Une mise en concordance automatique améliorée pour le recensement ?

Retour d’expérience sur l’expérimentation d’une nouvelle recette
La MCA++

Directement aux conclusions

- On peut imaginer un processus automatique meilleur que la MCARP standard
  - MCARP : 40 à 55 % selon les départements
  - MCA++ : 75 à 80 % selon les départements
- Un processus automatique peut résoudre des cas qui ne sont pas maîtrisables manuellement avec l'outillage standard du RP (RECAP)
- Il restera toujours des cas indécidables, même après expertise manuelle approfondie. En gros il y a trois cas, avec une forte marche à franchir pour sortir du premier et dépasser les 50 % de réussite
  - Trivial
  - Compliqué
  - Indécidable
- On peut toujours trouver un écho, mais encore faut-il s'assurer que c'est le bon écho ! ...dans un contexte où on n'a pas la bonne solution.

Les principes de la MCA++

- La finalité ultime est de pouvoir inférer un code NAF, le SIRET n'est qu'un intermédiaire.
- Lorsque cela se passe bien, les noms d'établissement fournis sont juste une présentation différente de ceux connus dans le répertoire
  - Plus riche ou plus pauvre
  - Pas forcément dans le même ordre

  ⇒ Ce sont eux qu'on va utiliser en majeur
- Les informations sur le lieu de travail ou l'activité peuvent au mieux être des informations annexes
  - Lieu de travail vs. Établissement de rattachement
  - Adresses multiples en milieu épars, partagées en milieu dense
  - Adresses mal remplies
  - Même très agrégée l'activité n'est pas codable à partir des informations du BI (son codage est d'ailleurs le but de l'opération!)

Reconnaître un nom d’établissement

- On est dans un contexte voisin de celui de la comparaison de deux chaînes de caractères
- Sauf que les libellés ont une signification : l’information élémentaire est le mot, pas le caractère

  ⇒ Détournement des algorithmes classiques (Levenshtein)
- et tous les mots n’ont pas la même valeur

  ⇒ Base de connaissances (variées) annexe

Architecture générale

Phase 0) Mise en forme de SIRENE, du RP

Phase 1) Comparaisons des libellés : repérage des dissimilitudes

Phase 2) Quantification des similitudes

Phase 3) Choix du meilleur écho

Mise en forme de SIRENE (1)

- Duplication sur plusieurs lignes selon les libellés possibles : enseigne, sigle, raison sociale
- Eclatement en mots
  - Fusion en un seul mot des lettres isolées ou séparées par des points
  - Suppression des articles
  - Expansion des abréviations (ECOL, MARIT...)
  - Compression de SAINT et SAINTE
- Pas d'élimination de mots « vides » à part les articles

Mise en forme de SIRENE (2)

- Remplacement des caractères spéciaux
- Réorganisation des noms propres, avec marquage différentié des mots des prénoms et des noms
- Repérage des sigles présents en début ou fin de libellé
  - catégorie juridique (on supprime)
  - recopies parfois tronquées de la raison sociale (on marque)
    - Exemple : DIFFUSION VENTE DISTRIBUTION DVD

Mise en forme du RP

- Modification MAIRIE -> COMMUNE, ajout du nom de la commune s'il n'est pas déjà présent
- Récupération des LYCEE, COLLEGE, ECOLE dans le libellé d'adresse
- Eclatements en mots comme SIRENE
- Codage sûr (pas SICORE!) de la NAF2

Phase 1
La comparaison

- Les *n* libellés du répertoire sont confrontés aux *m* libellés du recensement, donnant *n*x*m* résultats dont ne sont conservés que ceux ayant un niveau de similitude minimum.
- L'objectif de la comparaison n'est pas de calculer une distance, mais de repérer pourquoi les libellés diffèrent.

Comparer deux chaînes
L'algorithme de Levenshtein

- Mesure de la dissimilarité de deux chaînes de caractères, en comptant les modifications élémentaires
  - Suppression
  - Insertion
  - Substitution
- Donne à la fois une mesure et une suite de modifications
- Applications en génétique et pas restreint à la comparaison de chaînes de caractères dès lors qu'on sait faire un test d'égalité.

Déroulement de l'algorithme

		B					
		C	H	I	E	N	S
A	0	1	2	3	4	5	6
N	1	0	0	0	0	0	0
I	2	0	0	0	0	0	0
C	3	0	0	0	0	0	0
H	4	0	0	0	0	0	0
E	5	0	0	0	0	0	0

Initialisation

Puis on parcourt successivement chaque ligne en mettant dans case(i,j), le minimum entre :

- case(i-1,j) + coût de suppression (1)
- case(i,j-1) + coût d'insertion (1)
- case(i-1,j-1) + coût de substitution(1 si A[i]<>B[j])

