pAq : logiciel d'analyse des logs d'un « proxy web » d'accès aux périodiques électroniques

Application au proxy G@el de l'Université Joseph Fourier et de Grenoble INP

Urfist Lille - 28 novembre 2008

Serge.Rouveyrol@imag.fr – Service information Scientifique

- Projet démarré en 2002, a partir des idées du projet Eurydice (modules relieur-éditeur écrit avec des affectations d'expressions régulières en langage perl)

- 2006 : Sabine Barral, chargé de mission « indicateurs d'usage des ressources électroniques » par la la DGES, qui travaillait alors avec un logiciel équivalent basé sur des modules en XML (développé à l'UTC par David Lewis) a contacté Laurent Perillat et moi-même, nous avons travaillé ensemble à l'ajout de nouveau relieurs et de nouveaux compteurs à pAq, François Charbonnier a utilisé pAq et ajouté de nouveaux compteurs pour analyser les logs de l'université de de Lyon1.

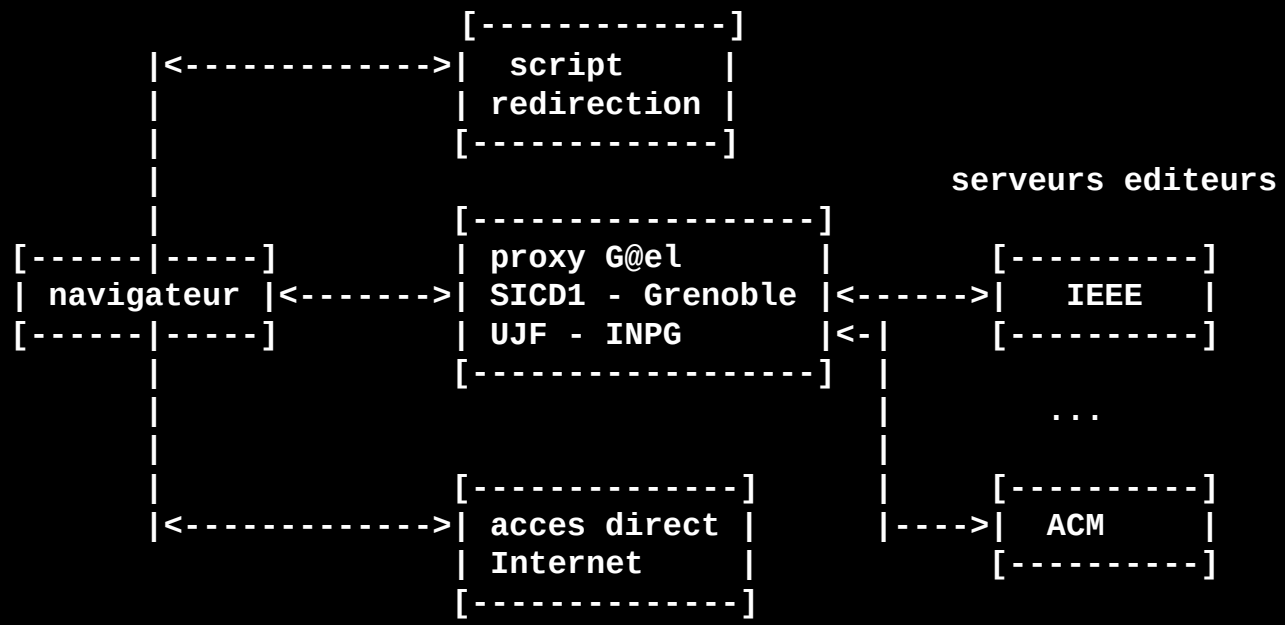
- But: Analyser les logs produit par un proxy web d'accès aux périodiques

  - ✓ essayer de savoir : qui lit quoi chez moi?
  - ✓ détecter les abus de proxy par une machine chez un éditeur
  - ✓ fournir les résultats sous forme de fichiers excel
  - ✓ les modules relieur-éditeur ne sont plus des programmes mais des affectation d'expressions régulières à des variables compteurs qui peuvent être développés par des documentalistes … connaissant les expressions régulières

