Représentation et Stockage des données de la numérisation du dictionnaire Trévoux

Ingrid Falk

Représentation et Stockage des données de la numérisation du dictionnaire Trévoux

Ingrid Falk

15 mars 2006

Résumé

Le dictionnaire dit de « Trévoux » est un « dictionnaire universel françois et latin » dont la première édition a été publié en 1704. Sa numérisation en vue à la foi de la préservation et de la valorisation fait partie du programme du CPER ILD-ISTC. Dans le cadre de mes travaux pour READ j'ai exploré les outils et ressources de normalisations mise en œuvre dans d'autre projets de création de collections numériques, ce qui m'amène a proposer un mode de stockage dans des formats XML standardisés en préservant les liens entre les images scannées, les objets reconnus par les OCR et des différentes versions du contenu textuel.

1 Introduction

Ce rapport présente mon travail au sein de l'équipe READ dans le cadre du projet CPER ILD-ISTC. Il s'est déroulé de janvier à mars 2006. Au cours de ce travail j'ai exploré comment représenter les données provenant de la numérisation du dictionnaire Trévoux dans un souci à la foi de préservation et de souplesse d'utilisation et de présentation. Dans ce rapport je présente d'abord (très brièvement) les standards existants et utilisés dans le domaine des collections numériques. Ensuite, on montre à travers d'un exemple - l'image scannée des pages 105 et 106 du dictionnaire Trévoux - comment on pourrait procéder pour représenter et stocker les ressources en résultant dans les formats mentionnés précédemment, avec le moins d'intervention humaine possible. Finalement, je présente un exemple d'application : à partir des fichiers XML composants l'objet numérique on génère une page web (en XHTML + CSS) pour visualiser les correspondances entre des parties de l'image initiale, les objets identifié par l'OCR et les textes codés dans le document de vérité.

2 Les différents standards

2.1 La TEI

La TEI (« Text Encoding Initiative »), v. [1], fondée à l'origine sur le SGML et s'appuyant désormais sur le XML, est un langage de marquage qui permet d'échanger des données textuelles, notamment pour les sciences humaines et les études sur les textes littéraires.

Mais en fait, la TEI n'est pas une DTD, mais un ensemble de recommandations (« Guidelines ») et d'éléments rassemblés en modules distincts (« tag sets ») dont l'utilisation et le choix forment une DTD particulière. Il n'existe donc pas une DTD TEI mais autant que les combinaisons de modules rendent possibles. Pour reprendre les mots de Lou Burnard (cf. [2]), la TEI est un système extensible, modulaire et polymorphe constituant un modèle abstrait.

Par exemple, la TEI convient parfaitement à la foi pour coder un document de vérité et les entrées d'un dictionnaire mais, on se servira de schéma différents :

Pour les documents de vérité on mettra l'accent plutôt sur les particularités typographiques :

- les sauts de lignes
- les marques de pages
- changements de polices
- glyphes inhabituels
- taches
- fragments illisibles

Par contre, pour le texte finalisé on voudra marquer :

- des entrées du dictionnaire
- des mots vedettes
- la description morphologique et syntaxique
- homonymes et synonymes

Les deux documents auront un set de balises de bases communes auxquelles on va rajouter celles qui conviendront le mieux au document et à l'utilisation envisagé de l'encodage.

La TEI est largement utilisée et documentée. Elle propose des formats de codage pour une large variété de textes et d'usages, mais si elle s'avère incomplète elle peut toujours être complété.

Si la TEI a été conçu pour représenter des structures textuelles, elle ne se prête pas aussi bien à la représentation détaillée et exhaustive de la mise en page physique d'un document imprimé. Mais pour la reconnaissance et structuration automatique de documents on a besoin de toutes ces informations physiques parce qu'on voudrait en inférer des unités logiques. Par exemple, dans le dictionnaire « Trévoux » les mots vedette sont toujours typographié en majuscule, les débuts de paragraphe sont indentés. Une fois que l'on dispose de cette information il conviendrait de la préserver, ainsi que le lien avec les images des originaux.

2.2 ALTO

Si la TEI est établie comme un standard pour la représentation structuré de texte, il m'a été beaucoup plus difficile de trouver un format (standardisé) pour représenter d'une manière exhaustive la composition d'une page d'un texte imprimé : en marges et un espace imprimé et en blocs physiques - textuels, illustrations ou composés.

ALTO (Analysed layout and text object), cf. [3], est le seul format ouvert et assez générique que j'aie trouvé pour représenter les résultats des OCR sur du texte imprimé.