		C	H	I	E	N	S
	0	1	2	3	4	5	6
N	1	1	2	3	4	4	5
I	2	2	2	2	3	4	5
C	3	2	3	3	3	4	5
H	4	3	2	3	4	4	5
E	5	4	3	3	3	4	5

suppression

insertion

substitution ou identité

Distance

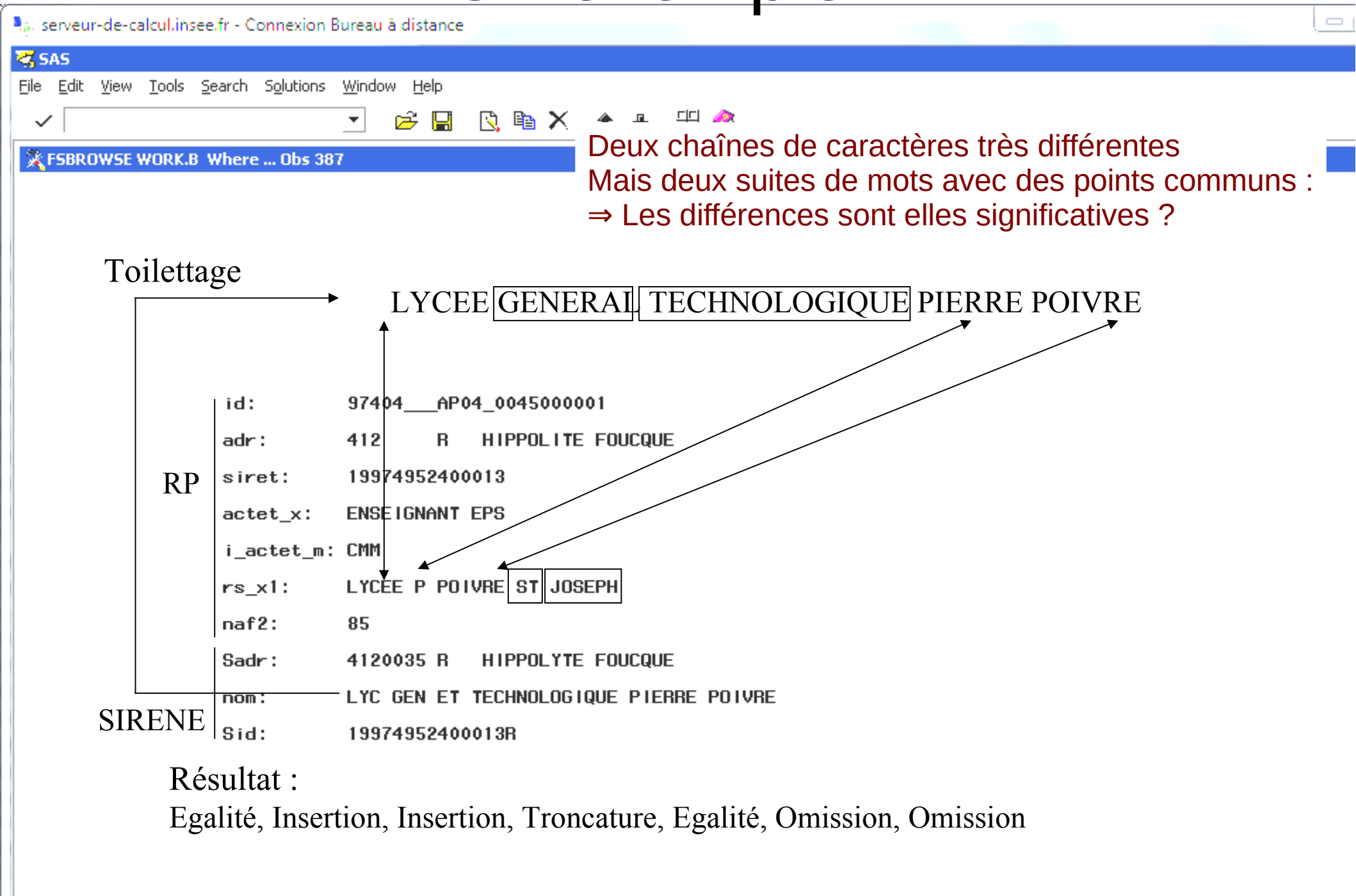
Un chemin de retour de pente selon la plus forte donne une suite d'opérations possible : --==+=++

NIche
CHe
ch**I**e
chi**E**
chie**NS**

Généralisations

- Dans l’exemple précédent, tous les coûts étaient à 1, ce n’est pas une fatalité :
  - Hirsberg : suppression 2, insertion 2, substitution 1
- On peut inclure d’autres opérations, elles ne font que compliquer la construction d’un chemin si c’est ce qu’on cherche :
  - mémoriser l’origine du minimum pour chaque case
  - mémoriser l’impact sur chaque chaîne
- Dameriau : permutations
- MCA++ : des mots plutôt que des caractères et des opérations élémentaires qui peuvent inclure plusieurs mots consécutifs, coûts différenciés
  - Égalités approchées (distance de Levenshtein entre deux mots!)
  - Racines communes, abréviations
  - Synonymes 1 =1, 1 = 2, 1 = 3, 2 = 2
  - Répétitions
  - Permutations
  - Fusions 1 = n, sigles

Un exemple



Résultat :

