OUTIL « EADIFICATION » ET « EXCELLIFICATION » : MODE DE FONCTIONNEMENT

Arnaud Millereux, informaticien MSH Dijon

Considérations générales

Dans le cadre des projets de recherche avec une dimension archivistique, il est préférable d'utiliser des formes normalisées pour l'organisation du matériel documentaire. Parmi cela, le standard de description ISAD(G) et son encodage XML-EAD sont les plus en mesure de répondre aux besoins du traitement archivistique qui, lui, suppose une hiérarchisation de l'information du général au particulier. Le schéma EAD est parfaitement compatible avec les pratiques archivistiques ce qui nous incite à adopter et à inciter les ingénieurs des laboratoires à adopter ce système d'encodage pour décrire leurs archives.
Puisque les logiciels d'édition et le langage XML son difficiles d'accès pour des non-initiés et en absence d'un logiciel de saisie avec des interfaces fonctionnelles, la MSH a développé dans le cadre du projet ANR PAPRIK@2F ou outil destiné à simplifier le travail des archivistes associés au projet : la structuration des données dans un tableur et sa transformation en fichier XML-EAD.
Ce document présente et décrit cet outil. Il est le fruit de la collaboration entre l'archiviste et l'informaticien du projet (Victor Lagarde et Arnaud Millereux).

EADification

« L'EADification » ou comment générer un fichier EAD à partir d'un tableur ?

Initialement, cet outil est développé pour contribuer à générer un document XML conforme à la spécification EAD à partir de données présentées de façon tabulaire. Ainsi, une personne ne connaissant pas la spécification peut facilement produire un document compatible et apprendre à partir du résultat obtenu. L'approche "tableur" peut apporter un gain dans la mesure où l'utilisateur ne se préoccupe pas de la partie encodage une fois le modèle défini.

Des solutions basées sur la mise en place d'un modèle XML avant l'exportation directe des données en XML via Microsoft Excel existent. Il suffit de définir un modèle qui servira de *mapping* entre les données tabulaires et le fichier XML à obtenir en sortie. Une autre possibilité est d'utiliser conjointement des feuilles de style XSL pour la transformation des données XML natives provenant du tableur. Nous avons choisi la possibilité d'avoir accès à une gestion plus fine des données en utilisant la force du langage de programmation embarqué dans le tableur : Microsoft Visual Basic for Applications (VBA). Ce langage a l'avantage certain d'être directement inclus dans tous les logiciels de la suite Microsoft Office et facile à apprendre pour ceux qui veulent étendre les fonctionnalités déjà présentes.

Hypothèses :

La magie repose sur une idée toute simple. Si l'on dispose de données métier correctement formatées sous forme tabulée, il existe une solution automatique pour produire un document hiérarchisé au format XML respectant la spécification souhaitée. Chaque colonne contient une donnée particulière qui correspondra à une balise XML correspondant à la spécification sortante choisie.

Les données des cellules peuvent être de deux natures :

- « monovaluée » : chaque cellule contient une donnée qui sera recopiée de manière brute sans traitement dans la balise correspondante

- « multivaluée » : chaque cellule contient un jeu de données qui, après traitement, pourra renseigner :

    - soit, un ensemble de valeurs de balises

    - soit, un semble de nuplets nom_atribut-valeur_attributs-valeur_balise

Données en entrée :

Un ensemble de données convenablement disposées sur une ou plusieurs feuilles selon un modèle défini au préalable. Dans une nouvelle version, il est tout à fait possible de créer une couche d'abstraction supplémentaire afin de rendre le système opérationnel parfaitement générique.

Données en sortie :

Au clic sur le bouton "EADification", nous obtenons un fichier texte encodé en UTF-8 et respectant le format XML EAD selon le modèle défini.

Résultat :

Ce fichier peut être publié directement dans une application compatible avec ce standard alors qu'aucune ligne XML n'a été saisie par l'utilisateur. Rien n'empêche l'utilisateur averti d'apporter ses propres modifications au fichier généré automatiquement, d'autant plus qu'il existe aussi un outil qui permet de rebasculer du format XML EAD vers Microsoft Excel ! (sous certaines conditions). Ce procédé est mis en œuvre dans le cadre du projet ESMR de la MSH de Dijon.