C'est un schéma XML issu du projet européen METAe (Metadata Engine Project - cf. [7]) qui est actuellement utilisé aussi par le National Digital Newspaper Program aux Etats Unis (cf. [8]).

2.3 Putting it all together : le format de métadonnées METS

METS (Metadata Encoding and Transmission Standard), v. [4], est un schéma XML élaboré à l'initiative de la Digital Library Federation. Il sert à encoder les métadonnées descriptives, administratives et structurelles de documents, ou plutôt d'objets numériques. Dans cette vision un objet numérique sera composé de plusieurs documents numériques (images, sons, textes en différents formats ...) et un fichier METS répertoriant les noms et la localisation des fichiers leur correspondant, des métadonnés descriptives et administratives, mais aussi une carte de structure pouvant représenter des correspondances entre des sousensembles de ces documents.

METS est déjà utilisé dans beaucoup de projets de numérisation, entre autre par l'École des chartes (cf. [9]) et la Bibliothéque Nationale de France.

3 Application au dictionnaire Trévoux

Les objets graphiques Dans cette exemple l'objet numérique sera construit autour de l'image scannée des pages 105 et 106 du dictionnaire Trévoux (édition ? ?). Le document numérique initial est un fichier image (DUT01_0082.tif) dans le format tif : (1).

En vue d'une présentation sur internet on pourrait convertir cette image en d'autres formats (comme png, gif ou jpeg) ou encore de prévoir des images plus petites en taille.

L'analyse physique L'OCR va analyser l'image et va stocker ses résultats dans un format XML propriétaire. Ce fichier XML pourra être convertit (automatiquement) dans un fichier XML alto DUT01_0082_alto.xml (2)

Les formats textuels Pour pouvoir analyser les résultats de l'OCR on a besoin d'un document de vérité. Ce document doit être produit manuellement, mais il peut être codé dans un format TEI approprié : DUT01_0082_tei_3.xml (3).

Les métadonnés pour monter l'ensemble Finalement on va regrouper dans un fichier METS les informations sur

- les fichiers composants cet objet numérique (4) :
  - les images : DUT01_0082.tif, DUT01_0082.jpg ...
  - le fichier ALTO avec l'analyse (physique) de l'OCR : DUT01_0082_alto.xml
  - le document de vérité (en TEI) : DUT01_0082_tei_3.xml
- des liens structurels entre ces documents et quelquesunes de leurs sous-divisions (5)

L'illustration 6 montre schématiquement le type d'informations que peut représenter un fichier METS et le fonctionnement de la carte structurale.
Le fichier METS peut être produit automatiquement (par des scripts ou transformation XSLT).

L'objet numérique Pour conclure, dans ce cas, l'objet numérique sera constitué :

- des fichiers images : DUT01_0082.tif, DUT01_0082.jpg, ...(cf. fig 1)
- du document de vérité (s'il y en a) : DUT01_0082_tei_3.xml (cf. fig 3)

blir un à Conſtantinople, qui après la mort du Saint fut transféré en Bithynie par Jean ſon ſucceſſeur. A Jean ſuccéda Marcellus, que Nicephore a cru être l'Inſtituteur des *Acoemétes*. Sous ce Marcellus ce pieux inſtitut s'étendit beaucoup, dit Bollandus; & c'eſt là apparemment ce qui a fait que Nicéphore l'en a cru fondateur. Ce fut de ſon tems que Studius vint de Rome à Conſtantinople y bâtir un Monaſtère, & y mit des Moines, qu'il tira des Monaſtères *Acoemétes*. Ce fut là l'origine des Studites, qui conſéquemment viennent des *Acoemétes*. Saint Jean Calybite ſe retira dans un Monaſtère d'*Acoemétes*: & non pas d'*Arométes*, comme a imprimé la Sauſſaye dans le Martyrologe de France. Quoique les *Acoemétes* ayent fleuri ſur tout en Orient, il y en a cependant eû quelques-uns en Occident. Le P. le Cointe prétend, à l'endroit que je citerai, qu'il n'y a eû que le Monaſtère de Luxeuil, *Luxovienſe*, celui de Remiremont, *Habendenſe*, & celui de S. Salabèrge à Laon, où l'on ait dit perpetuellement l'Office de la maniere que nous l'avons expliqué. Le P. Mabillon ſoutient qu'il y faut ajoûter celui de S. Maurice, *agauninſe*, fondé par Sigiſmond Roi de Bourgogne, celui de S. Marcel de Châlons, & celui de Saint Denys en France. D'autres ajoûtent encore celui de S. Riquiér &c. Il n'eſt pas vrai que S. Euchèr Évêque d'Orléans ſe fit Moine *Acoemète*, comme l'a dit Caniſius. Ce fut dans un Monaſtère de Bénédictins, à 5. lieuës de Rouen, qu'il ſe retira, comme l'a remarqué Bollandus. T. 1. de Janv. p. 1019.

