

Problèmes d'encodage de fichiers xml-ead lors de leur intégration dans Arkothèque

Camille Hodoul [chodoul@1egal2.com]

Envoyé : mercredi 13 juillet 2022 10:58

À : Le Goff Nolwenn; Franck Bernardet [franck@1egal2.com]

Cc : Veschambre Vincent; Gennuso Emanuela; Chabrier Margaux; Dupuis Murielle; Marion 1egal2 [marion.ristorcelli@1egal2.com]

Bonjour,

Suite à l'intégration de vos inventaires, mes collègues m'ont signalé un problème de rétroconversion sur certains IRs, notamment 069266FRAC_2Fi_Villeurbanne_en_Cartes_postales . Les IRs ont bien été récupérés mais des caractères indésirables sont présents dans certains champs (par exemple bioghist). Il est à noter que cela pouvait déjà être constaté en v6.

En regardant le contenu du fichier xml, voici un exemple de ce qui semble poser problème :

```
```xml
2 Fi 385 - 2 Fi 389

```
```

Spécifiquement

```
```xml


```
```

Ceci est techniquement un texte correct en XML, mais évidemment pas le texte attendu.
Le texte devait à l'origine être ``<BR>``, puis a été transformé par des outils intermédiaires :

- ``<BR>`` (à noter qu'il s'agit d'une balise HTML incompatible avec XML-EAD)
- les ``<`` et ``>`` sont encodés en leurs entités XML respectives ``<`` et ``>`` → ``<BR>``
- les ``&`` sont à leur tour encodés en l'entité XML correspondante ``&`` → ``&lt;BR&gt;``
- lors de la rétroconversion, Arkothèque reconvertit les ``&`` en ``&`` comme attendu → ``<BR>``

Ceci est donc le résultat de plusieurs problèmes en amont :

- l'utilisation de ``<BR>`` au lieu de ``<lb />`` (l'équivalent correct XML-EAD)
- le "double encodage" des ``&`` en entités XML.

Ces problèmes ont pu se produire dans votre SIA ou tout autre outil intermédiaire ayant participé à l'encodage de cet IR (éditeur XML, etc.), lors de la saisie et/ou de l'export en XML.

Idéalement, le correctif serait à apporter dans votre SIA ou, en tout cas, en amont de leur intégration dans Arkothèque.

Notre outil de pré-traitement nous permet habituellement de gérer ce genre de problèmes de validité ou d'interopérabilité entre systèmes, cependant cela sera difficile (voire, a priori, impossible) dans ce cas précis pour les raisons suivantes :

- Les textes ``<BR>`` et ``&lt;BR&gt;``, bien que n'ayant pas de sens, sont du contenu XML valide.
- Il est donc, a priori, impossible pour un programme de déterminer si une chaîne qui ressemble à ``&lt;BR&gt;`` est un mauvais encodage qui doit être corrigé ou une valeur valide, indépendamment

du contexte.

- Fournir le contexte au programme reviendrait à lister exhaustivement les cas d'erreur comme `<lb/><lb/>`, ce qui implique de décider au cas par cas pour le contenu de tous les IRs et ne serait pas plus efficace (ni moins risqué) qu'un rechercher-remplacer dans les fichiers.
- Un correctif global sans suffisamment de contexte serait contre-productif car le risque de faux-positif serait très haut ce qui résulterait en une baisse de la qualité de l'encodage.

En conclusion je recommande, par ordre de préférence :

1. Corriger dans l'outil qui fait le premier export XML, pour enlever les balises invalides et éviter les doubles-encodages d'entités XML. Cela me semble être la meilleure solution car elle résultera en la meilleure qualité des données sur toute la chaîne.
2. Corriger en "rechercher/remplacer" dans les fichiers XML (nous pouvons vous aider à lister les cas problématiques).
3. Développer un correctif spécifique dans notre outil de pré-traitement, qui fonctionnerait aussi sur la base d'une liste de cas à corriger.

Cordialement,

--

Camille HODOUL
PROFILS - 1 égal 2
900 Chemin de l'Aumône Vieille
Parc de l'Angevinière Bat5
13400 Aubagne
<http://www.legal2.com>
<http://www.arkotheque.fr>