OpenEdition

Cahier des charges OTX

Rendre OTX extensible et ajouter la gestion des docx

Date	Agent	Description de modification
13/09/2018	Roland Haroutiounian	Rédaction de la première version
24/09/2018	Roland Haroutiounian	Ajout d'informations sur la gestion des configurations

Objet du document

Ce document est le cahier des charges d'une prestation à réaliser pour OpenEdition

Contexte

OpenEdition Center est une infrastructure nationale de recherche pour la publication des Sciences Humaines et Sociales en Accès Ouvert (open access) sur 4 plateformes web : OpenEdition Books, Journals, Calenda et Hypothèses). La fréquentation atteint plusieurs millions de visites mensuelles. La plateforme accueille aujourd'hui près de 500 revues, 3000 carnets de recherche, 30 000 annonces d'événements scientifiques et 5 000 livres. Des milliers d'utilisateurs éditent et publient du contenu sur OpenEdition.

Les applications se basent principalement PHP et Python, ainsi que MySQL. Une partie des logiciels est publiée sous licence libre sur https://github.com/OpenEdition/ notamment Lodel, le CMS open source de publication scientifique d'OpenEdition, et OTX, pour convertir des documents électroniques transmis au format *doc* ou *odt* vers le format XML TEI qui est le standard utilisé dans l'édition électronique, en utilisant le format *odt* produit par Libreoffice comme intermédiaire; OTX est utilisé pour alimenter Lodel en contenus structurés.

Objectif

Afin de pouvoir répondre aux exigences technologiques actuelles et à venir (prise en charge des fichiers *.docx*, montée en version des outils annexes et des serveurs), il est devenu nécessaire de faire évoluer le logiciel OTX pour qu'il puisse récupérer les bonnes informations dans les *.odt* produit par la version de Libreoffice que nous utilisons (version 5.x.x).

L'objectif est donc double :

- Résoudre les problèmes de récupération de certaines données nécessaires au traitement d'un document : titre et sous-titre
- Ajouter la prise en charge du format *docx*

Etat de l'existant

OTX est un projet open source qui est disponible sur le compte github d'OpenEdition: https://github.com/OpenEdition/OTX. Il est inclus dans le processus de chargement de documents dans le logiciel Lodel (https://github.com/OpenEdition/lodel) également open source. Ces deux applications sont écrites en PHP.

Il accepte en entrée les formats *.doc* et *.odt* et produit des fichiers XML-TEI. Le principe est de convertir les fichiers *.doc*, de format propriétaire fermé, en *.odt* (on utilise LibreOffice pour cela) qui sont composés de fichiers XML, que l'on peut analyser et dont on récupère les informations pour construire le format XML-TEI.

Une feuille de styles est utilisée dans Microsoft Word pour permettre la construction structurée des documents. Ces styles constituent la base pour les conversions ultérieures. La particularité est qu'ils sont traduits dans la langue de l'utilisateur.

Une fois les fichiers *odt* générés, ceux-ci sont ensuite transformés en TEI qui est remodelée pour être comprise par un modèle éditorial correspondant à, par exemple, la revue dont fait partie le document.

Dans OTX, les configurations sont gérées dans une classe OTXConfig (qui est placée à la racine du projet). Cette classe lit un fichier xml de configuration (dont un exemple est donné dans le dépôt github).

Description du problème

Suite à une montée en version de Libreoffice, utilisé pour réaliser la conversion du format *doc* vers le format *odt*, les titres et sous-titres des documents ne sont plus présents dans le fichier *meta.xml* qui fait partie de l'archive *odt* produite. Cela pose problème pour le traitement des données du document dans le système d'information d'OpenEdition.

Il faut donc récupérer ces méta-données dans le fichier *content.xml* en analysant les noeuds contenant ces informations.

Toutefois, en fonction des documents, de la langue utilisée par le traitement de texte, de la version de LibreOffice qui fait la conversion, on obtient des structures variables dans les *.odt* produits. Certaines informations, comme le titre par exemple, deviennent difficiles à détecter. C'est à ce niveau qu'une modification du fonctionnement est à considérer.

Exemple avec le titre du document

Dans ce fichier, le titre du document se situe dans un noeud XML du type :

```
<text:p text:style-name="P7">
        Complejidades conceptuales sobre el colonialismo y lo postcolonial
</text:p>
```

Afin de comprendre qu'il s'agit du titre, il faut, toujours dans le même fichier, aller voir la définition de ce style "P7", ce qui donne :

<style:style style:name="P7" style:family="paragraph" style:parent-style-name="título" style:master-page-name="Standard">
<style:paragraph-properties fo:margin-top="0.212cm" fo:margin-bottom="0.212cm" loext:contextual-spacing="false" style:page-number="169"/>
</style:style>

L'élément distinctif est ici l'attribut *style:parent-style-name*. Celui-ci est alors dépendant de la langue du traitement de texte utilisé.

Modifications attendues

La modification envisagée est la suivante :

- Construire un mapping des styles où l'on fait correspondre le nom du style traduit avec un terme générique qui servira de base. Par exemple : título => title

- Modifier le code d'OTX pour utiliser ce mapping afin de récupérer les champs manquants (titre et sous-titre). La fonction qui permet de transformer un odt en TEI est ***lodelodt*** et se situe dans le fichier : ***server/otxserver.class.php*** (l. 487)

Il faut que l'on dispose d'un mapping centralisé qui soit facilement extensible au gré des ajouts de langues et/ou de modification des feuilles de styles utilisées dans Microsoft Word pour les documents destinés à OpenEdition. Ce mapping pourrait être ajouté dans la classe OTXConfig dans une propriété dédiée.

Actuellement, OTX utilise en entrée le *doc* et l'*odt*. Il faudra qu'il gère le *docx* qui s'intégrera dans la même chaîne, soit en passant par le format intermédiaire *odt*.

Des documents réels seront fournis afin de donner un aperçu des cas possibles : *doc*, *docx*, *odt*, dans plusieurs langues, afin de faciliter l'analyse nécessaire à l'élaboration du mapping et du traitement à appliquer.

Déroulement de la prestation

Eléments fournis en entrée:

- des documents à tester (différentes versions et versions linguistiques de Word, formats DOC, DOX et ODT)

- le code source est sur github, le prestataire pourra forker ce code et créer une branche correspondant à la prestation

- un serveur de développement Lodel associé à OTX sera utilisé pour permettre à des utilisateurs finaux de tester sur des documents réels, en plus des tests qu'effectuera le prestataire avec une instance locale d'OTX qu'il aura pu installer

Organisation:

Interlocuteurs : Roland Haroutiounian sera l'interlocuteur technique du prestataire, Bianco Tangaro sera l'interlocutrice représentant les Utilisateurs.

Une personne du secteur Edition sera également associée au suivi de la prestation en tant qu'utilisateur qui validera le logiciel livré.

Une réunion de lancement permettra de caler l'organisation et de valider la solution technique proposée par le prestataire.

Une réunion finale permettra de valider le livrable (code commenté, tests et documentation).

Les réunions pourront avoir lieu à distance.

Demande de devis

La proposition (technique et financière, annexe présentant les références du candidat) sera adressée par mail à roland.haroutiounian@openedition.org et bianca.tangaro@openedition.org avant le 19 octobre 2016 à midi.

La meilleure offre sera retenue selon les critères suivants:
- compétences techniques
- pertinence de la proposition et compréhension du besoin
- clarté de la proposition
- prix, délai