Egalité, Insertion, Insertion, Troncature, Egalité, Omission, Omission

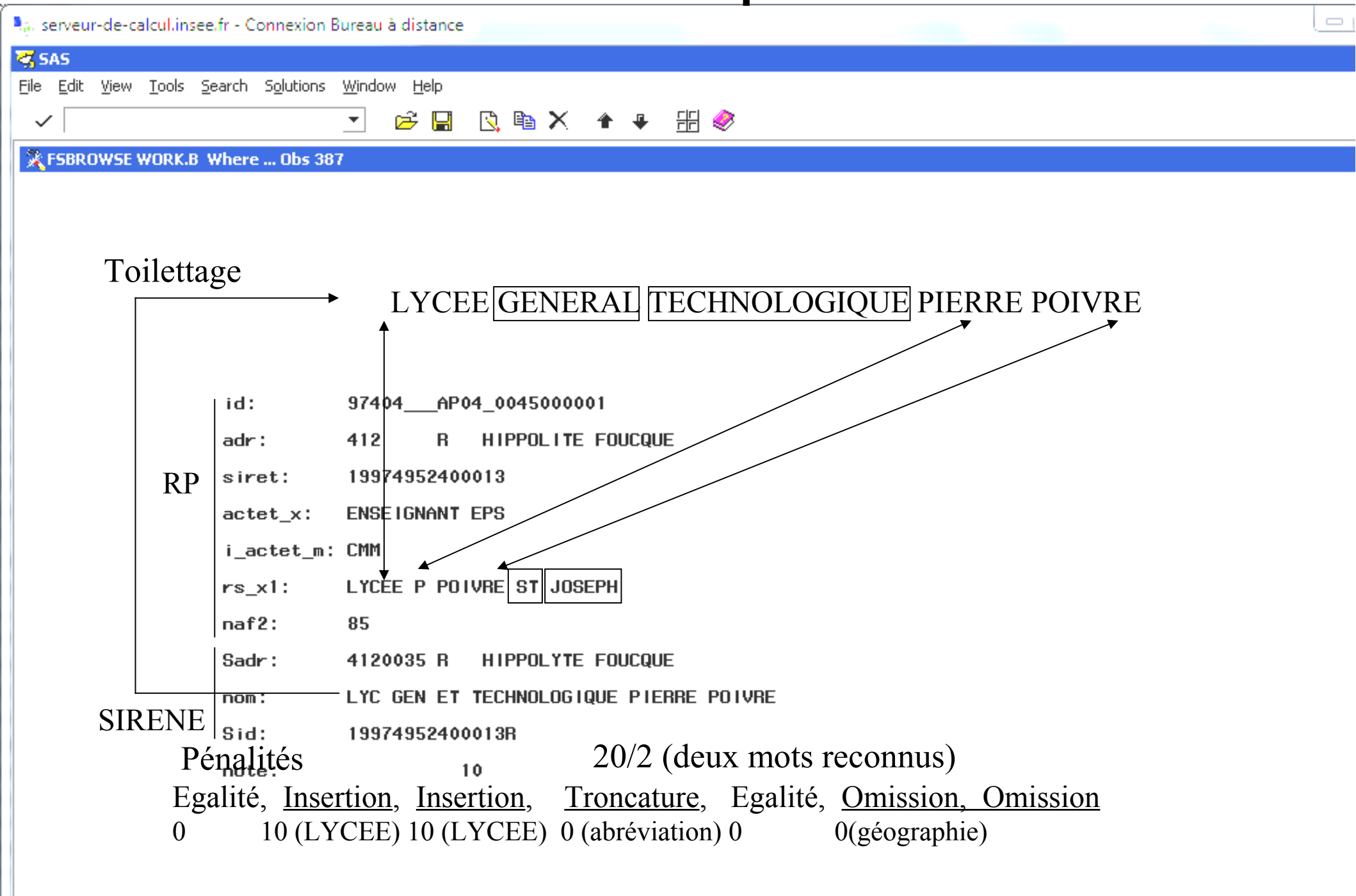
Phase 2 : Interprétation des dissimilitudes (1)

- Repérage des séquences d'insertions ou d'omissions en début et en fin
  - Catégories juridiques
  - ENTREPRISE, SOCIETE, ASSOCIATION…
  - Echelons géographiques redondants avec la localisation
    - SDIS 974
  - Eléments redondant avec l'activité
    - CENTRE HOSPITALIER GABRIEL MARTIN
- Quantification sur le principe de la fonction SPEDIS de SAS, même pénalités pour chacune des transformations, mais division par le nombre de mots communs
  - Insertion: 200 au début, 100 au milieu
  - Omission : 100 au début, 50 au milieu
  - Substitution : maxi 200 (SPEDIS), moitié moins au milieu

Phase 2 : Interprétation des dissimilitudes (2)

- Exceptions à la pénalisation
  - Neutralisation des éclatements (Substitution+Insertion),
  - Neutralisation des fusions (Substitution+Omission)
  - Neutralisation des permutations (Omission + Substitution + Insertion)
  - Neutralisation des pluriels/féminins
- Pénalisations moindres dans quelques cas
  - Omission de
    - GENERAL, TECHNOLOGIQUE… après LYCEE
    - MATERNELLE, CATHOLIQUE… après ECOLE
    - MEDICAL, DENTAIRE… après CABINET
  - Mots de une lettre
  - Prénoms vs. noms propres
- Abandon des mots vides dans le décompte des mots communs