Principe de fonctionnement du proxy web (serveur mandataire) G@el

un programme de redirection est exécuté sur le navigateur lors de chaque clic sur un

lien, ce programme force le navigateur à passer par le proxy web pour un accès à un périodique électronique, le proxy conserve dans des fichiers logs, toutes les informations d'accès aux périodiques éléctroniques



Périodiques électroniques g@el - Consultation - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

http://web.ujf-grenoble.fr/BUS/Ressources_Electroniques/Periodiques_Electroniques_Gael/Gael_PE.php

Démarrage Dernières nouvelles

- Nouveautés
- Bases de Données - g@el
- Périodiques Electroniques - g@el
  - Présentation et infos
  - Consultation
- Thèses en ligne
- Ressources Internet

Périodiques Electroniques g@el

Consultation

Liste des revues classées par ordre alphabétique des TITRES

A | B | C | D | E | F | G | H | I | J | K | L | M
N | O | P | Q | R | S | T | U | V | W | X | Y | Z

Liste des revues classées par EDITEURS

- Dernières mises en ligne
- Informations pratiques
- Revues en "open access" (DOAJ et Free Medical Journals)

...ans g@el par ...a REVUE
...mot (tronc...ature
...égé de la revue
OK

Liste des revues classées par DOMAINES
(ne concerne que les revues sélectionnées par la Commission de Sélection des Titres)

Biologie
Chimie
Electronique
Génie industriel

Afficher

L'accès est réservé aux membres de g@el

...lonnées g@el compte 9372 titres de revues au format électronique

Informations importantes

Paramètres de connexion

Configuration du serveur proxy pour accéder à Internet

- ○ Connexion directe à Internet
- ○ Détection automatique des paramètres de proxy pour ce réseau
- ○ Configuration manuelle du proxy :

Proxy HTTP : Port : 0

☐ Utiliser ce serveur proxy pour tous les protocoles

Proxy SSL : Port : 0
Proxy FTP : Port : 0
Proxy gopher : Port : 0
Hôte SOCKS : Port : 0

○ SOCKS v4 ◉ SOCKS v5

Pas de proxy pour : localhost, 127.0.0.1

Exemples : .mozilla.org, .asso.fr, 192.168.1.0/24

◉ Adresse de configuration automatique du proxy :

http://proxy.ujf-grenoble.fr/auto-proxy.pac Recharger

OK Annuler Aide

Terminé

Autorisation d'accès au proxy G@el

L' autorisation d'accès au proxy G@el par un laboratoire membre doit être configuré dans le fichier de configuration du proxy

```
...
# accès autorisé a toutes les machines des réseaux
# du  domaine imag.fr

acl gael-imag src 129.88.0.0/16         \
                  147.171.0.0/16        \
                  152.77.200.0/23       \
                  195.221.224.0/21      \
                  194.199.21.0/24       \
                  194.199.23.128/26     \
                  194.199.26.0/24       \
                  194.199.25.0/25
...
```

Accès aux périodiques via le proxy G@el

l'url fourni par G@el http://proxy.ujf-grenoble.fr/auto-proxy.pac
est à configurer dans le navigateur, elle référence un programme de redirection a exécuter sur le navigateur lors de chaque clic sur un lien.

```
function FindProxyForURL(url, host) {

   ...
   if (localHostOrDomainIs(host,"ieeexplore.ieee.org"))
         return "PROXY cw3-sicd1.ujf-grenoble.fr:3128";
   ...
   if (localHostOrDomainIs(host,"portal.acm.org"))
         return "PROXY cw3-sicd1.ujf-grenoble.fr:3128";
   return "DIRECT";
}
```

il y a actuellement 567 tests de redirection vers des serveurs d'éditeurs exécuté sur le navigateur pour chaque clic sur un lien