On a auſſi appellé *Acoemétes* les Stylites, & quelques autres Moines de la Paleſtine; mais dont l'Inſtitut étoit fort différent de celui des *Acoemétes*. On pourroit aujourd'hui appeller *Acoemétes* les Religieuſes du S. Sacrement, qui ont l'adoration perpétuelle, & ſe relèvent jour & nuit, enſorte qu'il y en ait toûjours devant le S. Sacrement à prier.

Outre Nicéphore & Bollandus, dont j'ai parlé, Théodore Lècteur L. 1. Evagrius L. 1111e. C. 18. & 21. Théophane, Cédrenus, l'Auteur de la vie de S. Alexandre dans Bolland. 15. Janv. & Jacobus Caniſius dans le Ribadeneira Latin au 20. Février, Baronius à l'an 459. M. du Freſne dans ſon Gloſſaire, le Cointe Annal: T. I. an. 536. n. 224, & ſuiv. Le P. Mabillon *Act. Sanct. Bened. Sæc. IV. p. 2. Præf.* ont écrit des *Acoemétes*.

ACOINT, TE, adj. Ce mot veut dire familiér, ſelon Nicod. *Amicus familiaris.*

ACOLALAN, ſ. m. Inſecte de l'Iſle de Madagaſcar. Il reſſemble à une punaiſe. Il eſt plus gros.

ACOLYTHAT, ſ. m. *Acolythatus*. Ordre, rang d'Acolythe: c'eſt le prémiér des moindres Ordres; c'eſt-à-dire, celui qui précede immédiatement le Souſdiaconat.

ACOLYTHE, ſ. m. Tèrme Eccléſiaſtique. *Acolythus*. Les Grècs donnoient ce nom à ceux qui étoient inébranſables dans leurs réſolutions. C'eſt par cette raiſon que les Stoïciens furent appellez *Acolythes*; parce que rien ne pouvoit leur arracher leurs ſentimens. Ils trouvoient même qu'il y avoit de la lâcheté à en changer. Depuis, l'Égliſe Chrétienne a conſacré ce nom, en l'appliquant à ceux qui ſe dévoüent au ſèrvice de Dieu. Anciennement les jeunes gens qui aſpiroient au miniſtère Eccléſiaſtique, accompagnoient & ſuivoient les Évêques par tout, ſoit pour les ſèrvir, ſoit pour être les témoins de leur conduite. Cette aſſiduité à ſuivre les Evêques les fit appeller *Acolythes*. Saint Cyprien dit lui-même, qu'il avoit des *Acolythes*. Aujourd'hui les fonctions des *Acolythes* ſont bien différentes de la prémiére inſtitution. Un *Acolythe* eſt celui qui a ſeulement reçû le prémier & le plus conſidérable des quatre Ordres Mineurs dans l'Égliſe; dont l'emploi eſt d'allumer les cièrges, de porter les chandeliérs, la navette où eſt l'encens, de préparer le vin & l'eau pour le ſacrifice, & de rendre d'autres ſèrvices à l'autel. Autrefois les *Acolythes* ramaſſoient dans un ſac ce que les fidéles avoient offert, & qui avoit été béni pendant la meſſe, & après qu'elle étoit finie ils le donnoient aux Prêtres, qui devoient le diviſer. Le devoir des *Acolythes* eſt d'accompagner l'Évêque, ou le Prêtre, & de leur rendre ſèrvice dans les fonctions Eccléſiaſtiques. C'eſt l'Ordre que les jeunes Clercs éxèrcent le plus. Il y avoit à Rome trois ſortes d'*Acolythes*. Les *Acolythes* du Palais. *Palatini*, qui ſèrvoient le Pape; les *Acolythes* Stationaires, *Stationarii*, qui ſervoient dans les Égliſes, où il y avoit Station, les *Acolythes* Régionaires, *Regionarii*, qui ſervoient avec les Diacres dans les différens quartiers de la ville. On trouve auſſi des *Acolythes* parmi les Officiérs Auliques de Conſtantinople; & Curopalates dit que le Capitaine, ou Chèf de la Cohorte Impériale de Bizance, étoit nommé *Acolythe*.