Exemple


serveur-de-calcul.insee.fr - Connexion Bureau à distance
SAS
File Edit View Tools Search Solutions Window Help
FSBROWSE WORK.B Where ... Obs 387
Toilettage
LYCEE GENERAL TECHNOLOGIQUE PIERRE POIVRE
RP
id: 97404___AP04_0045000001
adr: 412 R HIPPOLITE FOUCQUE
siret: 19974952400013
actet_x: ENSEIGNANT EPS
i_actet_m: CMM
rs_x1: LYCEE P POIVRE ST JOSEPH
naf2: 85
SIRENE
Sadr: 4120035 R HIPPOLYTE FOUCQUE
nom: LYC GEN ET TECHNOLOGIQUE PIERRE POIVRE
Sid: 19974952400013R
note: 10
Pénalités
20/2 (deux mots reconnus)
Egalité, Insertion, Insertion, Troncature, Egalité, Omission, Omission
0 10 (LYCEE) 10 (LYCEE) 0 (abréviation) 0 0(géographie)

En pratique : un bonus et un malus

- Différenciés suivant le type d'opération
  - Sigle expansé : bonus = 10, malus = 10
  - Omission, Insertion : bonus = 0, malus = 100 éventuellement minoré selon le contenu
  - Permutation : bonus = 100, malus = 10
  - Abréviation en une lettre : bonus = 10, en plus d'une lettre bonus = 50 (et malus)

  etc..

Valeurs obtenues de façon empirique au vu des suggestions produites

- In fine, il faudra (critère empirique) que **bonus>malus**

Phase 3 : arbitrage
Les critères de tri des échos

- Commune (identique, voisine, différente)
- Malus minimum
- Commune de travail = commune de résidence
- Activité cohérente
- Dernier mot de la voie identique
- Premier caractère du type de la voie identique
- Numéro dans la voie identique
- Etablissement noté comme employeur
- Taille de l'établissement

```
proc sort;
  by id descending eqdc malus descending bonus nelt
  descending eqac descending eqdm descending eqtp descending eqno
  descending empl_et descending eff3112_tr_et;

…
  if first.id ;
```

Des cas à traiter à part

- Dénominations inconnues du répertoire
  - DECATHLON, CARREFOUR

  ⇒ Apprentissage
- Noms d’employeur « génériques »
  - POLICE NATIONALE
  - EDUCATION NATIONALE
  - JUSTICE
  - DEPARTEMENT

  ⇒ Plutôt la localisation en majeur

Les ingrédients utilisés
L'intelligence du processus

- Géographie
  - Noms en clair
  - Indicateur de proximité de communes
- Catégories juridiques (code → sigle)
- Mots caractéristiques de certaines activités
- Prénoms
- Dictionnaire
  - Restreindre les calculs de distance entre mots aux cas où l'un des mots n'est pas connu
  - Interprétation des pluriels
- Synonymes
- Sigles usuels

Quelques exemples

- Mise au point sur La Réunion qui cumule quelques difficultés (présentes aussi sur la métropole)
  - Usage de noms de lieux dits, de grands ensembles
  - Noms de rue en double (ou plus!) sur certaines communes
  - Frontières floues
  - Interférences du créole
  - Forte présence de l’emploi public

L’analyse des résultats
Divergences avec l’automatique

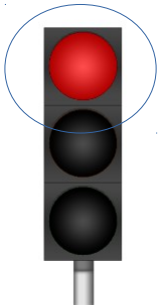
- Surtout des homonymes à la même adresse (groupes?)
- Des erreurs MCARP
- Des codages MCARP appuyés sur l’adresse avec des libellés reconnus approximativement

⇒ Net avantage codage MCA++

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help



= Pas bon pour la MCARP

COMITE ENTREPRISE est une séquence vide pour la MCARP

➔ activité erronée

Variables du Bulletin Individuel RP

Identifiant (dont commune de résidence : 97402)
Adresse déclarée du lieu de travail
SIRET issu du processus de codage
Indicateur de traitement manuel (ici aucun)
Nom déclaré de l'employeur

```
id:          97402___AD14_0006000002
adr:         4110002 RTE DU BOIS DE NEFLE
siret:       31355376000015
i_actet_m:   ___
rs_x1:       CLINIQUE SAINTE CLOTILDE
```

Description de l'écho donné par la MCA

Adresse dans le répertoire (dont commune 9741
Nom (enseigne ou sigle ou raison sociale)
SIRET et type de nom

```
Sadr:        4110127 R   DU BOIS DE NEFLES
nom:         CLINIQUE SAINTE CLOTILDE
Sid:         43156752800028E
```

FSBROWSE WORK.S1 Where ... Obs 5028 Screen 1

```
CJ_CP:          8310     rs:              COMITE ENTREPRISE CLINIQUE STE CLOTILD
SIGLE:                   _____                    APEN:          9420Z     EFF3112_TR:    00          NIC_
DEPCOM_SIEGE:   97411    SIRET:           31355376000015
enseigne:                _____                                                ETAT_ET:       1           EXPL
APET:           9420Z    EMPL_ET:         N          EFF3112_TR_ET: 00        DEP_ETAB:      9D          COM_
TYPE_UNITE:     PM       SOURCE:          SIR        Sadrmot:       NEFLES
Sadr:           411      R   DU BOIS DE NEFLES
nom:            COMITE ENTREPRISE CLINIQUE STE CLOTILD
Sid:            31355376000015R                     Sdc:           97411     Smot1:         COMITE
Smot2:          ENTREPRISE                          Smot3:         CLINIQUE
Smot4:          STE                                 Smot5:         CLOTILD
Smot6:                   _____                      Smot7:
```