Format d'une ligne de log sur le proxy

1225798193.217 11400 193.48.43.40 TCP_MISS/200 2034315 GET http://ieeexplore.ieee.org/stampPDF/getPDF.jsp?arnumber=4626008&isnumber=4625985 - DIRECT/140.98.193.112 application/pdf

- temps en secondes depuis le 1er janvier 1970
- temps de réponse à la requête en millisecondes
- adresse IP de la machine qui fait la demande
- taille en octets du fichier délivré
- pas dans le cache – code erreur (ici pas d'erreur)
- méthode pour passer les paramètres du navigateur au serveur
- url demandé
- adresse ip du serveur
- type de donnée

configuration initiale du proxy

- la fonction cache du proxy doit être désactivée
- le log de la QUERY_STRING (paramètres) doit être activé pour ne pas perdre les informations utiles

les urls seront alors sous la forme :

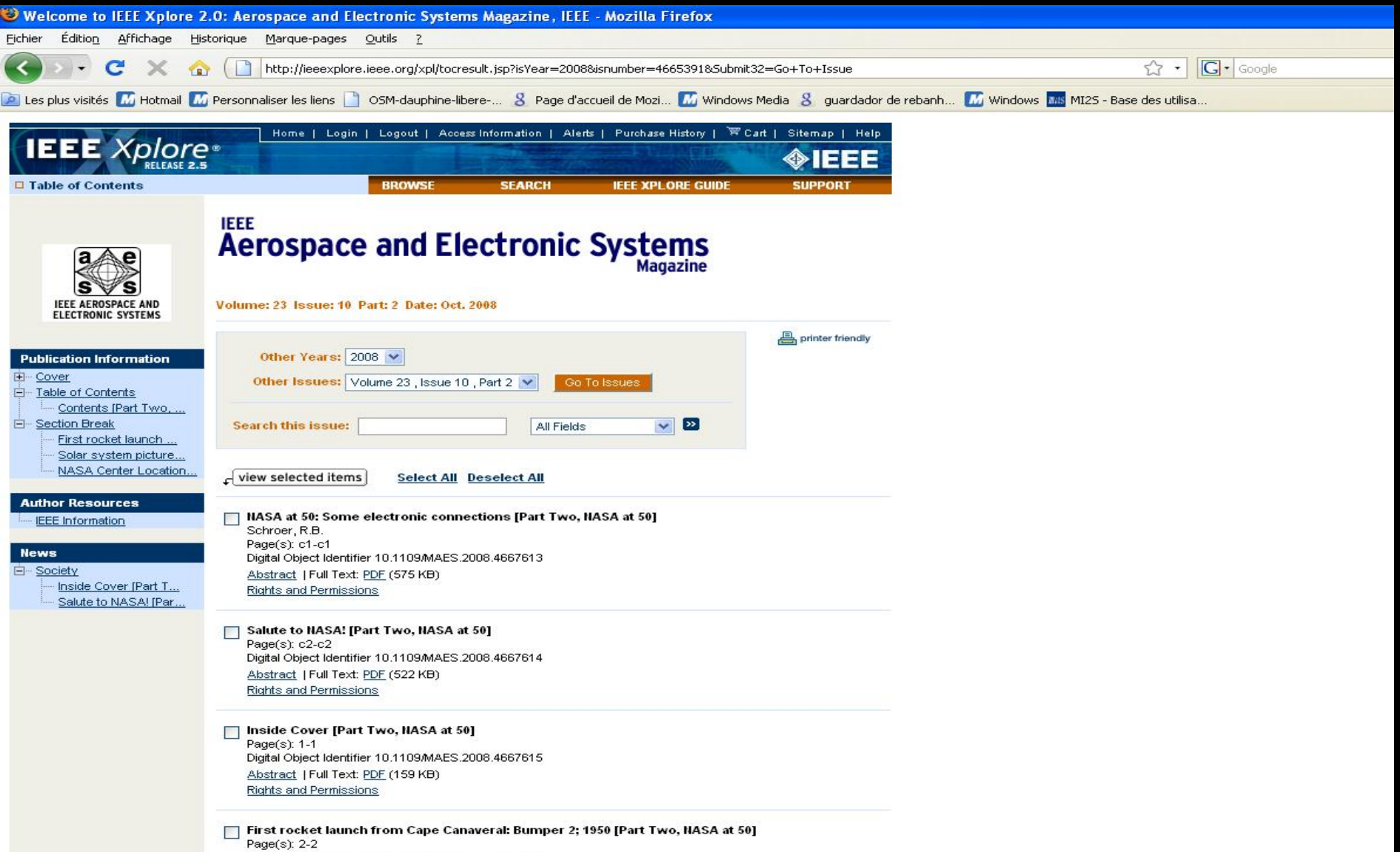
http://ieeexplore.ieee.org/stampPDF/getPDF.jsp?arnumber=4626008&isnumber=4625985

