Dé-duplication dans les distributions Linux

Les distributions Linux (et, en fait, la majorité des systèmes d'exploitation) sont distribuées sous la forme *d'images de disques,* aussi appelées images ISO. Concrètement, une image ISO est une collection de fichiers, encodée dans un certain format qui est techniquement simple à les graver sur un CD-Rom ou un DVD-Rom. De nos jours, l'installation de systèmes d'exploitation passe le plus souvent par des clés USB et/ou par le réseau, mais le format des images de disque est resté d'usage.

Un CD-Rom d'installation de Linux contient en réalité un système complet, prêt à démarrer sur un ordinateur et à être utilisé. On a placé, dans ce système, un programme d'installation qui permet de copier tous les fichiers de Linux sur le disque dur ; mais il est tout à fait possible de démarrer un ordinateur sur le CD, puis d'utiliser Linux sans l'installer. (Il suffit de ne pas lancer le programme d'installation.) On peut donc raisonnablement dire qu'un CD-Rom d'installation de Linux représente le contenu « typique » d'un système qui tourne sous Linux.

On peut s'attendre à trouver, dans une image ISO de CD-Rom d'installation Linux, beaucoup de *redondance.* En d'autres termes, on pense qu'il est fréquent de trouver un même fichier dupliqué à de nombreux endroits sur un même CD-Rom d'installation. Le but de ce projet est de mesurer cette redondance. Pour cela, on propose de télécharger une image ISO d'installation Linux, puis de calculer le *hash* de chaque fichier qu'elle contient. Si on utilise une fonction de hashage assez solide, comme une fonction de la famille SHA-2, on peut faire l'hypothèse que collision de hash implique fichiers identiques. Cela nous permettra donc de compter facilement les doublons.

On peut raffiner ce projet de diverses façons : en cas de collision, on peut aller vérifier que les deux fichiers sont bien identiques. On peut également envisager de hasher des blocs au sein des fichiers, afin de repérer les *parties* de fichiers en double. Si vous êtes en forme, on pourra aussi implémenter un système qui permet de dédupliquer les fichiers (ou les parties de fichiers), en créant notre propre format d'image disque.

Le projet sera de préférence implémenté en C (négociable). Il s'agit principalement d'écrire un *parser* pour le format des images ISO, puis d'implémenter une fonction de hachage de la famille SHA-2.

Réfrences.

- Wikipédia Image ISO : `https://en.wikipedia.org/wiki/ISO_image`
- Wikipédia, voir aussi : `https://en.wikipedia.org/wiki/ISO_9660`
- Doc de référence : `https://www.iso.org/standard/17505.html`
  (Payant, nous pourrons vous le fournir.)
- Wikipédia SHA-2 : `https://en.wikipedia.org/wiki/SHA-2`