule (mais également plusieurs) et encore un chiffre seront trouvés. Oui: dans la chaîne «123abc456» les cinq signes «3abc4» sont sélectionnés.

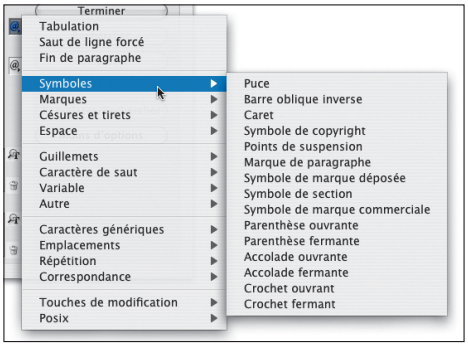


**Fig. 5 – Le menu «Symboles» étendu.**

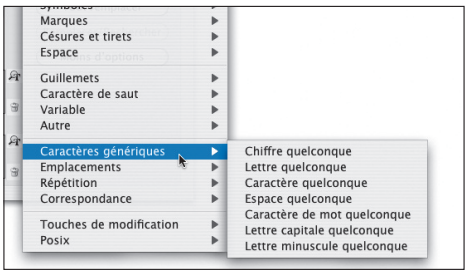


**Fig. 6 – «Caractères génériques»: de nouvelles possibilités pour le rechercher/remplacer dans InDesign.**

On peut aussi mettre plusieurs rangées ou des caractères isolés dans ces parenthèses. Le motif **[ -9;-ü]** s'occupe d'une première rangée, commençant par une espace (ASCII 32) et se terminant par le «9» (ASCII 57), puis d'une seconde, du point-virgule (ASCII 59) jusqu'au «ü» (ASCII 159). Ainsi, le deux-points (ASCII 58) est exclu de la recherche parce qu'il ne devrait pas être trouvé, pour une raison particulière.

### *Les parenthèses*

Celles-ci servent à former des groupes de chaînes. Le but n'est pas de rendre le motif plus lisible, mais d'obtenir des références (angl. *backreference*) aux groupes que l'on peut placer dans la chaîne de remplacement. Comment s'y prendre? Et, surtout, à quoi cela sert-il?

Sélectionnons le bloc texte qui contient les neuf caractères «123abc456». On peut, par exemple, taper **(\d+)([a-z]+)(\d+)** dans le champ de recherche et **$1 $3** dans le champ de remplacement.

Le motif rechercher contient donc trois groupes. Quand une correspondance est trouvée dans le texte, celle-ci est mémorisée dans trois variables temporaires. A l'aide du signe «$» suivi d'un chiffre, on établit une référence au groupe correspondant. Après l'exécution du rechercher/remplacer, le bloc texte ne contient plus que «123 456». Le deuxième groupe étant représenté par une espace, celle-ci est donc insérée à la place des trois caractères minuscules.

Une autre possibilité: l'ordre des éléments est inversé avec **$3$2$1**. Le résultat: «456abc123».

### *Les accolades*

Elles sont utilisées pour entourer des valeurs, indiquant le nombre de fois où les signes recherchés peuvent être présents. D'habitude, on utilise une valeur minimale et une valeur maximale. Mais il est possible d'écrire des combinaisons:

**{2}** = exactement deux positions
**{,3}** = aucune ou jusqu'à trois positions
**{3,6}** = au moins trois positions, mais au maximum six
**{4,}** = au moins quatre positions

Quelques exemples. Quand il n'y a qu'une valeur, les chaînes à trouver doivent correspondre exactement à la valeur définie: **[0-9]{6}** trouve des nombres à six positions. Mais il faut s'assurer que les chiffres peuvent être délimités, par exemple par des points, comme dans la figure 3.

Si cela n'est pas possible, il est impératif d'entourer la chaîne du métacaractère pour les limites des mots: **\b[0-9]{6}\b**. Ainsi, on évite que six chiffres soient extraits de nombres comprenant sept positions ou davantage.

**[0-9]{1,2}** signifie que le nombre doit consister d'au moins une position, mais au maximum de deux. Cela fonctionne également avec l'autre représentation d'un chiffre entre 0 et 9. Le motif de la figure 3 peut aussi être écrit ainsi:
**\d{1,2}\.\d{1,2}-\d{1,2}\.\d{1,2}:**

### *Le signe plus*

Il sert à déterminer que le signe précédent (ou la rangée) peut être présent au moins une ou plusieurs fois.

Un exemple très simple: rechercher **\s+\s**, remplacer par **\s**. Le résultat: des espaces multiples dans le texte sont remplacées par une seule.