Répertoire

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help

Adresse privilégiée faute de correspondance sur le libellé
➔ activité erronée

Sans commentaires…

```
id:         97404___AI11_0006000001

adr:        4140016 R   LAMBERT

siret:      34019165900010

i_actet_m:  ___

rs_x1:      HYPER U SAINT LOUIS

Sadr:       4140016 R   LAMBERT

nom:        HYPER U

Sid:        40236450900024E
```

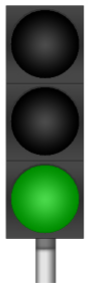
FSBROWSE WORK.S1 Where ... Obs 11266 Screen 1

```
CJ_CP:          9220     rs:              COMITE DES FETES DE ST-LOUIS
SIGLE:                   ________________          APEN:          9004Z     EFF3112_TR:     00          NIC_
DEPCOM_SIEGE:   97414    SIRET:           34019165900010
enseigne:                ____________________________________              ETAT_ET:        1           EXPL
APET:           9004Z    EMPL_ET:         N         EFF3112_TR_ET: 00       DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM       SOURCE:          SIR       Sadrmot:       LAMBERT
Sadr:           4140016 R   LAMBERT
nom:            COMITE DES FETES DE ST-LOUIS
Sid:            34019165900010R                    Sdc:           97414     Smot1:          COMITE
Smot2:          FETES                              Smot3:         ST
Smot4:          LOUIS                              Smot5:                   ____________________________
Smot6:                   ____________________      Smot7:                   ____________________________
```

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help



Adresse privilégiée faute de correspondance sur le libellé

Et coup de chance : TAMOIL semble être dans le groupe LIBYA OIL

- Mais sans conséquence sur le codage de l'activité par la MCA++

```
id:         97408___AM08_0165000001
adr:        4070093 R   JULES VERNE
siret:      31086445900030
i_actet_m:  ___
rs_x1:      TAMOIL REUNION
Sadr:       4070016 A   RICO CARPAYE
nom:        TAMOIL
Sid:        49989400400013E
```

FSBROWSE WORK.S1 Where ... Obs 3774 Screen 1

```
CJ_CP:          5710    rs:             LIBYA OIL REUNION
SIGLE:          ________________        APEN:           4671Z     EFF3112_TR:     11          NIC_
DEPCOM_SIEGE:   97407   SIRET:          31086445900030
enseigne:       ____________________________________________      ETAT_ET:        1           EXPL
APET:           4671Z   EMPL_ET:        0       EFF3112_TR_ET:  11      DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM      SOURCE:         SIR     Sadrmot:        VERNE
Sadr:           4070093 R   JULES VERNE
nom:            LIBYA OIL REUNION
Sid:            31086445900030R                 Sdc:            97407   Smot1:          LIBYA
Smot2:          OIL                             Smot3:          REUNION
Smot4:          ____________________________    Smot5:          ____________________________
```

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help

Contrexemple

```
id:         97408___AN14_0014000001

adr:        4110060 R   ALEXIS DE VILLENEUVE

siret:      48345445000014

i_actet_m:  ___

rs_x1:      BFCOI

Sadr:       4110058 R   ALEXIS DE VILLENEUVE

nom:        BFC OI

Sid:        330176470000955
```

FSBROWSE WORK.S1 Where ... Obs 54712 Screen 1

```
CJ_CP:          9220     rs:              ASS RETRAITES BFC OI
SIGLE:                   ________________          APEN:          9499Z     EFF3112_TR:     __          NIC_
DEPCOM_SIEGE:   97411    SIRET:           48345445000014
enseigne:                ____________________________________________      ETAT_ET:        1           EXPL
APET:           9499Z    EMPL_ET:         N         EFF3112_TR_ET: __       DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM       SOURCE:          SIR       Sadrmot:       VILLENEUVE
Sadr:           4110060 R   ALEXIS DE VILLENEUVE
nom:            ASS RETRAITES BFC OI
Sid:            48345445000014R                    Sdc:           97411     Smot1:          RETRAITES
Smot2:          BFC                                Smot3:         OI
Smot4:          ____________________________       Smot5:         ____________________________
Smot6:                                             Smot7:
```

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help

Le mot « groupe » est vide pour la MCARP
Mais c'est un synonyme potentiel d'établissement sanitaire pour la MCA++

```
id:        97413___DE08_0180000002
adr:       4070003 BD  DES MASCAREIGNES
siret:     47827439200016
actet_x:   CLINIQUE MEDICALE
i_actet_m: ___
rs_x1:     GROUPE LES FLAMBOYANTS
naf2:      86
Sadr:      4070003 B   DES MASCAREIGNES
nom:       CLINIQUE LES FLAMBOYANTS
Sid:       40146909300017R
```

FSBROWSE WORK.S2 Where ... Obs 50845

```
CJ_CP:          5710    rs:             SAS GROUPE LES FLAMBOYANTS
SIGLE:          ____________________            APEN:          7010Z     EFF3112_TR:    11
SIRET:          47827439200016                  enseigne:      ____________________
ETAT_ET:        1       EXPL_ET:        0       APET:          7010Z     EMPL_ET:       0          EFF3112_TR_ET: 11
TYPE_UNITE:     PM      SOURCE:         SIR     Sadrmot:       MASCAREIGNES
Sadr:           4070003 B   DES MASCAREIGNES
nom:            SAS GROUPE LES FLAMBOYANTS
Sid:            47827439200016R                 Sdc:           97407     Smot1:         FLAMBOYANTS
```

