TEI à la carte: travaux pratiques

Lou Burnard *July 2012*

1 Avant propos

Cet exercice vous permet d'expérimenter quelques uns des outils disponibles pour la création et le traîtement des fichiers TEI-XML. Nous allons faire beacoup avec oXygen, mais nous allons aussi regarder quelques autres. Vous aurez aussi besoin de quelques fichiers qui sont téléchargeable en format zip de l'adresse http://bit.ly/Nj0yce. Vous êtes invité de télécharger cet archive et de le dézipper avant de commencer l'exercice: Il va créer un dossier Travaux avec tous ce qui est nécessaire pour suivre cet exercice

En deux heures ce n'est guère possible de tout faire. Mais nous espérons vous fournir quelques idee sur les possibilités affordées par TEI-XML dans le domaine d'un projet de recherche notamment :

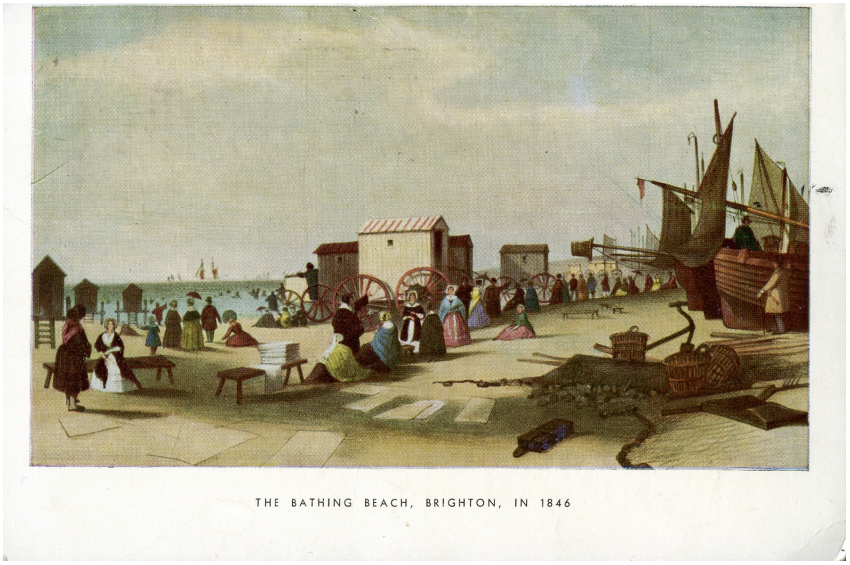
- l'utilisation d'une personnalisation TEI pour le balisage d'un fichier "plain text"
- l'utilisation de OxGarage pour la transformation et manipulation d'un document "bureatique"
- l'affichage des fichiers TEI XML en HTML, PDF, ePub etc.
- l'utilité du balisage pour des recherches dans un fonds textuel

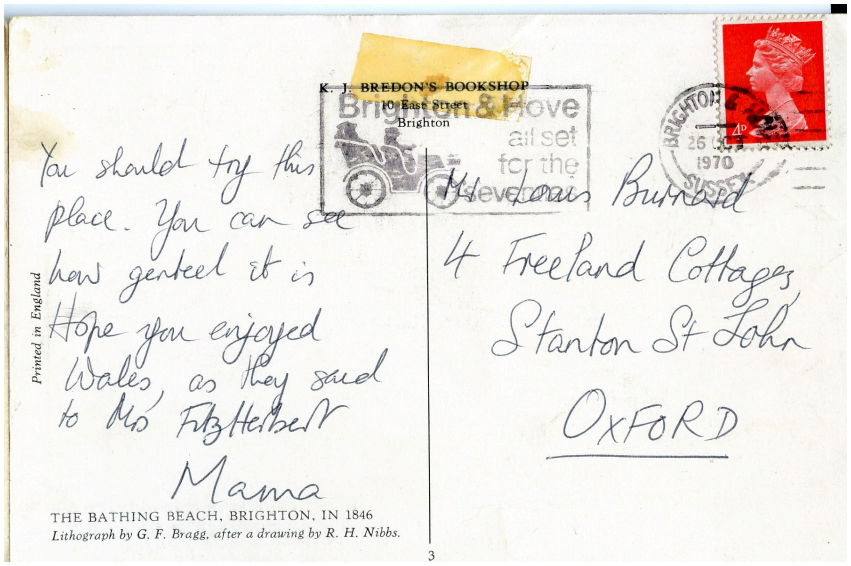
2 TEI à la carte : personnalisation

Vous l'ignorez peut-être, mais le feu Marcel Virgolos est mondialement reconnu comme l'un des pionniers des études carte-postaliques. Suite à sa regrettable disparition, nous avons été invités par ses exécuteurs testa,entaires de transformer sa collection personelle de 100,000 cartes postales en archive numérique, pour promouvoir l'étude de ce forme d'expression culturelle peu connu. Vous trouverez un tout tout petit échantillon de ce fonds patrimoniale dans votre dossier Travaux/Cartes.

