Système Hérisson : le pouvoir français veut tout surveiller sur Internet

21/04/2009 à 10h26

En ces temps d'HADOPI, d'EDVIGE ou de CRISTINA, un autre projet gouvernemental se révèle au grand jour : HERISSON. Le cahier des charges, du moins la liasse qui est présentée comme telle, laisse entrevoir un système de surveillance de grande ampleur sur tout ce qui transite sur les réseaux. Analyse.

La France serait-elle en voie de mettre en place l'équivalent d'un réseau Échelon dans l'hexagone pour surveiller et cataloguer tous les faits et gestes de tous les médias, radio, TV et internet compris ? Tout porte à le croire, selon une liasse de documents qui décrit le Système Hérisson, éponyme d'« Habile Extraction du Renseignement d'Intérêt Stratégique à partir de Sources Ouvertes Numérisées ». On pourra en découvrir la totalité sur ce lien Google Doc (Hérisson)

Selon ces documents qui nous ont été présentés comme étant le cahier des charges de ce système et daté de l'année passée, Hérisson tend vers la mise en place d'un système de collecte automatisée et de recherche d'informations dignes d'un scénario de film d'espionnage ou de Matrix. La paternité du dossier est attribuée à la DGA, Délégation Générale pour l'Armement dont la mission consiste au suivi et à la validation des moyens de défense militaires et territoriales.

Dans les références de ce fameux des charges, classées dans la rubrique « Biblio » il est ainsi fait mention de plusieurs textes officiels dont l'instruction générale interministérielle sur la sécurité des systèmes d'information « qui font l'objet d'une classification de défense pour eux-mêmes ou pour les informations traitées » (n°900/SGDN/SSD/DR ou 900/DISSI/SCSSI/DR du 20 juillet 1993), texte qui gère le secret défense. On trouve encore mention de diverses

directives ou instructions relatives à la protection contre les signaux parasites compromettants.

Un appel d'offres lancé en 2007

Chose certaine : le système est à rattacher à un appel d'offres lancé en 2007 par le ministère de la défense, la délégation générale pour l'armement, la direction des systèmes d'armes et le service centralisé des achats. Le marché public a pour objet « la Conception, réalisation et évaluation d'une plate-forme dédiée au traitement des sources ouvertes pour le renseignement militaire d'intérêt stratégique ». L'appel d'offres a une durée minimale de 36 mois à compter de la notification du marché et est en phase avec ce cahier des charges.

IRC, Newsgroup, FTP, P2P, HTTP, etc.

Dans ce fameux cahier des clauses techniques particulières (CCTP), on découvre ainsi que le système vise à glaner à peu près tout ce qui passe sur les réseaux aux fins notamment de renseignement militaire d'intérêt stratégique.

On peut le voir dans le tableau ci-dessous (plus lisible ici: tableau) que par exemple tous les contenus qui transitent sur les chats IRC, les mailings listes, les forums, les réseaux sociaux, les Newsgroups, les flux RSS, les blogs, podcasts, et les systèmes P2P feront l'objet d'une surveillance approfondie, avec une priorité soit « importante » soit « primordiale » selon les flux. La plate-forme est destinée à être installée au centre d'expertise parisien (CEP) d'Arcueil indiquent les pages.

Mais c'est surtout dans la section « Exigences » qu'on apprécie l'ampleur du système Hérisson. Il est calibré ainsi pour accéder à n'importe quel contenu ((texte, image, son et vidéo), quel que soit les langages utilisés « HTML, PHP, ASP…) ») en prenant en compte « la problématique des liens «cachés» en extrayant les liens a minima dans les applications flash contenues dans une page, les codes Java Script ». Hérisson sait aussi collecter et stocker le contenu complet d'un site FTP ou collecter simplement son arborescence, sait

télécharger les fichiers disponibles sur un réseau P2P.

Tous les types de fichiers

Notre curieux mammifère poursuit sa longue route en ayant « la capacité de collecter des données via les protocoles : MMS (flux vidéo type Windows Media Player) RSTP (flux vidéo type Real Player) POP3 (messagerie) », et donc des emails. Pour faire bonne mesure, « le système HÉRISSON a la capacité de gérer un éventail large, non restreint et évolutif de formats de documents de type : Vidéo (AVI, MPG, MOV, MP4, Real, FLV, OGM …) ; Audio (WAV, MP3, OGG…) ; Image (BMP, JPG, TIFF…) ; Texte (HTML, MHTML, Open Document, Open XML/Microsoft Office, Adobe PS/PDF, Flash).