On peut également insérer une espace normale, tandis qu'avec le **\s** (= toutes sortes d'espaces) des quarts de cadratins, etc., seront trouvés.

### *Le point d'interrogation*

Celui-ci est mis quand un signe recherché n'est pas forcément présent, ou alors qu'une seule fois.

Essayons-le: remplir un bloc texte avec du texte fictif, ajouter des deux-points à quelques mots, insérer dans deux ou trois cas une espace devant le deux-points. La chaîne à rechercher **([\l\u])( )?(:)** et celle du remplacement: **$1~|$3**. Le résultat: une espace fine est insérée entre les groupes un et trois, peu importe s'il y avait une espace.

### *L'astérisque*

Le rôle de ce signe est similaire à celui du point d'interrogation. Le signe précédant ne doit pas être présent, mais il peut s'y trouver plusieurs fois.

Un exemple. Taper «test(abc) test (abc) test» dans un bloc de texte. Entrer **\(.*\)** dans le champ rechercher du dialogue. Le

point trouve un signe quelconque, l'astérisque indique que ce signe peut être présent plusieurs fois ou pas du tout. Le pouvoir spécial des parenthèses est annulé par les *backslashes*. Ce motif trouve donc du texte délimité par des parenthèses et également ces dernières.

Quelle surprise: lors du clic sur le bouton «Rechercher» ce n'est pas la première occurrence qui est trouvée, mais «(abc) test (abc)».

L'explication: l'astérisque est un goinfre (angl. *greedy*). Il bouffe la totalité du texte se trouvant entre la première occurrence du début et de la dernière de la fin du motif. Ce résultat ne correspond que rarement aux intentions. La solution: insérer un point d'interrogation derrière l'astérisque. Ainsi, la machine RegEx est forcée de revenir en arrière, jusqu'à ce que les caractères trouvés correspondent au motif initial.

En modifiant le motif ainsi: **\(.*?\)**, la sélection est réduite à «(abc)», comme souhaité dès le départ.

D'ailleurs, le signe «+» est également *greedy*. Il suffit de le mettre à la place de l'astérisque dans l'exemple précédent pour constater le même comportement.

### *Le point*

Comme déjà expliqué, le point trouve des caractères quelconques, à l'exception de retour, shift-retour, colonne suivante, bloc suivant.

C'est bon à savoir: InDesign commence la recherche à partir de la position du curseur et sélectionne tout jusqu'à ce qu'il trouve une exception. Quand le curseur clignote à la fin du texte, la recherche commencera au début.

### *Le symbole pipe*

Ce trait vertical (= **|**) subdivise la chaîne de recherche. Chaque part peut trouver une correspondance lors du même passage. Avec un motif comme **abc|def|ghi** chacune des trois portions peut être trouvée et remplacée individuellement.

### *Le backslash*

Cette barre oblique inverse (= \) a déjà été mentionnée plusieurs fois. Elle est toujours insérée pour annuler la signification spéciale des métacaractères (fig. 4), pour ceux-ci puissent être traités littéralement.

Une précision: trois parmi les quatorze signes ne sont pas vraiment des métacaractères. Le crochet fermant (= **]**) et les accolades (= **{ }**) ne passent dans un statut spécial que dans un certain contexte.

Mais que faire quand il faudrait rechercher un *backslash* littéralement, puisque celui-ci fait partie des signes spéciaux? C'est simple: il faut annuler son statut exceptionnel en insérant, devant, un... *backslash*. Si l'on ne le fait pas, il ne se passe rien (au mieux). Mais il est probable que des bouts de texte soient modifiés involontairement.

### *Le caractère caret*

Celui-ci (= **^**) est connu spécialement de la recherche traditionnelle, comme faisant partie des métacaractères.

Il est utilisé avec RegEx pour marquer le début des paragraphes. Un exemple aide à mieux comprendre sa fonction. Taper dans un bloc texte «1. texte 2. texte¶3. texte», en saisissant un vrai retour à la place du symbole «fin de paragraphe». La chaîne de recherche **^\d\.** trouve le premier chiffre et le point qui le suit. Après le clic sur «Suivant» ce n'est pas le «2.» qui est sélectionné mais le «3.», puisque ce n'est que celui-ci qui se trouve au début d'un paragraphe.

Important: si le texte à modifier est toujours placé au début des paragraphes, il vaut mieux utiliser ce caractère caret. L'exécution sera accélérée, puisque la machine RegEx ne doit pas examiner la totalité d'un paragraphe si la chaîne n'est pas trouvée au début.

