TP 7 : encodage des apparences physiques

Aug 2012

1 Objectifs de l'exercice

Dans cet exercice nous allons explorer :

- l'encodage TEI des aspects spatiaux d'un manuscrit
- l'encodage de la "diachronie" de son écriture
- la liaison des informations spatiales et textuelles

2 Sommaire

Nous allons travailler sur l'image numérique d'un brouillon de l'auteur suisse Durrenmatt, manuscrite et tapuscrite, sur laquelle nous allons identifier des zones d'intérêt, selon l'ordre des écritures sur la page. Ensuite nous allons faire la liaison entre ces zones et une transcription diplomatique du texte contenu.

Nous allons nous servir des éléments suivants, du chapitre 11 "Transcription of Primary Sources" des Guidelines : `<surface>`, `<zone><line>`, et `<change>`.

Vous aurez besoin des fichiers fournis dans votre dossier Travaux, du logiciel Oxygen, et aussi de l'éditeur graphique Inkscape.

3 Import d'une image numérique dans Inkscape

Les images numériques sont produites dans des formats très variés (tiff, png, jpg...). Le format SVG (Scalable Vector Graphics) s'harmonise très bien avec la TEI – c'est un format XML ! Mais, puisque les images numériques sont stockées dans d'autres formats, nous allons commencer par l'import et la transformation d'un fichier JPG.

- Lancez le logiciel Inkscape.
- Séléctionnez la commande "Fichier > Importer" pour importer le fichier Durrenmatt.jpg que vous trouverez dans votre dossier Travaux
- Choisissez "lier" dans la boîte de dialogue qui s'affiche
- Sélectionnez la commande "Fichier > Propriétés du Document" : Dans cette boîte de dialogue, cliquez sur "Redimensionner la page au contenu..." puis sur le bouton "Ajuster la page au dessin ou à la sélection". Fermez la boîte de dialogue.
- Sélectionnez la commande "Fichier > Préférences d'Inkscape" et scrollez la liste qui s'affiche jusqu'à Sortie SVG, enlevez la coche pour "Autorise les coordonnées relatives" et fermez la boîte de dialogue.

4 Distinguer et marquer les régions ; noter leur coordonnées

Dans le fac-similé numérique, vous verrez une partie tapuscrite, des modifications manuscrites, un (ou deux ?) dessins, et peut-être d'autres choses.

Nous vous proposons de distinguer au moins quatre parties

1. le bloc de texte tapuscrit avec ses ajouts
2. les ajouts ou corrections manuscrites sur le bord droit

3. les additions manuscrites en dessous du bloc de texte
4. le dessin (comme une unité) dans la moitié inférieure de la page

Notre but est de marquer ces zones avec Inkscape, puis d'insérer leurs coordonnées dans un document TEI.

- Sélectionnez la commande "Affichage > Zoom > Zoomer" pour zoomer sur l'image afin de travailler avec plus de précision :
- Dans la barre d'outils à gauche, sélectionnez l'outil "Créer des rectangles et des carrés" (F4) et créez un rectangle autour du bloc de texte tapuscrit.
- Par défaut le rectangle que vous définissez ainsi est rempli de couleur : Pour modifier cela cliquez avec le bouton droit pour afficher la boîte de dialogue Remplissage et contour et cliquez sur le bouton en forme de croix "Pas de Remplissage"

5 Addition des coordonnées à un document TEI

Nous avons maintenant besoin d'insérer les coordonnées de ce bloc de texte dans un document TEI. Lancez oXygen et ouvrez le fichier Durrenmatt-start.xml. Ce fichier contient un petit document TEI minimal. Nous allons ajouter à son élément `<sourceDoc>` des informations concernant les zones de l'image

Ce document décrit la structure de notre source au niveau minimal. Il fournit des métadonnées pour l'identifier et il définit les dimensions de la surface physique sur laquelle le texte est inscrit. Les coordonnées de cette surface sont définies par quatre attributs de l'élément `<surface>`

- ulx définit la valeur X du coin supérieur gauche (0px distance horizontale entre le bord de l'image)
- uly définit la valeur Y du coin supérieur gauche (0px distance verticale de l'écran)
- lrx définit la largeur de l'image (800px)
- lry définit le hauteur de l'image (1220px)

En outre, le `<surface>` est assorti d'un identifiant, surface-1, de sorte que nous puissions clairement l'identifier plus tard.