L’analyse des résultats
Divergences avec RECAP

- Pour RECAP

  Traitement spécifique de libellés non significatifs

  EDUCATION NATIONALE, RECTORAT, DEPARTEMENT

  Holdings, reprises

  Noms d’usage

  CHU

  Sigles non enregistrés

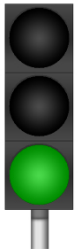
  Arbitrage adresse/libellé

- Contre RECAP

  Des erreurs d’identification de nom

On peut encore améliorer!
Jusqu’à équilibrage des + et des -

Libellés non significatifs



Divergence employeur / lieu de travail

```
id:         97401___A005_0011000001
adr:        404          SIMON LUCAS
siret:      19974813800013
i_actet_m:  CMM
rs_x1:      DEPARTEMENT REUNION
Sadr:       4040117 A    RAYMOND BARRE
nom:        DEPARTEMENT DE LA REUNION
Sid:        22974001400365R
```

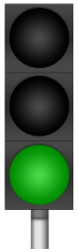
FSBROWSE WORK.S1 Where ... Obs 941 Screen 1

CJ_CP:	7331	rs:	COLLEGE SIMON LUCAS					
SIGLE:				APEN:	8531Z	EFF3112_TR:	12	NIC_
DEPCOM_SIEGE:	97404	SIRET:	19974813800013					
enseigne:						ETAT_ET:	1	EXPL
APET:	8531Z	EMPL_ET:	N	EFF3112_TR_ET:	12	DEP_ETAB:	9D	COM_
TYPE_UNITE:	PM	SOURCE:	SIR	Sadrmot:	STADE			
Sadr:	4040025 R	DU STADE						
nom:	COLLEGE SIMON LUCAS							
Sid:	19974813800013R			Sdc:	97404	Smot1:	COLLEGE	
Smot2:	SIMON			Smot3:	LUCAS			
Smot4:				Smot5:				

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help



Un vrai « plus » de l'expertise humaine
Enrichir la base de connaissance de la MCA++

```
id:         97415___EW11_0011000002

adr:        4110002 R   JOSEPH WETZELL

siret:      18000602500167

i_actet_m:  CMM

rs_x1:      L'UNIVERSITE DE LA REUNION

Sadr:       411     A   DES AIGUES MARINES

nom:        UNIVERSITE DE LA REUNION

Sid:        19974478000248R
```

FSBROWSE WORK.S1 Where ... Obs 584 Screen 1

```
CJ_CP:          7389    rs:             INSTITUT RECHERCHE POUR LE DEVELOPPEME
SIGLE:          IRD                             APEN:           7219Z     EFF3112_TR:     51          NIC_
DEPCOM_SIEGE:   13202   SIRET:          18000602500167
enseigne:       IRD LA REUNION                                            ETAT_ET:        1           EXPL
APET:           7219Z   EMPL_ET:        0       EFF3112_TR_ET:  12        DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM      SOURCE:         SIR     Sadrmot:        WETZELL
Sadr:           4110002 R   JOSEPH WETZELL
nom:            INSTITUT RECHERCHE POUR LE DEVELOPPEME
Sid:            18000602500167R                 Sdc:            97411     Smot1:          INSTITUT
Smot2:          RECHERCHE                       Smot3:          POUR
Smot4:          DEVELOPPEME                     Smot5:          _____________________________
```

Défaut d'outillage du coté RECAP
Trivial pour le Levenshtein généralisé
confirmé par la géographie (B DES MASCAREIGNES dans la ZAC BELVEDERE)

- NAF2 52 au lieu de 71

```
id:         97408___AT03_0066001001

adr:        407          ZAC BELVEDERE

siret:      51834584800015

i_actet_m:  CMM

rs_x1:      T TRAM

Sadr:       407      B    DES MASCAREIGNES

nom:        TTRAM

Sid:        310850078000425
```

FSBROWSE WORK.S1 Where ... Obs 75337 Screen 1

```
CJ_CP:          5710     rs:              TRAM'TISS
SIGLE:          _____                     APEN:          7112B     EFF3112_TR:     __          NIC_
DEPCOM_SIEGE:   97407    SIRET:           51834584800015
enseigne:       ____________________________________________________ ETAT_ET:      1           EXPL
APET:           7112B    EMPL_ET:         N         EFF3112_TR_ET: __ DEP_ETAB:     9D          COM_
TYPE_UNITE:     PM       SOURCE:          SIR       Sadrmot:       VERLAINE
Sadr:           4070106 R   PAUL VERLAINE
nom:            TRAM'TISS
Sid:            51834584800015R                    Sdc:           97407     Smot1:          TRAM
Smot2:          TISS                               Smot3:         _____________________________
Smot4:          ____________________________       Smot5:         _____________________________
```

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help

L'adresse est incomplète, et le libellé est parfaitement reconnu : arbitrage ?
Le répertoire semble incorrect...