Exemple :

Prenons l'exemple du fonds 517/1 traité au cours du projet ANR Paprik@2F. Après extraction des données issues de la base 4D au format texte brut, les données sont importées dans le tableur par l'archiviste. On définit le rôle de chaque colonne en entête pour plus de lisibilité.

Cote RGASPI	Cote 1	Cote 2	Cote 3	Titre	*ScopeContent*	**Nom**	**Géo**	**Sujet**
517/1/0022	517	1	22	Affiches, tracts du PCF, du Comité français de l'ICJ, des organisations syndicales révolutionnaires sur les meetings pour la défense de la Russie soviétique, en relation avec les élections parlementaires et d'autres évenements en France.	Coupure de presse signé par Un Groupe d'Anciens Combattans et de Mutilés intitulé "Appel aux Soldats et aux Marins Français" publié le 12 octobre 1920. Affiche du Comité d'Action contre le Fascisme mettant en avant le slogan "Classe Ouvrière, défends-toi". Affiche de L'Internationale Communiste des Jeunes dont le slogan est "L'ennemi se trouve dans votre propre pays". L'affiche est en français et en allemand avec un dessin.	Sadoul,Jacques;;Marly;;Cachin,Marcel;;Daudet;;Liebknecht,Karl	Silésie(Allemagne);;Cilicie(Adana;Turquie);;Turquie;;Anatolie(Turquie);;Verdun(Meuse;France)	Guerre;;Bloc National;;Traité de Versailles;;PGM;;Bataille de Verdun;;Fascisme
517/1/0023	517	1	23	Journal La bonne guerre n°49 édité à Tours par Jean Sartori, bulletin du syndicat de l'éducation l'Ecole emancipée, coupure de journaux socialistes et de l'Humanité sur la question de la scission du PS et l'adhésion	Extrait de coupure de traitant de l'application de la motion de Strasbourg, de la scission entre la SFIO et la SFIC, des conditions d'adhésion à la IIIe IC.	Frossard,Ludovic-Oscar;;Longuet,Jean;;Zinoviev,Grégori;;Sembat;;Verfeun;;Reinhold,Louis;;Delattre,Alfred;;Caussy,		Parlementarisme;;Exclusion;;Lien entre syndicat et parti;;Syndicat;;Question agraire;;Question coloniale
517/1/0024	517	1	24	Résolutions du 3° congres de l'IC sur la dissolution du comité français pour la 3°Internationale. Resolution du CEIC sur le travail du PCF dans les syndicats.				
517/1/0025	517	1	25	Directives, lettres et télégrammes du secrétariat du CEIC à la direction du PCF sur la question de la tactique du parti suivant la décision du 3° congres de l'IC sur	Télégramme de Garinoviev à Marcel Cachin et Ludovic-Oscar Frossard du 1er août 1921, où il dresse le bilan du IIIe congrès de l'IC, les perspectives de travail et l'amélioration de la	Cachin,Marcel;;Frossard,Ludovic;;Monmousseau,Gaston;;Monatte,Pierre;;Garinoviev;;Trostky,Lé	Japon	Impérialisme Japon;;Résolutions 3e IC sur la question syndicale;;Modification
517/1/0026	517	1	26	Correspondance du CEIC et du secrétariat du CEIC avec la direction du PCF sur la situation intérieure du parti, la crise de la direction, le mouvement syndical révolutionnaire. Rapport de Loriot sur l'activité du PCF après le 3° congrès de l'IC.	Lettre de Marcel Cachin à Boris Souvarine du 7 septembre 1921 à propos des difficultés du PCF à se développer, sur le développement de L'Humanité et les erreurs de publications faites. Lettre de Marcel Cachin à Boris Souvarine du 19 octobre 1921 traitant de L'Humanité, du CSR et de la correspondance. Lettre de Fernand Loriot et Ludovic-Oscar Frossard au CEIC du 5 octobre 1921. Ils adressent une réponse à la lettre N°86 du 1er septembre 1921, où ils donnent la date du premier Congrès National du PCF (25 au 30 décembre 1921) et son ordre du jour. Ils déclinent ensuite l'assistance au peuple russe, la presse officielle du Parti, le nombre de membre du Parti et le recrutement, la politique du Parti avec les syndicats.	Souvarine,Boris;;Revo;;Humbert-Droz,Jules;;Gouralski,Abraham;;Sadoul,Jacques;;Loriot,Fernand;;Frossard,Ludovic-Oscar;;Cachin,Marcel;;Rosmer,Alfred;;Tommasi,Joseph;;Gaye,Georges;;Labonne,Victor;;Henchel;;Sirolle,Henri;;Jouhaux,Léon;;Monatte,Pierre;;Munzenberg,Willy;;Smeral,Bohumir;;Koenen;;Hernle;;Ionov;;Mikhalski;;Sokolski;;Mayer;;Zetkin,Clara;;Dunois,Amédée;;Rappoport,Charles;;Vaillant-Couturier,Paul;;Paul,Louis;;Treint,Albert;;Serrati,G	Prague(Bohême;Tchécoslovaquie);;Berlin(Allemagne);;Marseille(Bouches-du-Rhônes;France);;Mer Noire(Atlantique);;Algérie (Algérie Française;Département);;Alger(Alger;Algérie Française);;Dordogne(France;Département);;Bas-Rhin(France;Département);;Bouches-du-Rhônes(France;Département);;Luxembourg;;Belgique;;Italie	Commerce de riz;;Élection;;Répression communiste;;Presse du Parti Communiste;;Propagande;;Semaine de recrutement international;;Semaine de propagande international;;Diffusion de la presse;;Question syndicale

