Prestation de conseil
« Archivage à long terme du format TEI »

Contexte

Le TGE-Adonis (UPS 2916 en cours de fusion avec la TGIR corpus pour former la nouvelle TGIR Huma-Num) propose à sa communauté d'utilisateurs un service d'archivage à long terme basé sur les ressources et l'infrastructure du CINES (Centre Informatique National de l'Enseignement Supérieur http://www.cines.fr).
La communauté des utilisateurs de la TEI (Text Encoding Initiative - http://www.tei-c.org) est très développée en France dans le domaine des Sciences Humaines et Sociales et de nombreux corpus sont encodés en TEI (e.g. corpus de la presse quotidienne de l'ATILF, ressources du CLEO, Bibliothèque Virtuelle des Humanistes de Tours, Base de Français Médiéval à Lyon, Ressources de l'école des chartes etc.).
Actuellement, il est possible d'archiver le format TEI au CINES en le considérant comme un fichier XML pur, la validation du format se faisant à l'aide d'un schéma XML classique.
La difficulté est que dans ce cas, nombre d'informations importantes liées à l'utilisation de la TEI dans le cadre d'un document spécifique, sont alors perdues. Or le but de la pérennisation est de pouvoir restituer le document numérique dans son environnement le plus fidèlement possible sur le long terme. Plus l'on dispose d'informations sur la version originale, meilleure en sera la restitution.

Prestation : Définir une version archivable de la TEI

Le TGE-Adonis souhaite étudier, en collaboration avec le CINES et avec le soutien de la communauté TEI, la possibilité d'un archivage enrichi des documents en format TEI. Le projet Monk (http://monkproject.org/) a réalisé une étude sur des documents en format TEI provenant de différentes sources et a mis en évidence une grande variabilité de la TEI utilisée. Par ailleurs, une centaine de balises sont employées en moyenne pour le format TEI.
L'idée serait de poser les bases d'un format minimal d'archivage (un format TEI-A en somme à l'instar du PDF/A) qui serait acceptable pour l'archivage à long terme de la TEI : ce format serait un compromis entre les possibilités de vérification effectuées par le CINES et le foisonnement naturel de la TEI.
Cette prestation se fera en concertation avec les personnels du CINES en charge des formats, de la communauté TEI et des utilisateurs déjà candidats à l'archivage pérenne. Il sera nécessaire d'organiser les réunions avec ces partenaires suivant des modalités à définir avec le TGE-Adonis.
Le résultat final attendu se compose :

- d'une description précise du ou des formats acceptés pour un document à archiver ;
- d'une description des documents d'environnement nécessaires à la bonne compréhension du document (e.g. « guidelines » à utiliser) ;
- de la définition des procédures de vérification.

Activités à réaliser et livrables

Préparer et animer les différentes réunions de travail

Au minimum 5 réunions sont à organiser avec les partenaires pilotes et le CINES

Livrables : Ordres du jour et compte-rendu des réunions

Assurer le lien avec la communauté internationale de la TEI

Identifier les utilisateurs de la TEI qui sont intéressés et/ou qui travaillent sur l'archivage de données en format TEI
Faire le lien avec les instances de la TEI (TEI Board et conseil scientifique)

Livrables : Compte-rendu des réunions ou échanges avec la communauté

Document final

Le document final comportera

- La description précise du ou des formats acceptés pour un document à archiver ;
- La description des documents d'environnement nécessaires à la bonne compréhension du document ;
- La définition des procédures de vérification.

Livrables : Document final

Suivi du processus au CINES

Préparer des jeux de données représentatives pour les tests et suivre les procédures de vérifications

Livrables : jeux de documents test et compte-rendu des essais

Nombre de jours estimés pour la prestation

- 5 jours pour les réunions
- 10 jours pour les actions de coordination et de rédaction

Modalités de paiement

30% à la signature du contrat
30% en milieu de mission
40% à la livraison du rapport final.

Dates de début et de fin de la prestation

La prestation commencera dès l'acceptation de l'appel d'offre et devra être intégralement réalisée avant le 31 décembre 2013.