2.1 Première carte

Nous allons travailler sur une seule carte d'abord. La voici, au recto :





Et voila son verso :

Nous avons deja discuté de la nécessité de bien choisir les balises qu'il nous faut, et comment nous servir de Roma pour créer un schéma qui reflet ces discussions.

2.2 Création de document nouveau

On va d'abord créer un nouveau document XML qui sera conforme à notre schéma tei_cartes.

- Démarrez oXygen.
- Cliquez sur l'icône Nouveau en haut à gauche (ou sélectionnez Nouveau du menu Fichier, ou tapez CTRL-N) pour ouvrir la boîte de dialogue Nouveau
- Choisissez Nouveau document, ensuite Document XML
- Cliquez sur le bouton Personnaliser en bas. Le dialog Customize Editor s'affiche
- Cliquez sur le petit triangle jaune à l'extreme droite du champs URL de schéma. Un sous-menu s'affiche.
- Sélectionnez Parcourir les fichiers locaux et naviguer jusqu'au fichier tei_cartes.rnc dans votre dossier Travaux. Sélectionnez ce fichier, et cliquez sur Ouvrir
- Vous revenez dans le dialog Nouveau. Cliquer sur Créer en bas.

oXygen vous propose un gabarit à compléter. Commencons avec les métadonnées :

- Nous vous proposons comme titre (`<title>`) 'The Bathing Beach, Brighton, 1846 : édition numérisée de la carte 19701026_0004 du fonds Virgolos'. Pour le publication statement, une phrase commme 'Echantillon inédit créé à l'atelier TEI, Lille 2013' servira.
- Pour le source description, nous vous proposons d'inclure tous les renseignments bibliographiques du verso, de cette manière :

```
<sourceDesc>
    <bibl><title level="m">The Bathing Beach, Brighton, in 1846</title>
        <respStmt>
            <resp>Lithograph by</resp>
            <name>G. F. Bragg</name>
        </respStmt>
        <respStmt>
            <resp>after a drawing by</resp>
            <name>R. H. Nibbs</name>
        </respStmt>
        <publisher>K. J. Bredon's Bookshop</publisher>
        <pubPlace>10 East Street, Brighton</pubPlace>
    </bibl>
</sourceDesc>
```

2.3 Ajout de texte au document

Rien ne vous empêche de taper à la main toute la carte directement. Mais pour gagner du temps on vous propose la démarche suivante :

- Assurez-vous que le curseur soit toujours entre les deux balises `<body>` et `<</body>>` de votre document vide
- Dans le menu Document, sélectionnez Fichier, et ensuite Insèrer un fichier
- Naviguez jusqu'au fichier card-0004.txt dans votre dossier Cards, et sélectionnez-le. Cliquez sur Open.
- Votre document est rempli de taches rouges! Pas de panique... on va régler ça petit à petit.

2.4 Structuration du document

Il est possible d'identifier dans ce document plusieurs sous-parties. En particulier, il contient :

- deux divisions physiques : à baliser `<<div type="recto">>` and `<<div type="verso">>` réspectivement
- au verso (en ce cas) on peut aussi distinguer deux parties: l'une contenant le message, l'autre contenant des information relatives à l'envoi de la carte (l'addresse du déstinataire, le timbre, l'oblitération etc.)
- Pour notre projet il nous semble utile de distinguer ces choses. Notons que nous ne tenons pas autant à encoder l'apparence physique de la carte: pour cela, l'image sert mieux.

Allons-y !

- Avec la souris, sélectionnez tout le texte que vous venez d'insérez, y compris la balise `<graphic>` au début.
- Tapez CTRL-E (ou sélectionnez XML Refactoring et ensuite Entourer des balises dans le menu Document)
- oXygen vous propose toutes les balises disponibles à cet emplacement : sélectionnez `<div>` et cliquez Accepter.
- Il faut ajouter des attributs au `<div>` : avec le cursor juste avant le > de sa balise ouvrante, tapez un blanc pour voir la liste des attributs disponible, et choisissez type (il est en gras parce qu'il est obligatoire). Tapez RETOUR pour l'insérer
- Une liste des valeurs possibles pour cet attribut s'affiche. Choissisez `recto` et tapez RETOUR pour l'inserer.