- sans conséquences sur l'activité

```
id:         97408___AC22_0237000003

adr:        407     R   LABOURDONNAIS

siret:      49100584900014

i_actet_m:  CMM

rs_x1:      AUTO ECOLE DU MARCHE

Sadr:       4070009 R   LOUISE MICHEL

nom:        AUTO ECOLE DU MARCHE

Sid:        34118477800031E
```

FSBROWSE WORK.S1 Where ... Obs 58783 Screen 1

```
CJ_CP:          5499    rs:             VIRAGE SARL
SIGLE:          ________________                APEN:           8553Z     EFF3112_TR:     02          NIC_
DEPCOM_SIEGE:   97407   SIRET:          49100584900014
enseigne:       ____________________________________                      ETAT_ET:        1           EXPL
APET:           8553Z   EMPL_ET:        0       EFF3112_TR_ET:  02        DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM      SOURCE:         SIR     Sadrmot:        LABOURDONNAIS
Sadr:           4070062 R   LABOURDONNAIS
nom:            VIRAGE SARL
Sid:            49100584900014R                 Sdc:            97407     Smot1:          VIRAGE
Smot2:          ____________________________    Smot3:          ____________________________
Smot4:          ____________________________    Smot5:          ____________________________
```

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help

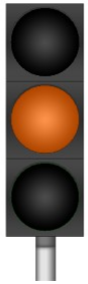
Attention à l'interprétation des adresses !
Le C DU TOUR DES ROCHES est à SAVANNAH
 Pas de chance pour RECAP c'était une laiterie

```
id:         97408___A018_0022005002

adr:        415          SAVANNAH

siret:      33233238600272

i_actet_m:  CMM

rs_x1:      CILAM

Sadr:       415      C    DU TOUR DES ROCHES

nom:        CILAM

Sid:        310864038000575
```

FSBROWSE WORK.S1 Where ... Obs 9548 Screen 1

```
CJ_CP:          5599     rs:              VINDEMIA DISTRIBUTION
SIGLE:          _____________________     APEN:          4711D     EFF3112_TR:     51          NIC_
DEPCOM_SIEGE:   97418    SIRET:           33233238600272
enseigne:       JUMBO SCORE                                        ETAT_ET:        1           EXPL
APET:           4711F    EMPL_ET:         0         EFF3112_TR_ET: 32        DEP_ETAB:       9D          COM_
TYPE_UNITE:     PM       SOURCE:          SIR       Sadrmot:       SAVANNAH
Sadr:           4150004 R    DE SAVANNAH
nom:            VINDEMIA DISTRIBUTION
Sid:            33233238600272R                     Sdc:           97415     Smot1:          VINDEMIA
Smot2:          DISTRIBUTION                        Smot3:
Smot4:          ____________________________________ Smot5:       ____________________________________
Smot6:                                              Smot7:
```

…et des matchs nuls



FSBROWSE WORK.S2 Where ... Obs 7085 Screen 1

CJ_CP:	6599	rs:	SOGIM GUILLAUME IMMOBILIER					
SIGLE:	SOGIM			APEN:	6832A	EFF3112_TR:	__	
SIRET:	32245286300028			enseigne:				
ETAT_ET:	1	EXPL_ET:	0	APET:	6832A	EMPL_ET:	N	EF
TYPE_UNITE:	PM	SOURCE:	SIR	Sadrmot:	BOURBON			
Sadr:	4150045 A DE BOURBON							
nom:	SOGIM							
Sid:	322452863000285			Sdc:	97415	Smot1:	SOGIM	

id: 97401___AP06_0015000001

adr: 415

siret: 34907384100026

actet_x: OUVRIER

i_actet_m: CMM

rs_x1: SOGIM ST PAUL

naf2: __

Sadr: 4150045 A DE BOURBON

nom: SOGIM

Sid: 322452863000285

Deux échos possibles
avec des activités différentes

CJ_CP:	5710	rs:	SOC GENERALE INVESTISSEMENT MASCAREIGN					
SIGLE:	SOGIM			APEN:	1101Z	EFF3112_TR:	03	
SIRET:	34907384100026			enseigne:				
ETAT_ET:	1	EXPL_ET:	0	APET:	1101Z	EMPL_ET:	0	EFF3112_TR_ET: 03
TYPE_UNITE:	PM	SOURCE:	SIR	Sadrmot:	SUD			
Sadr:	415	LA PLAINE CHABRIER SUD						
nom:	SOGIM							
Sid:	349073841000265			Sdc:	97415	Smot1:	SOGIM	

serveur-de-calcul.insee.fr - Connexion Bureau à distance

SAS

File Edit View Tools Search Solutions Window Help


On fait ce qu'on peut…

```
id:        97415___ET09_0024002002
adr:       4110040 R   ALEXIS DE VILLENEUVE
siret:     70201631200168
actet_x:   FINANCE
i_actet_m: CMM
rs_x1:     SOCIETE GENERALE
naf2:      __
Sadr:      4110169TR   DU GAL DE GAULLE
nom:       LE GENERAL
```

En fait c'est la
BANQUE FRANCAISE DE L'OCEAN INDIEN
détenue à 50 % par la SOCIETE GENERALE
Au no 58, du bon coté de la rue
Et enlever 'SOCIETE' était une mauvaise idée

FSBROWSE WORK.S2 Where ... Obs 92787

```
CJ_CP:        5599    rs:            COMPAGNIE GENERALE D AFFACTURAGE
SIGLE:        CGA                    APEN:          6619B     EFF3112_TR:    32
SIRET:        70201631200168         enseigne:      CGA
ETAT_ET:      1       EXPL_ET:       0         APET:          6499Z     EMPL_ET:       0        EFF3112_TR_ET: 03
TYPE_UNITE:   PM      SOURCE:        SIR       Sadrmot:       VILLENEUVE
Sadr:         4110047 R   ALEXIS DE VILLENEUVE
nom:          CGA
Sid:          70201631200168E                  Sdc:           97411     Smot1:         CGA
Smot2:        ________________________________ Smot3:        ________________________________
Smot4:        ________________________________ Smot5:        ________________________________
```

A propos d’outils...