Annexe : Compte Rendu de la réunion préparatoire du 15/02/2013 qui s'est tenue au CINES

Introduction

Le format TEI est un format XML utilisé pour la description et l'échange de textes électroniques.
Il comporte environ 500 balises mais une personnalisation plus fine du format est toujours possible : on peut utiliser un sous ensemble des balises existantes et en définir de nouvelles.

Un changement de version important a eu lieu en 2001 lors du passage de la version P4 à la version P5 puisque la technologie est passée de SGML vers XML.
Le passage de P4 vers P5 a également facilité la personnalisation qui devient automatisable en grande partie.
La version P4 n'est plus supportée depuis fin 2012, et des outils de migration existent.

Depuis la version P5, le document ODD contient la description d'un ensemble de documents : il décrit les balises employées dans une collection ou un ensemble de documents et la manière de les employer.
L'idée, pour l'archivage, serait donc d'imposer a minima une version P5 pour les documents associés à un document ODD.

Processus d'évolution de la TEI

La communauté TEI est composée :
- Des utilisateurs de la TEI, dont le nombre est difficile à évaluer
- Du consortium TEI qui finance le fonctionnement de la TEI et dispose de droits de vote
- Du TEI Board qui s'occupe de l'administration
- Du Conseil scientifique qui est élu pour 2 ans et s'occupe de l'évolution technique de la TEI

Les "guidelines" sont mises à jour environ deux fois par an.
Les "guidelines" contiennent les définitions des schémas de la TEI ainsi que la description du contexte d'utilisation.
Les "guidelines" sont exprimés en format ODD, qui est lui-même une personnalisation de la TEI.
Tout comme un document TEI, un fichier ODD fait donc lui aussi référence à une version de "guidelines".

La TEI intègre également les propositions de groupes extérieurs travaillant sur des sujets particuliers : les "Special Interest Groups" (e.g. édition génétique)
A noter qu'une partie de la TEI a été normalisée à l'ISO.

Proposition de méthodes de validation

Plusieurs niveaux de validation sont proposés pour tester un document :

- Vérification des balises du document TEI, classiquement en utilisant le schéma XML de la version de TEI utilisée (basée sur une version de guideline)
- Vérification de la validité du schéma XML du document ODD en utilisant le schéma XML de la version de l'ODD utilisée (basée sur une version de guideline)
- Comparaison de l'ODD généré à partir du document par rapport à l'ODD fourni
- Utilisation de feuilles de style pour effectuer une transformation "standard" (à définir) du document TEI

D'autres pistes de validation sont envisageables :
- L'évaluation du contenu des attributs n'est pas réalisée de manière systématique dans la TEI, mais le type de ces attributs peut être défini dans le document ODD, ce qui permettrait d'envisager une validation de ce type.
- Les balises sont regroupées dans des classes. Ces classes permettent de définir à quel niveau dans la structure du document cette balise doit se situer (e.g. dans un paragraphe). Les classes et sous-classes permettent également de qualifier cette balise en l'associant à des balises de même type/signification (e.g. appartient à l'ensemble des balises contenant une description bibliographique). Il serait donc possible d'utiliser les classes pour effectuer des vérifications plus fines.

Enfin, il est à noter que la plupart des projets utilisent l'un des odd pré-existants qui sont fournis par la TEI.

Les outils disponibles

Des outils de validation sont déjà disponibles dans la communauté TEI, il reste à les évaluer dans le contexte d'utilisation du CINES.
En ce qui concerne les documents de type ODD, le logiciel ROMA est un outil d'aide à la définition de ODD.
D'autres outils, basés sur les technologies standard associées à XML (XSLT/CSS) seront peut-être à développer en fonction des besoins.

Migration éventuelle du format

Le CINES mène une veille pour disposer des éléments de décision pour la migration des formats stockés dans leur système.
Parmi les critères retenus, la disponibilité des outils nécessaires à la restitution des documents ainsi que la péremption de la version employée (e.g. version plus utilisée, plus d'outils disponibles, alertes de type "économique" pour les outils qui deviennent commerciaux ...).
Il sera utile de préciser les aspects à surveiller pour le format TEI dans cette optique.