Le caractère caret est doté d'une seconde signification. Placé à l'intérieur des crochets, il exerce une négation. Un exemple: **[^a-z]+** trouve des chaînes qui ne sont pas constituées de caractères minuscules. Le début du paragraphe n'est plus considéré.

### *Le signe dollar*

Comme dans d'autres environnements «grep», le signe dollar signifie le contraire du caractère caret. Il marque la fin des paragraphes. Un exemple pratique: **\s+$** supprime des espaces inutiles devant les retours.

Une autre mise en œuvre: comme déjà démontré, un chiffre (de 1 à 9) placé derrière effectue une référence à un groupe (désigné par des parenthèses) de la chaîne de recherche. Quand il n'y a pas de groupe, on peut toutefois utiliser **$0** pour reprendre le texte trouvé. C'est pratique quand il faut insérer quelque chose devant ou derrière la chaîne. Ce raccourci peut être placé plusieurs fois dans la chaîne de remplacement. Le texte sera répété en conséquence.

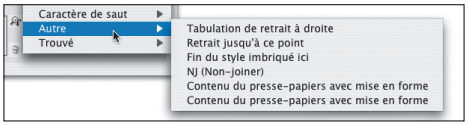


**Fig. 7 – Le contenu du presse-papiers peut être inséré dans le motif de remplacement. Mais, contrairement aux apparences (deux articles de menu identiques), en utilisant le dernier on obtient le métacaractère ~C et le contenu est inséré *sans* mise en forme...**


**Fig. 8 – Un dialogue simple s'ouvre après le clic sur le symbole disquette. Il permet de nommer et de sauvegarder les motifs rechercher/remplacer. Les noms des fichiers apparaissent dans le menu «Requête».**

```
<?xml version="1.0" encoding="UTF-8"
standalone="yes"?>
<Query>
<Header>
<Version major="5" minor="1">
</Version>
<Application value="Adobe InDesign">
</Application>
<QueryType value="Grep" qid="1">
</QueryType>
</Header>
<Description>
<FindExpression value="\d?\d\.\d?\d-\d?\d\.\d?\d:">
</FindExpression>
```

**Fig. 9 – Un petit extrait du fichier XML contenant le motif de la figure 3.**

### Les caractères génériques

Retournons, enfin, dans le menu des signes spéciaux. Le «chiffre quelconque» (= **\d**) a déjà été mentionné. Mais la «lettre quelconque» (= **[\l\u]**) est nouvelle. Entouré de crochets, un «l» (= *lowercase*, angl. pour caractères minuscules) et un «u» (= *uppercase*, angl. pour caractères majuscules).

Le *backslash* qui les précède les désigne comme métacaractères. Ce motif trouve une seule lettre qui peut être soit minuscule, soit majuscule. La bonne nouvelle: les lettres accentuées sont aussi de la partie.

Faisons un test. Taper dans un bloc texte les caractères «123à456». Entrer **\d[\l\u]\d** dans le champ de recherche et cliquer sur «Rechercher». En effet: les signes «3à4» sont sélectionnés.

### Le remplacer par...

Jusque-là, on s'occupait le plus souvent de la recherche de texte. Mais les articles offerts par InDesign pour le remplacement correspondent assez largement à ceux de la recherche. Le menu est un peu plus court.

Une bonne nouveauté: dans l'article «Autre» (fig. 7), on découvre qu'il est possible d'inclure le contenu du presse-papiers dans la chaîne de remplacement.

### Une sauvegarde pratique

InDesign CS3 offre une autre nouveauté bienvenue: les motifs de la recherche et du remplacer sont non seulement mémorisés dans le dialogue, mais on peut les enregistrer dans des fichiers sur le disque dur.

Ceux-ci se trouvent dans le dossier des préférences de l'utilisateur qui les a créés: «Adobe InDesign/Version 5.0/Find-Change Queries/GREP». Ces fichiers sont éditables. Pour l'utilisation, ils apparaissent comme articles de menu dans le dialogue.

### Ce n'est pas tout...

Non, parce qu'il était impossible de tout mentionner. Il reste donc de la matière pour un autre article.

Quelques mots-clé: Lookbehind positif, Lookbehind négatif, Lookahead positif...

### De la lecture

On l'a déjà dit: il faudrait se renseigner à fond sur les possibilités des *Regular Expressions*. Pour un bon départ, lancer «Google», entrer RegEx, et on obtient plein de liens valables. En anglais, en allemand, mais presque rien en français...