Et non sous la forme:

http://ieeexplore.ieee.org/stampPDF/getPDF.jsp

Cela provoque une augmentation d'environ 20% de la taille du fichier log

Log généré sur le proxy lors du premier accès correspondant à l' affichage de cette page du sommaire d'une revue IEEE



1227631405.577 202 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/styles/ieee_menubar/javascript_main.js - DIRECT/140.98.193.112 -
1227631405.580 205 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/styles/css/StyleSheet.css - DIRECT/140.98.193.112 -
1227631405.697 116 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/ieee_menubar/browser_detection.js - DIRECT/140.98.193.112 -
1227631405.807 109 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/ieee_menubar/activemenu.js - DIRECT/140.98.193.112 -
1227631405.938 131 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/ieee_menubar/init_cssposition_guest.js - DIRECT/140.98.193.112 -
1227631405.957 12 129.88.27.135 TCP_IMS_HIT/304 289 GET http://ieeexplore.ieee.org/styles/ieee_menubar/javascript_main.js - NONE/- application/x-javascript
1227631406.001 1730 129.88.27.135 TCP_MISS/200 72496 GET http://ieeexplore.ieee.org/xpl/tocresult.jsp?isYear=2008&isnumber=4665391&Submit32=Go+To+Issue - DIRECT/140.98.193.112 text/html
1227631406.076 113 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/xplorejs/Xplorejs.js - DIRECT/140.98.193.112 -
1227631406.206 129 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/xplorejs/tree.js - DIRECT/140.98.193.112 -
1227631406.316 110 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/xplorejs/tree_tpl.js - DIRECT/140.98.193.112 -
1227631406.425 108 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/styles/cmsjs/62/lss62.js - DIRECT/140.98.193.112 -
1227631406.544 119 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/1px_white.gif - DIRECT/140.98.193.112 -
1227631406.545 101 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/transparent_spacer.gif - DIRECT/140.98.193.112 -
1227631406.650 105 129.88.27.135 TCP_REFRESH_MISS/200 3452 GET http://ieeexplore.ieee.org/styles/cmsjs/62/4665391/4665391.js - DIRECT/140.98.193.112 application/x-javascript
1227631406.651 105 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/viewselected.gif - DIRECT/140.98.193.112 -
1227631406.659 204 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/icon_cart_wh.gif - DIRECT/140.98.193.112 -
1227631406.660 215 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/top_left.gif - DIRECT/140.98.193.112 -
1227631406.665 203 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/branding_images/aes_m_l.gif - DIRECT/140.98.193.112 -
1227631406.670 209 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/L2_pageid_square.gif - DIRECT/140.98.193.112 -
1227631406.756 103 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/top_right.gif - DIRECT/140.98.193.112 -
1227631406.757 106 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/society_images/sl_aess.gif - DIRECT/140.98.193.112 -
1227631406.766 105 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/empty.gif - DIRECT/140.98.193.112 -
1227631406.768 107 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/nav_bg.gif - DIRECT/140.98.193.112 -
1227631406.774 108 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/line.gif - DIRECT/140.98.193.112 -
1227631406.778 108 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/empty2.gif - DIRECT/140.98.193.112 -
1227631406.880 102 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/joinbottom.gif - DIRECT/140.98.193.112 -
1227631406.885 106 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/arrow_submit.gif - DIRECT/140.98.193.112 -
1227631406.885 105 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/icon_print_toc.gif - DIRECT/140.98.193.112 -
1227631406.885 106 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/plus.gif - DIRECT/140.98.193.112 -
1227631406.886 106 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/plusbottom.gif - DIRECT/140.98.193.112 -
1227631406.895 115 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/join.gif - DIRECT/140.98.193.112 -
1227631406.981 100 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/empty3.gif - DIRECT/140.98.193.112 -
1227631406.982 201 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/minus.gif - DIRECT/140.98.193.112 -
1227631406.985 206 129.88.27.135 TCP_REFRESH_HIT/304 321 GET http://ieeexplore.ieee.org/images/minusbottom.gif - DIRECT/140.98.193.112 -
1227631406.987 101 129.88.27.135 TCP_REFRESH_HIT/304 229 GET http://ieeexplore.ieee.org/images/inspec_003366.gif - DIRECT/140.98.193.112 -