L'œuvre d'observation et de collecte « prend [aussi] en compte la problématique du Web invisible en proposant des moyens, méthodes et outils, permettant l'accès à ces sous-ensembles du Web Invisible ». Le tout est archivé et chaque mise à jour « d'un gisement Web, FTP ou NMTP » fait l'objet d'un stockage différencié pour tenir compte des évolutions des contenus amassés.

Le système s'attaque aussi bien au web qu'aux radios et aux TV. Ainsi, les langues des documents audio devant être identifiées dans le système HÉRISSON sont :

[P]rimordial : le français, l'anglais, l'arabe, le russe, le farsi.

[I]mportant ; l'espagnol, l'allemand, , le chinois mandarin, l'italien,k le serbo-croate, l'hindi, le japonais, le coréen, le turc, l'ukrainien, l'hébreu, l'urdu, l'albanais, le macédonien.

[S]ouhaitable : le néerlandais, le grec, le portugais, le polonais.

Pour les images, Hérisson permet l'accès aux caractéristiques brutes et aux métadonnées d'une image, « comme le type et les caractéristiques d'encodage, la résolution, le taux de compression, les champs EXIF, les statistiques usuelles liées à l'image histogramme, moments centrés…), … il note aussi la provenance de l'image (URL pour les images obtenues via Internet) et éventuellement le nom du photographe ».

Hérisson « reconnaît les images transformées » et « lorsqu'une photographie a subi des transformations (rotations, changement d'échelle, compression, modification des contrastes,…), [il] sait reconnaître l'image originale parmi les images qui en sont issues ».
Pour pousser un peu plus loin l'excavation des données glanées, Hérisson permet « la détection et classification d'objets contenus dans une image (personne, véhicule, meuble…). Cette fonctionnalité sera évaluée avec les méthodologies utilisées dans les évaluations Techno-Vision/ImagEval, tâche n°4, PASCAL VOC (Visual Object Classes challenge), et CLEAR (tâche « 2D Face détection ») ». Sur les vidéos, HÉRISSON « permet la détection et l'identification de personne dans une vidéo. Cette fonctionnalité sera évaluée avec la méthodologie utilisée dans l'évaluation CLEAR (tâche « person identification ») »

Reconnaissance de caractères, faciale et d'objets

S'ajoutent à la [très] longue liste, la reconnaissance des caractères (détection du texte, détection des colonnes, détection de zones de texte, détection de tableaux, détection de formats de caractères et de polices, détection de graphiques, etc.), la détection d'éléments « de contexte d'une image (photo couleur, photo noir&blanc, reproduction artistique, visage, scène intérieur/extérieur, ville/campagne/bord de mer/montagne/…, jour/nuit; présence/absence d'objets manufacturés où d'artefacts d'origine humaine, …) »…

Des données exploitées par un moteur multicritère

Une fois la masse de données stockées, l'exploitation se fait avec une fonction de recherche avancée (ou multicritère) qui permet a minima de préciser une combinaison booléenne des paramètres suivants:
Date (avec intervalle),
Taille (avec un intervalle),
Type de document (texte, image, graphique, vidéo, audio),
Format de fichier,
URL d'origine.

Les propriétés du contenu :
L’exclusion de mot,
Une expression exacte,
La combinaison booléenne entre les mots-clés,
Les mots commençant par une chaîne de caractères donnée,
Distance entre les mots dans le texte,
Expression régulière,
Localisation dans la page (titre / corps de texte),
Une écriture phonétique,
Une orthographe approchée,
Dans une page (URL) précise,
Dans les pages contenant un lien précis.
Exemple piquant, « Le système offre la possibilité de reconnaître l’entité elle-même de plusieurs manières différentes: Soit par mention explicite. Exemple : « Nicolas Sarkozy », Par mention relative (« le nouveau président français »), par mention partielle (« Monsieur Sarkozy », « le Président »), par anaphore (« il », « celui-ci »), par surnom (« Sarko ») ».
(Source: pcinpact.com)