ICA
International Council on Archives
Conseil International des Archives

ICA
Africa Programme
Programme pour l'Afrique

ICA
Training Programme
Programme de Formation

PCOM ICA
Programme Commission
Commission du Programme

Profilage des données

Alain Dubois

Profilage des données

- Définition
  - processus qui a pour objectif de collecter des informations sur les données et notamment de disposer de mesures sur la *qualité* des données et sur la *conformité* de ces dernières par rapport aux standards. Ce processus permet notamment d'évaluer les *risques* engendrés par l'intégration de ces données dans de nouvelles applications et de disposer d'une vue globale des données dans une perspective de gouvernance des données
- Profil d'un fichier
  - *Identification* du format de fichier
  - *Validation* du format de fichier
  - *Caractérisation* du format de fichier

Profil d’un fichier

- *Identification* du format de fichier
  - Processus qui permet de déterminer le format auquel se conforme un objet numérique
  - Réponse à la question suivante: *“Je dispose d’un objet numérique. Quel est son format?”*

Profil d’un fichier

- *Validation* du format de fichier
  - Processus qui permet de déterminer le niveau de conformité d’un objet numérique avec les spécifications du format de fichier dont il se réclame
  - Réponse à la question suivante: *“Je dispose d’un objet supposément de tel format. Est-ce réellement le cas?”*
  - La conformité d’un objet numérique se mesure à l’aune de trois critères:
    - Un objet numérique est *bien formé* s’il correspond en tous points aux spécifications de ce format
    - Un objet numérique est *valide* s’il est bien formé et s’il répond en tous points aux spécifications de validation du format
    - Un objet numérique est consistant si son information de représentation interne est conforme à l’information de représentation externe

Profil d’un fichier

- *Caractérisation* du format de fichier
- Processus qui permet de déterminer les propriétés significatives spécifiques d’un objet numérique enregistré dans un format donné
- Réponse à la question suivante: *“Je dispose d’un objet numérique du format X. Quelles sont ses propriétés intrinsèques?”*

Somme de contrôle

- Une somme de contrôle est une longue chaîne de caractères alphanumériques qui établit une “empreinte numérique” pour les objets numériques
- Exemple: 96b13dbbc9f3bc569ddad9745f64b9cdb43ea9ae
- Création à partir d’algorithmes utilisant les sommes de contrôle (function de hachage cryptographique), tells que MD5, SHA-1 (Secure Hash Algorithm 1), SHA-256 ou SHA-512

Registres techniques

- L’identification des formats de fichiers dépend de registres techniques
- Les registres techniques sont utilisés dans le domaine de l’archivage électronique pour que les institutions de conservation du patrimoine puissent se fonder sur des référentiels de formats, propriétés des formats et autres outils validés
- Exemple: Archives nationales anglaises (*PRONOM*)

DROID (Digital Record Object Identification)

- Développé par les Archives nationales anglaises
- *Identification* précise de plus de 1400 formats de fichier (même si l'extension du fichier est fausse ou manque)
- Informations notamment sur les éléments suivants :
  - type (fichier ou dossier)
  - nom de fichier
  - extension du fichier (y compris détection de toute erreur dans le nommage de l'extension)
  - format de fichier (nom et version du format, type MIME et identifiant PRONOM)
  - emplacement
  - taille du fichier
  - date et heure de la dernière modification
  - méthode d'identification du format (extension, signature, conteneur)
  - contenu de hachage (somme de contrôle (cf. algorithme MD5))

JHOVE (JSTOR/Harvard Object Validation Environment)

- Développé par JSTOR *(Journal Storage* (bibliothèque numérique de 2000 journaux fondée en 1995)) et la bibliothèque de l'Université d'Harvard
- But : développer un outil pour l'*identification* (de quel format s'agit-il?), la *validation* (est-ce bien un fichier de tel format?) et la *caractérisation* (quelles sont les propriétés intrinsèques de tel format?) des formats de fichier
- Outil qui permet d'analyser la conformité des fichiers GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, MP3, XML et ZIP

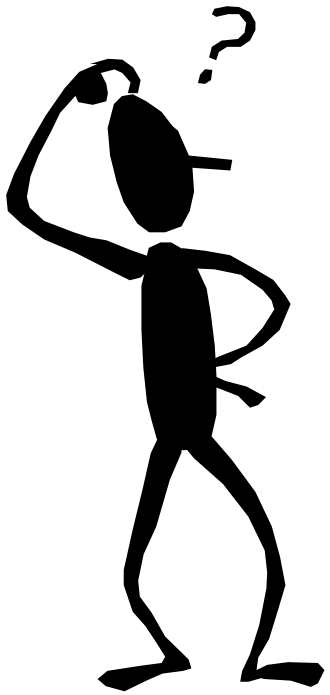
Format Description Service

- Service de description de format développé par le *Dark Archive In The Sunshine State* (DAITSS)
- Outil fondé sur la combinaison de DROID (version 3.0) et de JHOVE (version 1.11)
- Outil permet en effet
  - d'*identifier* le format d'un fichier (en utilisant DROID)
  - de *valider* le format de fichier (en utilisant JHOVE)
  - de *caractériser* le format de fichier (en utilisant JHOVE)
- Outil génère ensuite un fichier construit selon PREMIS
- Lien : http://description.fcla.edu

Atelier pratique

- Utilisez DROID, JHOVE et DAITSS

Merci!



Questions?

ICA Africa Programme
Programme pour l'Afrique

ICA Training Programme
Programme de Formation