- Déplacez le curseur juste apres le mot '1846' and tapez ALT-MAJ-D (ou sélectionnez XML Refactoring et ensuite Élément de division dans le menu Document) pour effectuer une division.
- Les mots 'Beach view...and mats' fournissent une description de l'image ; ils ne figurent pas sur la carte. La balise prévue pour cela est `<figDesc>`. Sélectionnez cet empan, et tapez CTRL-E pour l'emballer dans un `<figDesc>`.
- Les mots 'The Bathing ... 1846' constituent le titre du graphie. La balise prévue pour cela est`<head>`. Sélectionnez cet empan, et tapez CTRL-E pour l'emballer dans un `<head>`.
- Ces trois éléments (`<graphic>` `<figDesc>` `<head>`) ensemble constituent un élément `<figure>`. Sélectionnez tous les trois, et entourez-les d'une balise `<figure>`. Plusieurs lignes rouges disparaissent ... on fait du progrès!

Procédons au verso ... Notre but initial est de séparer la partie contenant le message (`<div type="message">`) de la partie concernant l'envoi de la carte (`<div type="destination">`) ; nous allons nous servir des éléments `<p>` pour des paragraphes de texte, `<stamp>` pour les timbres, et `<address>` et `<addrLine>` pour l'adresse et ses lignes. D'autres balises supplémentaires sont envisageables, mais nous commençons simple.

Notons d'abord que nous disposons de deux versions du verso: une version en mode image, et également une version transcrite. Nous allons nous servir d'un attribut `<facs>` pour indiquer la correspondence entre les deux. Cet attribut est disponible pour tout élément dans une transcription pour le lier à sa représentation numérique en mode image.

- Changez en `verso` la valeur `recto` à l'interieur de notre deuxième `<div>`. Puis tapez un blanc pour voir les autres attributs disponibles.
- Choisissez facs dans cette liste. Sa valeur devrait être la chaine de caractères `197001026_004v.jpg`, actuellement présente comme valeur de l'attribut *@url* du deuxième `<graphic >`. Transférez cette chaine au bon endroit avec copier-coller, et puis supprimez ce qui reste de l'élément `<graphic>` ; nous n'en avons plus besoin.
- Avec la souris, sélectionnez tout le texte (i.e. de "You" jusqu'a "OXFORD") et tapez CTRL-E pour l'entourer d'un seul élément `<p>`. Répétez cette manipulation pour entourez ce `<p>` d'un `<div>`.
- Tapez un blanc à l'interieur de la balise ouvrante du `<div>`, et sélectionnez *@type* dans la liste d'attributs disponible qui s'affiche. Cette fois ci, specifier `message` comme valeur pour cet attribut.
- Presque tous les lignes rouges disparaissent. Est-ce que vous comprenez pourquoi ces esperluettes nous posent toujours une problème? Regardez le message en bas. Effectivement, dans un document XML les caracteres < et & doivent être représentés indirectement. Vous n'avez qu'à remplacer chaque esperluette avec la séquence `&amp;` : par exemple `Brighton &amp; Hove`
- Le petit carré vert apparaît ! Avons-nous terminé ? Hélas non : un document peut être valide, tout en contenant des mensonges ! Voyez-vous des mensonges ? Cliquez sur le bouton Indentation (ou tapez CTRL-MAJ-P).

Parce qu'elles ne sont pas explicitées par le balisage, plusieurs distinctions implicites dans la mise en forme de l'originel ne sont plus affichées. Il faut donc les baliser.

Nous devons d'abord separer les paragraphes au sein du message et sa signature.

- Ré-établissez l'affichage originel en tapant CTRL-Z.

- Diviser le paragraph en plusieurs, en tapant ALT-MAJ-D quand le curseur est positionné après 'is', 'Fitzherbert', 'Mama', et 'vermilion'

- Pour diviser le `<div>` il faut mettre le curseur entre la fin d'un paragraphe et le debut du paragraphe suivant, i.e. *entre* le `</p>` apres 'Mama' et le `<p>` qui le suit.