Maintenant, nous pouvons enregistrer la première zone :

- Insérez un nouvel élément `<zone>` dans le document, après l'élément `<graphic>`, mais toujours au sein de `<surface.>`
- Cet élément devrait avoir un identifiant, ajoutez donc une valeur comme "zone-1" pour son attribut xml:id
- Les coordonnées d'une zone sont spécifiées par les attributs ulx, lrx etc, de la même manière que les coordonnées d'une surface. Pour connaître les valeurs à ajouter, il faut revenir dans Inkscape. Sélectionnez la commande Édition > Éditeur XML et une nouvelle perspective s'affiche sur votre fichier, qui ressemble beaucoup à la vue structurée présente dans oXygen. Cliquez sur le dernier objet `<svg:rect>` dans la liste et le panneau à droite vous propose les attributs associés avec ce rectangle
- Notez bien les valeurs pour x, y, height, et width: nous allons nous en servir dans notre document TEI : ces valeurs sont à transférer directement dans notre document TEI : elles servent respectivement comme valeurs des attributs ulxulyheight et width :
- Dans notre `<zone>` vous n'avez donc qu'à transférer la valeur de x à l'attribut ulx, etc.

6 Ajout d'une transcription

Bien sûr, nous voudrions compléter cette description topographique du document par une transcription des mots qui constituent chacune de ses zones. Nous allons traiter chaque zone comme une série de `<line>`

- Tapez "Ctrl + Maj + P" dans oXygen pour voir plus clairement la structure de votre fichier actuel. Actuellement, notre élément `<zone>` est vide, et donc oXygen l'affiche d'une manière abrégée. Nous allons y ajouter du contenu, donc il faut le développer.

- Supprimez le slash à la fin de l'élément `<zone/>`, et insérez un élément `<line>` entre la balise ouvrante du `<zone>` et sa balise fermante.

- Ouvrez dans oXygen le fichier Durrenmatt-Text.txt qui se trouve dans votre dossier Travaux. Vous verrez une transcription très basique du contenu de ce manuscrit. Nous allons laisser de coté toutes ses complexités pour l'instant ! Sélectionnez tout le texte contenu dans la zone en question et le copier avec Ctrl + C

- Revenez au document "Durrenmatt-start.xml" et collez ce texte au bon endroit. Cliquez avec Ctrl + Maj + D à la fin de chacune des lignes pour diviser l'élément `<line>` en plusieurs.

7 Combinaison de transcription et image

Un des aspects important de SVG comme format est le fait qu'il s'exprime en XML, comme la TEI. Donc, il est possible de "fusionner" notre transcription TEI-XML avec la version SVG de l'image numérique ; c'est-à-dire, de créer un nouveau document XML qui contiendra les deux. Ce document intégré aura des possibilités de traitement intéressant.

Pour effectuer cette intégration, nous allons nous servir d'une feuille de style XSLT. On peut configurer oXygen pour faire la transformation automatiquement :

- Sélectionnez la commande "Document > Transformation > Configurer les scénarios de Transformation"

- Dans la boîte de dialogue qui s'affiche, cliquez sur "Nouveau" et sélectionnez "XML Transformation with XSLT" ; ensuite, cliquez sur la petite icone en forme de dossier à droite de la fenêtre labellisée "XSL URL" et naviguer jusqu'au fichier xml2svg.xsl qui se trouve dans votre dossier Travaux.

- Vérifiez que l'option sélectionnée pour le "Transformateur" soit bien "Saxon EE 9.4.0.3" et cliquez sur "OK".

- Pour faire la transformation, il faut cliquer sur "Appliquer Associés" (ou cliquez sur le triangle rouge dans la barre d'outils). Si tout va bien un petit carré vert apparait en bas de l'écran avec le message Transformation réussie.

Dans votre dossier Travaux, vous devrez maintenant trouver un nouveau dossier "out" (pour "output"), dans lequel se trouve un autre dossier "time" ... et dans ce dossier vous allez trouver un document SVG intéressant...

- Ouvrez ce fichier avec Firefox, Safari ou Chrome et vérifiez que vous voyez bien une représentation de la zone que vous venez d'identifier et du fac-similé de la page.