Filtrage des logs bruts

On ne conserve que :

- Les lignes qui correspondent aux serveurs définis dans le programme de redirection
- Les lignes de type html, pdf ou postscript
- les lignes dont le code erreur est TCP_MISS/200 (OK)

On supprime les lignes « clics rapides » (Counter)
- même ip , même url à moins de 30s d'intervalle pour le type pdf , moins de 10s pour le type html

log G@el octobre 2008

log brut : 1,3 Go (8 milions de lignes) (Log brut compressé : 214 Mo)

log filtré : 59 Mo (295000 lignes)
durée filtrage : 1 mn 27 s

Suppression log « clics rapides » : 55 Mo (271000 lignes)
Durée filtrage : 13 s

Structure de l’analyseur pAq

Analyse des logs filtrés (écriture d'un relieur-éditeur)

En étudiant les logs d'un éditeur on essaie d'écrire des « expressions régulières pour différents types de compteurs :

EDITEUR_SEARCH

TITRE_TOC
ou peut-etre ... EDITEUR_TOC
ou peut-etre rien

TITRE_ABSTRACT
ou peut-etre ... EDITEUR_ABSTRACT
ou peut-etre rien

TITRE_REF
ou peut-etre ... EDITEUR_REF
ou peut-etre rien

TITRE_FULL_HTML
ou peut-etre ... EDITEUR_FULL_HTML
ou peut-etre rien

TITRE_FULL_PDF
ou peut-etre ... EDITEUR_FULL_PDF
ou peut-etre rien

rappel sur les expressions régulières

Les expressions régulières sont des modèles créés à l'aide de caractères ASCII permettant de manipuler des chaînes de caractères, c'est-à-dire permettant de trouver les portions de la chaîne correspondant au modèle.

[ ] une liste de caractères ex : [a-d] signifie caractères a,b,c,d
. Le caractère point représente n'importe quel caractère
* indique la répétition 0 a n fois de l'élément la précédant
+ indique la répétition 1 à n fois de l'élément la précédant (identique à ..*)
^ - Placé en début d'expression il signifie "chaîne commençant par .. "
- Utilisé à l'intérieur d'une liste il signifie "ne contenant pas les caractères suivants... »
$ - fin de chaîne
() Les parenthèses définissent un élément composé de l'expression régulière qu'elle contient et mémorise la sous-chaîne sélectionnée
| Occurrence de l'élément situé à gauche de cet opérateur ou de celui situé à droite

Exemples:
^[a-c].*[^Z]$: toutes les chaînes dont la 1ere lettre est a, b ou c et dont le dernier caractère n'est pas Z et de longueur >= 2