- Le mot 'Mama' n'est pas strictement une partie du message – c'est une signature, pour laquelle nous préferons utiliser l'élément `<signed>`. Vous pourriez retaper les balises, ou bien mettre le curseur à l'intérieur de la balise ouvrante du `<p>`, et puis tapez ALT-MAJ-R (Document - XML refactoring - Rename élément) pour renommer l'élément.

- Le `<div>` que vous venez de créer contient trois descriptions de timbre, et une addresse. Selon notre schéma son attribut *@type* devrait avoir la valeur `destination` : faites en sorte!

- Balisez chacune des descriptions de timbre en utilisant l'élément `<stamp>`. à vous de decider si vous le faites en sélectionnant le texte de chaque description et l'entourant d'un `<stamp>` l'une apres l'autre, ou bien en faisant cette manipulation qu'une fois, et ensuite en divisant l'élément en trois.

- Il ne nous reste qu'à traîter le destinataire. Sélectionnez le texte de l'adresse, tapez CTRL-E, et sélectionnez `<address>`. Les lignes rouge retournent parce qu'il faut baliser aussi les composants d'une adresse, en se servant des balises `<name>`, `<street>` ou `<addrLine>` selon votre gout.

2.5 Epreuve de la realité

oXygen peut afficher la structure hiérarchique du document que vous êtes en train de créer. Regardez dans la fenetre sommaire à gauche. Vous devrez voir quelque chose qui ressemble à ceci :

.

Avons nous fini de baliser?

Malheureusement, il n'existe aucune réponse certaine à une telle question. Nous avons balisé la structure essentielle du document pour faciliter l'affichage et la manipulation de ces composants en plusieures manières (nous allons voir cela un peu plus tard). Mais nous n'avons pas balisé *tout* les objets d'interêt sur cette carte. Nous revenons sur ce point à la fin de l'exercice.

2.6 Transformation d'un fichier Word

Vous savez maintenant comment créer un document TEI XML ex nihilo, ou à partir d'un document `.txt`. En réalité la plupart des documents sont créés avec un outil bureautique, comme Word ou Libre Office. Est-ce que nous pourrons travailler avec cette sorte de document aussi?

Dans votre dossier Cards, vous trouverez un petit document Word nommé Les deux soeurs.docx. Il contient une transcription jolie d'une carte postale : vous aimerez peut etre l'ouvrir avec Word pour vous assurez qu'il s'agit d'un véritable fichier Word, bien formatté, sans aucun chevron visible.

Quoi faire pour transformer ce joli fichier en TEI-XML et travailler la-dessus avec oXygen?

Nous pourrions l'exporter au format « plein text ». Nous pourrions aussi faire copier-coller. Mais dans chaque cas, nous risquons la perte du formatage qui distingue (par exemple) l'adresse des parties qui la précedent.

Si le fichier Word est bien stylé, on peut le transformer en XML, sans perte d'information. L'outil OxGarage nous aidera.

- Ouvrez votre Internet browser (Firefox de préférence), et allez sur le site
- Cliquez sur Text Documents. Dans la liste des formats de saisi proposée, cochez la case pour Microsoft Word Document(.docx).
- Une liste des formats cibles proposés apparaît. Cochez la case pour TEI P5 XML.
- En haut de l'écran un bouton Browse apparaît. Cliquez sur ceci, et naviguez jusqu'au fichier Les deux soeurs.docx dans votre dossier cartes.
- Cliquez sur le grand bouton Convert et patientez.
- Le site vous renvoie (après un bref délai) un archive nommé Les deux soeurs.zip
- Retrouver cet archive et cliquez là dessus pour le dézipper. Il contient un dossier Media et un fichier tei.xml. Ce dernier représente le contenu du fichier docx transformé en TEI XML.

2.6.1 Traduction des balises

Voyons ce que le Garage aura fait de notre fichier Word. En tout cas, il est devenu un document TEI valide, même s'il pourrait etre amélioré.

- Relancez oXygen (si nécessaire), et ouvrez le fichier tei.xml que vous venez de recevoir du garage.
- oXygen vous avertit que ce fichier ne contient pas de retours à la ligne : ce n'est pas grave. Cliquez sur "format".

Ne regardez pas le TEI Header pour l'instant. La conversion à pu distinguer les trois composants du verso de la carte (le message, les timbres, et l'addresse) parce qu'ils étaient formattés différemment dans le fichier Word, en se servant des stylages word divers. Ce fait nous aide beaucoup.

Par exemple, l'adresse à ete transformé en tableau, avec un rang contenant une seul cellule pour chaque ligne.