8 Ajouter de nouvelles <zone>s

OK, laissons ce document pour l'instant, et revenons sur le document TEI-XML pour y créer encore des <zone>s ! Il y a au moins trois autres parties à distinguer dans le fac-similé numérique... Nous vous proposons d'ajouter encore trois éléments <zone> en suivant la même procédure qu'auparavant :

- Délimitez la zone dans Inkscape (voir section 4)
- Entrez ses coordonnées dans le document TEI (voir section 5)
- Ajoutez sa transcription (voir section 6)
- (facultativement) Exécutez à nouveau la transformation (voir section 7)

9 Définition de la séquence d'écriture

Sur http://research.cch.kcl.ac.uk/proust_prototype/index.html vous verrez un affichage très joli d'une analyse temporelle de l'évolution de l'œuvre de Proust, effectué à partir d'un balisage TEI semblable à ce que vous venez de faire. Il n'y a qu'une chose à ajouter : des indications de la séquence de production de ces morceaux d'écriture.

Bien sûr, l'identification de telles "étapes" dans la production d'un document est en général un sujet de recherche controversé, et parfois non vérifiable. Il n'empêche que l'explicitation des hypothèses de recherche (ce qui est, n'oublions pas, l'un des buts primaires de l'encodage) à ce sujet peut contribuer aux avancées scientifiques en ce domaine, comme dans d'autres.

Nous posons donc les hypothèses suivantes au sujet de la séquence d'écriture pour notre texte.

1. La partie tapuscrite doit précéder les annotations et corrections faites à la main par dessus
2. Dans les parties manuscrites on peut distinguer au moins deux épisodes, mais à part le fait qu'ils suivent forcément la partie tapée à la machine, leur succession n'est pas évidente
3. On peut supposer que les graphies en bas de page ont été ajoutées plus tard

Pour documenter ces trois étapes, il faut d'abord définir des éléments <change> dans l'entête du fichier TEI.

- Si vous avez réussi à définir toutes les zones du document, vous pouvez continuer avec votre propre version du fichier TEI XML. Sinon, pour s'assurer que nous travaillons tous sur la même version, nous vous proposons de prendre la version corrigée que vous trouverez dans le fichier Durrenmatt-change.xml dans votre dossier Travaux. Vous allez noter que ce fichier contient un balisage de quelques ratures ou d'additions (balisages <del> ou <add>) pour compléter le marquage des zones. Prenez un petit moment pour vous assurer que vous êtes d'accord avec le balisage proposé !
- Regardez dans l'élément <teiHeader> à la fin du <fileDesc>. Il faut ajouter un élément <profileDesc> pour documenter les aspects non-bibliographiques du document, notamment la séquence d'écriture.
- Dans le <profileDesc> ajoutez donc un <listChange>. Cet élément va contenir une liste des évènements ou étapes que nous désirons distinguer dans la séquence d'écriture, chacun étant décrit par un élément <change> portant un identifiant unique.
- L'élément <listChange> porte aussi deux attributs : ordered, avec la valeur "true", et n avec la valeur "writing". Nous indiquons d'abord que la séquence des éléments n'est pas aléatoire, et ensuite que c'est la séquence d'écriture que nous documentons, et non pas (par exemple) la séquence d'une lecture idéale.

- Allez donc ajouter un <listChange> ; contenant un <change> pour cette première étape

```
<profileDesc>
<creation>
<listChange ordered="true" n="writing">
<change xml:id="stage-1">texte dactylographié y compris le numéro de
page </change>
</listChange>
</creation>
</profileDesc>
```

- Ajoutez (au moins) deux éléments <change> supplémentaires, pour documenter les étapes que vous identifiez.
- N'oubliez pas d'ajouter à chacun de vos <change>s un xml:id ("stage-2", "stage-3"...) et une brève description. Par exemple :
  - stage-2 : "corrections de texte dactylographié, ajout des notes manuscrites"
  - stage-3 : "dessins en bas de page"

Maintenant, nous allons associer des éléments dans notre transcription TEI-XML avec l'étape définissant sa position dans la séquence d'écriture. Pour cela vous allez vous servir de l'attribut change, qui peut s'attacher à n'importe quel élément, et dont la valeur indique l'identifiant de l'élément <change> qui lui correspond.

- Regardez d'abord les cinq éléments <zone>. Ex : hypothèse, "zone-1" et "zone-2" sont associées à la même étape, celle que nous avons décrite dans l'élément <change> qui porte l'identifiant "stage-1". Ajoutez donc à ces deux éléments un **attribut**change avec la valeur "#stage-1" (le dièse indique qu'il s'agit d'un pointeur)
- Associez de la même manière les zones "zone-3" et "zone-4" avec l'étape "stage-2", et la zone "zone-5" avec l'étape "stage-3"
- Bien sûr, vous pourriez associer d'autres éléments de votre encodage (par ex. les ajouts) avec d'autre étapes. C'est à vous d'éviter les incohérences éventuelles !