Fiche de cours

Contenus

Le profilage des données est un 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.

En somme, le profilage des données est essentiel pour assurer l'intégrité des données et, partant, leur valeur de preuve. Dans le domaine de l'archivage électronique, il se concrétise à travers les registres techniques, le profil d'un fichier et la somme de contrôle.

1. Les registres techniques

L'identification des formats de fichiers dépend de registres techniques qui proposent de décrire les formats de fichiers utilisés dans le monde. L'objectif est de permettre aux institutions en charge de la préservation numérique de disposer de référentiels validés sur lesquels ils peuvent se fonder. Le registre technique utilisé aujourd'hui par la plus grande partie des services d'archives est le registre PRONOM, développé et maintenu par les Archives nationales anglaises (https://www.nationalarchives.gov.uk/PRONOM/).

2. Le profil d'un fichier

Lorsque l'on souhaite déterminer le profil d'un fichier, on va chercher à

- *identifier* le format du fichier (quel est le format de ce dernier ?)
- *valider* le format de fichier (ce format correspond-il aux spécifications du format de fichier dont il se réclame ?)
- *caractériser* le format de fichier (quelles sont les propriétés intrinsèques d'un objet numérique se réclamant de tel format ?)

Il existe actuellement plusieurs outils qui permettent de déterminer le profil d'un fichier :

- DROID (Digital Record Object Identification). Cet outil permet d'*identifier* près de 1400 formats de fichiers, quand bien même l'extension du fichier est fausse ou manque (https://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/)
- JHOVE (JSTORE/Harvard Object Validation Environment). Cet outil permet d'*identifier*, de *valider* et de *caractériser* les formats de fichier suivants : GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, MP3, XML et ZIP (https://openpreservation.org/technology/products/jhove/).
- DAITSS (Dark Archive in the Sunshine State). Cet outil en ligne permet d'*identifier*, de *valider* et de *caractériser* les formats de fichier suivants (http://description.fcla.edu). Il permet de générer un fichier XML contenant les résultats de l'analyse.

3. La somme de contrôle

La somme de contrôle est une longue chaîne de caractères alphanumériques qui établit une empreinte numérique pour les objets numériques. L'empreinte numérique se calcule à partir d'algorithme (MD5, SHA-1, SHA-256 ou SHA-512).

Types d'évalution

- Essai réalisé individuellement
- Présentation d'un dossier en groupe

Ressources bibliographiques

Introduction à la problématique

File-format analysis tools for archivists, 2016 (https://lwn.net/Articles/688396/).

Shala Lavdërim, Shala Ahmet, Technology and Culture TECIS 2016*, File Formats – Characterization and Validation*, 17th IFAC Conference on International Stability (http://www.sciencedirect.com/science/article/pii/S2405896316324880).

TÖWE Matthias, GEISSER Franziska, SURI Roland E, iPRES 2016, *To Act or Not to Act. Handling File Format Identification Issues in Practice*, 2016.

Pronom

Siegfried – a PRONOM-based, file format identification tool, 2014. (http://openpreservation.org/blog/2014/09/27/siegfried-pronom-based-file-format-identification-tool/).

DROID

Hoppenheit Martin, *Minimizing the DROID signature file*, 2017 (https://martin.hoppenheit.info/blog/2017/minimizing-the-droid-signature-file/).

JHOVE

Lindlar Michelle, TunnatT Yvonne, *How valid is your validation? A closer look behind the curtain of JHOVE*, dans *12th International Digital Curation Conference: Upstream, Downstream: embedding digital curation workflows for data science, scholarship and society*, 2017.