^(.)$1.+(.)$2$: toutes les chaines ayant les 2 premiers et les 2 derniers caractères identiques et de longueur superieure > 2

exemple du module relieur-éditeur pour sciencedirect.com

```
package SCIENCEDIRECT_COM;

$regexp_EDITEUR_SEARCH_1 = "^http://www\.sciencedirect\.com/science?.+ob=MiamiSearchURL";
$regexp_EDITEUR_SEARCH_2 = "^http://www\.sciencedirect\.com/science?.+ob=QuickSearchURL";
$regexp_EDITEUR_SEARCH_3 = "^http://www\.sciencedirect\.com/science?.+ob=ArticleListURL.+\&.
+sort";
$regexp_EDITEUR_SEARCH = "$regexp_EDITEUR_SEARCH_1|$regexp_EDITEUR_SEARCH_2|
$regexp_EDITEUR_SEARCH_3";

$regexp_TITRE_ABSTRACT_HTML_1 = "^http://www\.sciencedirect\.com/.+fmt=summary.+\&_cdi=([0-
9]+)";
$regexp_TITRE_ABSTRACT_HTML_2 = "^http://www\.sciencedirect\.com/.
+ob=ArticleURL.+fmt=full.+\&_cdi=([0-9]+)";
$regexp_TITRE_ABSTRACT ="$regexp_TITRE_ABSTRACT_HTML_1|
$regexp_TITRE_ABSTRACT_HTML_2";

$regexp_EDITEUR_REF = "^http://www\.sciencedirect\.com/science?.+ob=BrowseListURL";
$regexp_EDITEUR_TOC = "^http://www\.sciencedirect\.com/science?.+ob=PublicationURL\&_tockey";

$regexp_TITRE_FULL_HTML = "^http://www\.sciencedirect\.com/science?.+ob=ArticleURL.
+fmt=&_orig.+\&_cdi=([0-9]+)";


$regexp_TITRE_FULL_PDF_1 = "^http://www\.sciencedirect\.com/.+\&_cdi=([0-9]+)\&_.+sdarticle.pdf";
$regexp_TITRE_FULL_PDF_2 = "^http://www\.sciencedirect\.com/science/MiamiMultiMediaURL.+(/[0-
9]+/)";
$regexp_TITRE_FULL_PDF = "$regexp_TITRE_FULL_PDF_1|$regexp_TITRE_FULL_PDF_2";
```

tableau d'association construit par programme à partir des pages html du serveur web de l'éditeur
(2366 associations numéro - titre)

4874 ACC Current Journal Review
6108 ACOG Clinical Review
12985 Academic Radiology
5794 Accident Analysis & Prevention
6679 Accident and Emergency Nursing
20463 Accounting Forum
6023 Accounting, Management and Information Technologies
5957 Accounting, Organizations and Society
5679 Acta Astronautica
20189 Acta Biomaterialia
33065 Acta Ecologica Sinica
33027 Acta Genetica Sinica
18099 Acta Histochemica
5556 Acta Materialia
33028 Acta Mathematica Scientia
…..

Logs G@el octobre 2008

log brut : 1,3 Go (8 milions de lignes)
Log brut compressé : 214 Mo
durée compression : 1mn 49

log filtré : 59 Mo (295000 lignes)
durée filtrage : 1 mn 27 s

Suprresision lignes « clics rapides » : 55 Mo (271000 lignes)
Durée filtrage : 13 s

Type html : 223000 lignes
Type pdf : 48000 lignes
Type postsript : 71 lignes

Durée analyse : 5mn 31s

Exemple de résultats obtenus

fichiers excel résultats

Conseil de lecture:

Rapport de Sabine Barral « Mission Indicateurs d'usage des ressources électroniques »

Google: indicateurs Barral

http://www.adbu.fr/IMG/pdf/usage_des_ressources_electroniques.pdf

Distribution de pAq