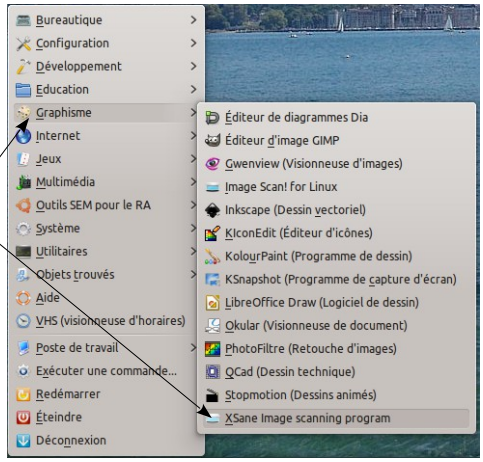
Introduction

La reconnaissance optique de caractères désigne les procédés informatiques pour la traduction d'images de textes (imprimés, articles de journaux, livres, ...) en fichiers de textes.
Il s'agit, à partir d'une image issue par exemple de la numérisation d'une page via un scanner, d'extraire le contenu textuel.
Pour les personnes désirant récupérer, sans avoir à les recopier, d'anciens documents papier, une solution OCR est installée sur les configurations pédagogiques des PC (du secondaire).

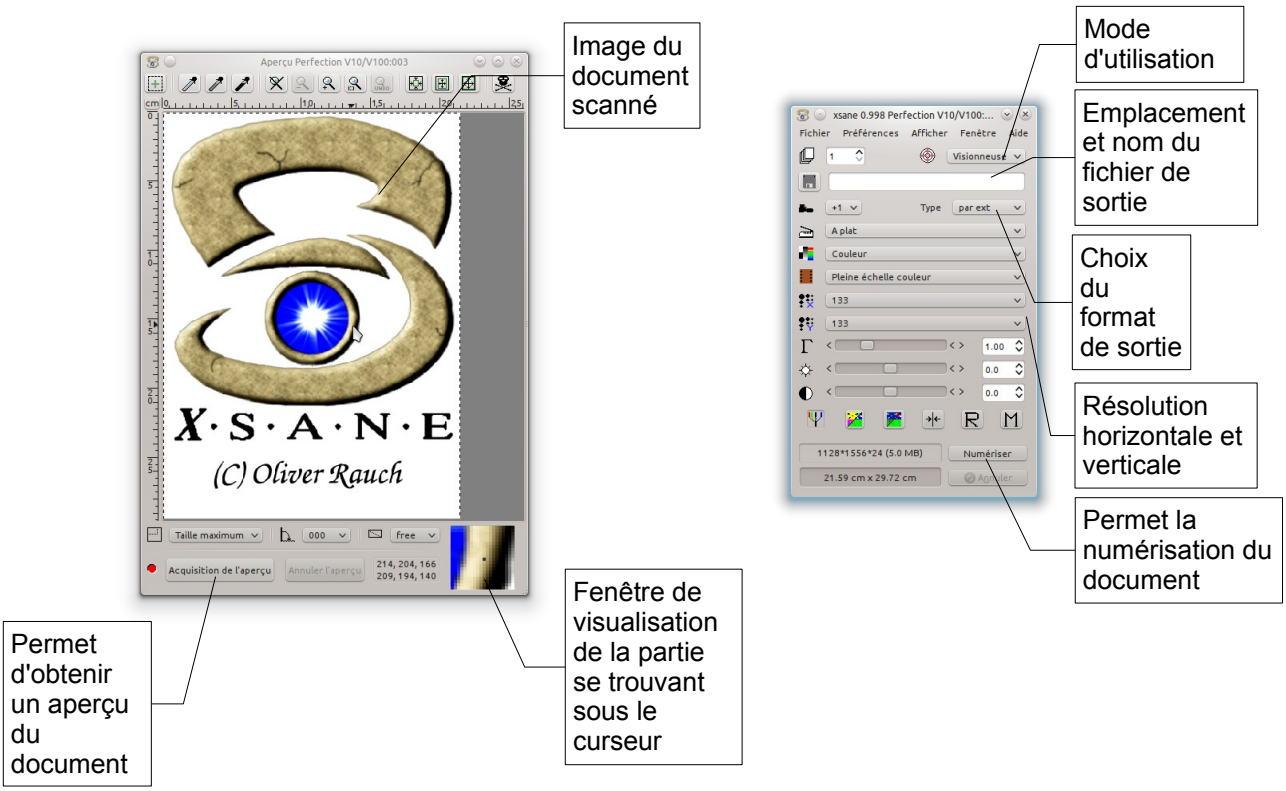
Utilisation

Sur un PC équipé d'un scanner, dans le menu de démarrage des applications qui se trouve en bas à gauche de l'écran (icône) :
sélectionnez ***Graphisme***
puis ***Xsane Image scanning program***.
Ce logiciel permet de scanner un document et, en fonction des paramètres sélectionnés, de créer une image du document ou un texte utilisable avec un traitement de texte.



<u>Paramètres importants pour l'OCR</u> :

- **Mode d'utilisation** : permet de sélectionner le mode d'utilisation de Xsane, il faut choisir **Enregistrer** ;
- **Type** : permet de sélectionner le type de format de sortie, il faut choisir **TEXT** ;
- **Résolution horizontale et verticale** de l'image, prendre 300 ou, à défaut, la valeur immédiatement supérieure.



Remarques

Lors de la numérisation, il y a deux étapes ; la numérisation proprement dite puis le processus de reconnaissance des caractères, cette deuxième étape peut prendre du temps et il ne faut pas ouvrir trop tôt le fichier obtenu, car il peut être encore vide. Il n'y a pas, malheureusement, d'indication de fin de numérisation de la part du logiciel.

La fenêtre Xsane de droite dépend du scanner branché ; elle peut être légèrement différente de celle ci-dessus.

Les résolutions horizontale et verticale n'ont pas besoin d'être à une valeur très élevée, cela n'apporte pas d'améliorations notables et surtout ralentit passablement la numérisation.

Il faut bien être conscient que la reconnaissance ne sera pas parfaite à cent pour-cent ; il faudra toujours contrôler le texte obtenu pour corriger les caractères mal interprétés.

Comme le fichier de sortie obtenu est un fichier texte, les mises en pages particulières, tels les caractères gras ou soulignés, les colonnes, ... ne seront pas gérées. Il est à noter que la lecture des colonnes est gérée correctement dans la plupart des cas, c'est-à-dire que le texte sera bien lu colonne par colonne et non ligne par ligne.

Naturellement toutes les images, photos, dessins, ... ne seront pas pris en charge par la reconnaissance. Il faudra simplement scanner « normalement » le document (type PNG, JPEG ou TIFF dans le choix de l'image de sortie) et récupérer les images avec par exemple un logiciel de dessin.