- Mettez le curseur à *l'interieur* d'un des éléments `<cell>`
- Tapez ALT-MAJ-R (Document -> XML-Refactoring -> Renommer l'élément) et changez `cell` en `addrLine`
- Cochez la case Renommer tous les éléments ayant le même nom et cliquez OK.
- Tous les `<cell>` deviennent des `<addrLine>`s. (Le document n'est plus valide, mais il devient plus honnet!)
- Il faut enlever toutes les balises `<row>` et `<table>`. Vous pouvez faire cela en les supprimant comme n'importe quel autre caractère. Ou, d'une manière plus fiable, mettez le curseur sur une des balises `<row>`, et sélectionnez la commande Document -> XML-Refactoring -> Effacer les balises des éléments .
- Pour terminer, emballer votre séquence de `<addrLine>` avec un `<address>` et votre document sera de nouveau (presque) valide.
- Les descriptions de timbre sont convertis dans une liste, plutot qu'un table, mais la démarche est pareille. Renommer tous les éléments `<item>` en `<stamp>`s et emballez les dans un élément `<p>` comme auparavant.

Nous vous laissons compléter le balisage de cette carte. N'oubliez pas d'introduire des `<div>` éléments de bon type au bon endroit, ni de transférer des informations dans l'entête. La carte déjà faite peut vous servir comme modèle. Si il vous reste du temps, essayez de transcrire d'autres cartes : vous en trouverez encore trois exemplaires dans le dossier Cards.

Vous trouverez notre suggestion pour toutes les cartes dans le fichier postcard-archive.xml!

3 A quoi sert le balisage?

Le balisage qu'on introduit si soigneusement et avec tant d'effort n'est pas là juste pour le plaisir. On espère en profiter. Parce que nous aimons tous la lecture, il y a une tendence naturelle de se limiter a la production des choses simples (même agréables) à lire à partir de ce document balisé Mais on peut aussi profiter du balisage pour faire l'analyse des traits ou de la structure du document, pour faire des recherches intelligentes à travers plusiueurs documents, ou simplement pour les indexer.

3.1 Transformation pour l'affichage

Vous le saurez peut être déjà : en mode Auteur oXygen transforme votre document pour l'afficher sans balises, sous controle d'un feuille de style CSS. Il est également possible de le transformer en HTML ou en PDF.

- En oXygen, ouvrez n'importe quel fichier XML TEI contenant une seule carte postale
- Cliquez sur le bouton à droite de la grande flèche rouge (ou tapez CTRL-MAJ-C, ou sélectionnez Document -> Transformation -> Configurer Scénario(s) de Transformation)
- Le terme 'Scénario de Transformation' en oXygen s'applique à une association prédéfinie entre un document et une feuille de style pour le traîter. Quelques scénarios sont fournis pour des formats souvent utilisés, notamment TEI P5 XHTML , TEI P5 EPUB, TEI P5 PDF, TEI P5 DOCX.
- Sélectionnez l'un de ces scénarios. Cliquez sur le bouton Appliquer associes button. Qu'est-ce qui se passe?
- Bien sur, on peut configurer ces transformations à volonté, et en créer des nouveaux. Pour vous en donner le goût, nous proposons une transformation qui n'a rien à faire avec la visualisation.

La feuille de style qui controle une transformation s'écrit dans un autre langage XML, qui s'appelle XSLT : sujet fascinant, mais peut être pas pour les débutants. Vous trouverez quelques exemplaires de feuilles de styles XSLT dans votre dossier. Pour en servir, il faut créer une nouvelle transformation en oXygen

- Configurez un scénario (CTRL-MAJ-C) de nouveau
- Cliquez cette fois le bouton Nouveau et choisir XML transformation with XSLT pour ouvrir la fenetre Nouveau Scenario. Dans cette fenêtre :
  - choisir un nom pour le scénario : on vous propose "texte-brut"
  - specifier la location du fichier XSL souhaité, en cliquant sur le petit icone dossier jaune à droite du champs XSL URL. Ceci vous permet de naviguer au dossier Travaux : sélectionner le fichier texte-brut.xsl et cliquer Ouvrir.
  - sélectionnez Saxon HE 9 du menu Transformateur
  - Cliquer Accepter pour terminer, et Transformer Maintenant pour voir le résultat.

Comme vous voyez, c'est possible de transformer notre document XML ou bien dans un format d'affichage, ou bien dans un format d'analyse simplifié.