Expressions régulières (regex)

- Finalité : décrire le contenu d'une chaîne de caractères de façon concise
  - sous forme de chaîne de caractères
  - incluant des caractères spéciaux précisant ce qu'on attend
- Exemple : une adresse e-mail insee simple (prénom éventuellement composé, nom non composé)

  [a-z]+(-[a-z]+)?.[a-z]+@insee.fr
- Mini langage de programmation permettant de faire des recherches, des substitutions en spécifiant la forme de ce qu'on attend plutôt que la façon de l'obtenir (programmation par assertions, cf. Prolog).
- Origine : PERL, implémentations diverses
  - SAS : prxparse…
  - Bibliothèque PCRE interfacée en R, Python, Lisp...

Familles de caractères

.	N’importe quel caractère
\d	Un chiffre
\D	N’importe quel caractère qui ne soit pas un chiffre
[a-z**]** **[**abd**]**	Un caractère entre « a » et « z » : alphabétique et minuscule Un « a », un « b » ou un « d »
[^a-z**]** **[^**abd**]**	N’importe quel caractère qui ne soit pas alphabétique minuscule
\[	Le caractère «[» dépourvu de sa signification syntaxique

Directives sur le positionnement

^	Début de la chaîne	^a	**a**bracadabra
$	Fin de la chaîne	a$	abracadabr**a**

Répétitions et alternatives

<*expr*>**?**	<expr> éventuellement manquant	essais?	essai essais
<*expr*>*****	<expr> éventuellement manquant ou répété	es*ai	essai eai
<*expr*>**+**	<expr> éventuellement répété		
<*expr1*>**|**<*expr2*>	<expr1> ou <expr2>	(a|b|c)	ess**a**i

Juste pour voir..

(?=<*expr*>**)**	Ce qui suit doit convenir pour <expr>	a(?=c)	abr**a**cadabra
(?!<*expr*>**)**	Ce qui suit ne doit pas convenir pour <expr>	a(?![bc])	abrac**a**dabra
(?<=<*expr*>**)**	Ce qui précède doit convenir pour <expr>	(?<=bra).	abra**c**adabra
(?<!<expr>**)**	Ce qui précède ne doit pas convenir pour <expr>	(?<!a).	ab**r**acadab**r**a

Groupes de capture

Référence à un précédent match	(.)\1	e**ss**ai
Référence à un précédent motif	(\{([a-z]+|(?1))+\})	f**{a{b}}**g

Exemples

Une portion du code de toilettage du nom de l'employeur

```
i2 = prxparse("s/^(MINISTERE )?(DE )?(L('|' | ))?EDUCATION( NATIONALE)?/éducation/");
i3 = prxparse("s/^(MINISTERE )?(DE )?(L('|' | ))?INTERIEUR/intérieur/");
i4 = prxparse("s/^(MINISTERE )?(DE )?(LA )?DEFENSE/défense/");
i5 = prxparse("s/^(MINISTERE )?(DE )?(LA )?JUSTICE/justice/");
i6 = prxparse("s/^(MINISTERE )?((DE L')?ECONOMIE ET )?(DES )?FINANCES/finances/");
i7 = prxparse("s/^(MINISTERE )?(DU )?TRAVAIL/travail/");
i8 = prxparse("s/(AUTO ?ENTREPRENEUR|INTERIMAIRE)//");
i9 = prxparse("s/(PLUSIEURS|DIVERS|PARENTS?|PARTICULIERS?)( EMPLOYEURS?)?/inconnu/");
```

Une portion des données utilisées
par le code de transformation du libellé d'activité en code d'activité

^(RE)?VEN(DEU(R|SE)|TES?)(DE MARCHANDISES)?$	COMMERCE
^(COMMERCE VENTE|VENTE COMMERCE)	COMMERCE
^(VENTES?(A DISTANCE)?)?E ?(\.|- ?)?COMMERCE	COMMERCE PAR INTERNET
(VENTE|COMMERCE) (EN LIGNE|A DISTANCE)	COMMERCE PAR INTERNET
^(COMMERC(ANTE?|IALISATION)|MARCHANDE?) ?(de|EN)?	COMMERCE
^COMMERCE ?(de)?DETAIL ET (de)?GROS	COMMERCE
^COMMERCE ?(de)?GROS ET (de)?DETAIL	COMMERCE
^(COMMERCE|VENTE) ?(de|EN)GROS ?(de|EN)?	COMMERCEg
^(COMMERCE|VENTE) ?(de|EN |AU)?DETAIL ?(de|EN)?	COMMERCEd
^(VENTES? |COMMERCE (de)?)(.*) EN GROS$	COMMERCEg\1
^(.*) EN GROS$	COMMERCEg\1

Merci de votre attention