Formats de données et pérennité

C.Huc, N. Lormant, J. Masanès, D. Courtaud

sommaire

- Aptitude des formats de données au regard de la pérennité
  C. Huc
- un exemple d 'étude pour le format PNG de représentation des images raster
  N. Lormant
- les initiatives en cours au plan international
  J. Masanès
- les travaux au niveau du groupe PIN - vers une base XML
  D. Courtaud

Introduction

- On propose :
  - un glossaire
- Démarche
  - consolider la méthodologie d 'analyse proposée
  - définir le format des fiches d 'analyse
  - expérimenter le travail sur quelques formats - organiser ces résultats dans une base de données
  - organiser le relais au niveau international

Glossaire (1/3)
Définitions du Modèle ISO OAIS

- ***Donnée :*** *représentation formalisée de l'information permettant la ré-interprétation future de cette information à des fins de communication, d'analyse ou de traitement.*
- **Information** : tout type de connaissance pouvant être échangée. Lors d'un échange, l'information est représentée par des données. Par exemple : une chaîne de bits (les données) accompagnée d'une description permet d'interpréter cette chaîne de bits comme des nombres représentant des observations de températures mesurées en degrés Celsius (c'est l'information de représentation).
- **Information de Représentation** : information qui établit une correspondance entre un Objet données et des concepts plus significatifs. Exemple : le code ASCII qui décrit comment une séquence de bits est convertie en caractères.

De la donnée à l 'information
l 'information de représentation

- L'information est toujours représentée par des données
- En général, on peut dire que "les Données interprétées à l'aide de leur **information de représentation** génèrent de l'information"
- Pour que la pérennité d 'un **Objet Information** soit garantie, il est indispensable que l 'archive préserve **l'Objet Données** et l **'Information de Représentation** qui lui est associée.

Glossaire (2/3)

- **Format (de données)** : *(il est assez difficile de définir ce qu'on entend par là bien que le terme soit très répandu Est-ce qu'il existe une définition normalisée quelque part ?)*

Le format de données peut être défini par l'ensemble des règles et algorithmes permettant d'organiser l'information dans un objet numérique.

Par ex

- spécifier le codage des couleurs des pixels d'une image, définir un algorithme de compression des données et l'organisation de ces données dans un fichier, (formats PNG, TIFF…)
- spécifier l'organisation et la structuration d'informations textuelles à partir de l'encodage élémentaire des caractère (formats SGML, XML)
- définir comment les quatre informations élémentaires que sont la mantisse (nombre entier positif), l'exposant (nombre entier positif) et le signe de l'exposant et le signe de la mantisse (caractères + et -) sont organisés pour représenter un nombre réel sous forme numérique(cf. standard ANSI/IEEE 754-1985)

Glossaire (3/3)

- **Norme** : Ensemble de règles approuvées par des instances officielles en charge de la normalisation. Elles offrent une certaine garantie de stabilité et de pérennité
- **Standard** : les standards sont définis par des groupes privés, en général industriel ou commerciaux (par ex les standards PostScript ou PDF de Adobe). Ces groupes peuvent aussi être collégiaux comme le W3C et le consortium Unicode

Condition de base

Un format de données ne peut être acceptable pour la pérennisation de l'information que s'il n'introduit aucun élément manquant dans l'Information de Représentation (au sens du Modèle OAIS) indispensable pour passer du train de bits à une information intelligible.

- **L'Information de Représentation constitue un réseau d 'informations de représentations imbriquées**
- **le format de données ne constitue de l'un des éléments de ce réseau**
  - **exemple : suite de nombre codés en ISO646 (texte brut)**

123.345643.54654.4576.765.
1 23.34 5643.5 4654.4 5 76.765.

Longueur des champs / norme ISO 646 / Signification des champs

Condition de base

- La condition de base vise à éliminer les formats non publiés qui ne peuvent en aucun cas être retenus pour un archivage long terme
  - par exemple les fichiers au format Word 97 créés par la suite Microsoft Office). Toute solution qui ne respecterait pas cette condition ne peut être qu'une solution d'attente valide à court terme.
- La description du format de données **ne représente qu'une partie de l'information de représentation**.
  - Dans le cas simple d'un fichier au format texte ne contenant que des nombres codés avec le jeu de caractère ISO 646, la signification des nombres présents dans le fichier doit par ailleurs être renseignée. Cette information indispensable ne dépend pas du format.

La pérennisation de l'information par le moyen de techniques d'émulation n'est pas rejetée a priori. Nous considérons qu'elle est en dehors du champ de réflexion sur les formats de données puisqu'elle repose sur une toute autre logique.

critère 1 : le format doit pouvoir représenter toute l'information à pérenniser

Aptitude du format à représenter l'information dans toute sa richesse et sa complexité.

- Certains formats sont réducteurs. La représentation d'une information textuelle dans un format de type texte pur (TXT) dans lequel le texte à plat est codé avec le jeu de caractère ISO8859-1 (ISO Latin 1) conduit à mettre au même niveau et sans distinction particulière, tous les éléments structurants du texte
- L'évaluation du format ne peut donc pas être faite indépendamment du type d'information qu'il va représenter
  - l'usage du format PNG ou TIFF pour représenter une image raster peut être un bon choix,
  - l'usage de ces même formats pour conserver l'aspect original d'un document textuel peut également être retenu,
  - ce même choix sera réducteur si l'objectif est de préserver l'information contenue dans un document textuel initialement créé sur un ordinateur

critère 2 : format normalisé

Chaque fois que le choix sera possible, on recommandera l'usage de formats normalisés et on évitera l'usage de formats propriétaires ou d'éléments propriétaires au sein d'un format normalisé. On favorisera également l'usage de formats définis par des groupes collégiaux ouverts (par exemple le W3C ou UNICODE).

- n'interdit pas formellement l'usage de formats propriétaires publiés mais néanmoins le déconseille. Il est clair que la politique commerciale du propriétaire d'un format peut changer au gré du marché et poser à l'archive des problèmes inattendus. Par ailleurs, les spécifications du format peuvent évoluer de façon régulière et parfois très rapidement en fonction des besoins du propriétaire.
- nous devons cependant prendre en considération le fait qu'au delà d'une certaine période, un format propriétaire tombera dans le domaine public. Ce format ne sera pas normalisé pour autant mais son usage ne sera plus soumis à des restrictions juridiques particulières.

critères complémentaires

- Outils et facilités de création et de manipulation : disponibilité, coût…
- Complexité/simplicité du format
- Disponibilité ou faisabilité de logiciels de contrôle de conformité