3.2 Recherches dans la structure

Une des motivations importantes pour l'application du balisage est la possibilité d'identifier et donc de retrouver les composants individuels d'un document independemment du document lui même. Par exemple, on souhaite avoir une liste des timbres, ou rechercher les lieux d'où ont été expédié un ensemble de cartes postales à une date precise. Voici quelques exemples tres simples pour vous donner un peu l'idée de ces possibilités, toujours en nous servant d'oXygen.

A gauche en haut, il y à une petite fenetre labelisée XPath 2.0. XPath c'est un standard W3C qui permet d'identifier des parties d'une structuration XML. Nous n'entrons pas dans les détails, mais nous démontrons quelques exemples.

- Ouvrez le fichier postcard-archive.xml en oXygen. Ce fichier contient cing cartes postales déjà balisés.
- Tapez `//salute` dans la fenêtre XPath, et tapez RETOUR. Au fonds de l'écran, un tableau s'affiche contenant un rang par occurence d'élément `<salute>` dans le document: chaque rang indique d'abord la location exacte de l'occurrence, et ensuite son contenu. Par example, la premiere ligne nous informe que le texte 'Love Kath.' se trouve dans le premier `<salute>` contenu par le premier `<div>` contenu par le deuxieme `<div>` contenu par le premier `<body>`contenu par le premier `<TEI>` contenu par l'élément racine `<TEIcorpus>`.
- Essayons de voir tous les titres des cartes. Tapez `//title`dans la fenêtre XPath, et tapez RETOUR.
- Hmm. On voit le titre de la collection, le titre d'une source bibliographique, le titre de la carte elle-même. Essayons d'être plus précis en tapant plutot `//TEI//titleStmt/title`, c'est-a-dire 'retrouvez les `<title>`s qui sont directement contenus par un `<titleStmt>` qui sont eux-memes directement contenu par un `<TEI>`', ce qui est plus exacte. Vous ne devrez voir que cinq rangs de resultats.
- Maintenant tapez `//stamp` pour voir tous les éléments `<stamp>`.
- On peut sélectionner selon la valeur des attributs. Essayez donc `stamp[@type='postage']` pour n'afficher que les timbres postes, excluant les oblitérations. Qu'est-ce qu'on ferait pour l'invers?
- Enfin, sauriez vous comment extraire les messages?

3.3 Avons nous fini *maintenant*?

Voici quelques propositions supplémentares, supposant que vous souhaitez continuer de travailler avec de tels documents :

- Comme tout objet manuscrit, la carte postale peut comprendre des erreurs, des ajouts, des corrections, des passages illegibles etc. La TEI propose des balises pour tous ces cas, et d'autres.
- Une carte postale souvent fait référence à de vraies personnages, et à des lieux existants. Nous pourrions indiquer ces 'entités nommés' (comme on dit) en les balisant avec `<name>`, distinguant par ex nom de personne, nom de lieu, nom d'evenements etc. avec l'attribut *@type*
- Nous pourrions enricher la ressource avec une normalisation et des explications sur les entités nommés eux meme. Par ex. qui était 'Mrs FitzHerbert'? il y a un petit jeu de mots à commenter à son egard.

- Notre balisage n'a pas tenté d'indiquer la mise en page originelle, par ex l'orientation de l'écriture. Cela pourrait être intéressant en quelques cas.
- Les entités nommés que nous trouvons sont également réferenciés par d'autres ressources numériques (ou pas) : des catalogues de timbres, des index géographiques ou onomastiques, des listes d'entreprises commerciales, de maisons d'édition etc. Il serait tres utile d'ajouter des liens vers de telles informations.
- La carte elle même pourrait être d'interêt : les metadonnées à ajouter pourraient traiter par exemple le genre d'image, ou bien les moyens de production, combien de cette carte ont été imprimés, sa valeur pour les collecteurs actuels ou anciens, d'autres exemplaires conservés dans d'autre collections etc...
- Ou nous pourrions etre intéréssés par la carte comme objet linguistique. Son texte est d'habitude bref, informel, et formulaique, ces formules auront donc d'interet de point de vu analyse discours et semiotique (on pourrait les comparer avec des tweets par exemple)...
- ... and so on!

Prenant position sur toutes ses possibilités est évidemment une tache à completer avant d'investir beacoup d'effort dans la saisie de nos documents XML et la definition de notre schéma. Heureusement, ils sont tous prévus dans la TEI, donc on peut s'offrir la possibilité d'enricher notre archive sans trop de perturbation à l'avenir...