# **Objectif**

L'objectif du transcodeur est de **transcoder un fichier textuel** dont les lignes peuvent avoir été encodées *avec différents charsets* en un **fichier de sortie textuel intégralement encodé dans un charset de sortie.**

Par exemple, le transcodeur produira à partir d'un fichier textuel en entrée comportant 3 lignes en ANSI et 2 lignes en IBM-850 un fichier textuel en sortie comportant 5 lignes toutes encodées en UTF-8.



Diagramme N° 1 : Objectif

# **Exigences**

Les pré-requis sont :

* le File en entrée doit être un fichier simple (pas un répertoire).
* Le fichier simple en entrée doit être un fichier textuel (pas une photo, un mp3, ...).
* Une ligne d'un fichier simple textuel en entrée doit être intégralement encodée dans un unique charset.

On exige en outre que le fichier simple textuel en sortie soit intégralement encodé dans le charset de sortie.

Toutes les lignes dans le charset de sortie doivent donc être encodées en charsetSortie.



Diagramme N° 2 : Exigences

# **Fonctionnement du Transcodeur**

## **Principe général**

Le transcodeur doit pour chaque ligne du fichier en entrée :

1. lire la ligne n du fichier textuel simple en entrée avec le charset supposé "charsetEntree".
2. Détecter avec quel charset "charset\_n" cette ligne n a été encodée.
3. Décoder cette ligne n avec son charset d'encodage "charset\_n".
4. Ré-encoder le résultat du décodage précédent avec le charset de sortie voulu "charsetSortie".
5. Copier le résultat du ré-encodage précédent à la position de ligne n dans le fichier de sortie.



Diagramme N° 3 : Principe général



Diagramme N° 4 : Principe général

## **Principe de la détection du "charset\_n"**

Le "charset\_n" est le charset avec lequel la ligne n du fichier textuel simple en entrée a réellement été encodée à l'origine.

Un fichier réputé encodé en ANSI peut en effet parfaitement contenir 3 lignes en ANSI, puis 2 lignes en IBM-850, ...

Le principe de la détection du "charset\_n" consiste à :

1. décoder la ligne n avec un charsetDecodage.
2. inspecter la chaîne de caractères résultant du décodage avec charsetDecodage caractère par caractère
3. s'assurer qu'aucun des caractères n'est un caractère "indésirable" (caractère qu'un humain n'aurait pas pu écrire au clavier).

Si aucun caractère indésirable n'est détecté, "charset\_n" est susceptible d'être charsetDecodage.

Si un caractère indésirable est détecté, "charset\_n" ne peut pas être charsetDecodage. Il faut donc recommencer la tentative de décodage avec un autre charset.



Diagramme N° 5 : Principe de la détection du "charset\_n"

# **Objets participant au transcodage**