La routine informatique traite les données présentes en fonction des relations établies entre les colonnes sources et la balise destination attendue, conformément au standard XML EAD.
Lecture du tableau :

- les données issues des colonnes B, C et D forment la cote : on retrouve cette valeur dans l'attribut id de la balise <c />, comme valeur de la balise <unitid />, dans l'attribut href des balises <dao />

- la colonne Titre permet de remplir la balise <unittitle />

- le contenu de la balise <scopecontent /> provient des données des cellules de la colonne Scopcontent.

Les colonnes suivantes décrivent les mots-clés associés à la notice. Ces éléments sont « multivalués » pour économiser les lignes et simplifier l'affichage par la même occasion. Un caractère permettant de séparer les occurrences est présent : le symbole ";;", en règle générale au cas où des contenus disposeraient déjà de ";" dans leur texte. Ainsi, il sera créé autant de balises que d'éléments obtenus par découpage.

Le découpage est réalisé par macro intégrée (programme intégré développé pour le projet). Cela peut aussi être réalisé par transformation XSL à partir des données brutes du fichier converti au format XML.

Un avantage certain pour les contenus longs repose sur une meilleure gestion des contenus multilingues au moment de l'exportation.

La technique :

Deux possibilités pour générer une macro qui exécute cette tâche:

- utiliser la concaténation de chaînes de caractères pour générer les contenus

- utiliser la boîte à outils MSXML (accès moins immédiat)

Le choix s'est porté sur la première méthode : plus simple à aborder et à modifier pour des non spécialistes. Ne demande aucune activation de librairie supplémentaire. Les algorithmes sont extrêmement simples : pas de lecture dans un arbre, etc.

EXCELLIFICATION

Il s'agit ici de fournir l'opération inverse à la création d'un fichier EAD à partir d'un fichier Microsoft Excel, à savoir, récupérer les données issues d'un fichier XML conforme à la spécification EAD dans un logiciel de type tableur respectant les mêmes préconditions définies dans la partie précédente.

Hypothèses :

Données en entrée :

Nous disposons d'un fichier au format XML EAD et d'un modèle de représentation de ces données sous forme tabulaire.

Données en sortie :

On obtient un fichier au format Microsoft Excel, librement modifiable qui pourra de nouveau être rétroconverti en XML.

Cet outil est disponible en mode web, via le portail Paprika, accessible uniquement sur le réseau interne de la MSH de Dijon. En revanche, il est utilisable pour n'importe quel fonds puisqu'il la méthode s'appuie sur une organisation hiérarchique proposée par la spécification XML EAD.