Dans l'Euchologe des Grècs on trouve les leçons qu'on lit lorſqu'on ordonne des Lècteurs; mais il n'y eſt point parlé des autres moindres Ordres, qui ſont, l'Ordre de Portiér, d'Éxorciſte, & d'*Acolythes*: ce qui pourroit faire croire que les Grècs ne confèrent point ces Ordres-là aujourd'hui. Le Père Goar dans ſes notes ſur l'Euchologe, répond qu'on ne peut pas douter que ces trois moindres Ordres n'ayent été connus de l'ancienne Égliſe Grècque, & qu'elle n'ait eû des Miniſtres qui les avoient, puiſqu'il en eſt fait mention dans Saint Denis, S. Ignace Martyr, Saint Epiphane, dans les Conciles de Laodicée & d'Antioche, dans les Novelles de Juſtinien, dans Photius &c. Il ajoûte, qu'il ſemble que les Grècs d'aujourd'hui ont des *Acolythes* ſous le nom de Députez & de Céroféraire. Les Miſſionnaires Latins qui ſont en Grèce diſent que les Grècs ont aujourd'hui des *Acolythes*, & les autres moindres Ordres; & on doit plus les en croire que le Père Martène, qui aſſûre que l'Ordre des *Acolythes* a été tout à fait inconnu à l'Egliſe d'Orient. Voyez le P. Goar ſur l'Euchologe, le P. Martène de Pontifical, l'Ordre Romain, &c.

Ce mot vient du Grèc ἀκολουθεῖν, qui ſignifie ſuivre. Et *Acolythe*, un ſuivant. C'eſt ainſi que l'éxplique le Gloſſaire Grèc & Latin, & Macér; mais Dominique ſon frère le tire de l'*a* privatif, & de κωλύω, *empêcher*.

ACOMAS, ſ. m. Arbre qui croît dans les Iſles Antilles, & dont le bois s'employe aux ouvrages de Menuiſerie. Cet arbre eſt à peu près de la hauteur de nos pommiérs; ſes feuilles ſont aſſez longues & liſſes; ſon fruit eſt de la groſſeur d'une prune, il devient jaune dans ſa maturité. Son amèrtume empêche qu'on ne le mange: il n'y a que les pigeons ramiers qui puiſſent s'accommoder de ſon fruit; mais leur chair en retient ſi fort le goût, qu'on ne peut les manger dans le tems qu'ils s'en nourriſſent: l'écorce de cet arbre eſt raboteuſe, cendrée, & elle donne un ſuc laiteux lorſqu'on l'inciſe. Son bois eſt peſant, de couleur rouge, tirant ſur le jaunâtre; le cœur eſt d'un rougé tirant ſur le violet. Ces couleurs varient ſuivant ſon âge; & tout le bois prend fort bien le poli. Rochefort. Le Père du Tèrtre rapporte qu'un Nègre l'avoit guéri d'un grand mal de dents en lui frottant les tempes & le dèrriére des oreilles avec le lait qui ſe tire de l'écorce de l'*Acomas* franc. Car ce Père, Hiſtoire des Antilles, Traité 3. C. 4. §. 3. diſtingue trois ſortes d'*Acomas*; l'*Acomas* franc, qui eſt un des plus gros, & des plus hauts arbres des Antilles, & le meilleur de tous pour les bâtimens; l'*Acomas* bâtard, qui croît à la Capſtèrre de la Guadeloupe, qui n'eſt ni ſi beau, ni ſi bon à bâtir que le précédent; & le troiſiéme, qui croît aux environs de la grande Ance, ſemblable au prémiér, ſinon que le cœur en eſt rouge.

ACOMMICHER, Vèrb. act. Vieux mot François, qui vouloit dire *Communier*, *donner la Communion*. Et fit le Roi dire grand planté de Meſſes, pour *accommichér* ceux qui devotion en avoient. Froissard.

ACOMPARAGER, Vèrbe act. Ce mot, ſelon Nicod, veut dire *comparer*. *Conferre*, *comparare*,

ACOMSIGT, part. & adj. Ce mot dans Pèrcéval veut dire *pourſuivi*.

ACON, ſ. m. *Cymba*. Tèrme de marine. Petit bateau à fond plat, dont on ſe ſèrt pour aller ſur la vaſe quand la mèr eſt retirée.

ACONIT, ſ. m. *Aconitum*. n. Plante venimeuſe. Les anciens Botaniſtes ont attribué ce nom à pluſieurs plantes de différens genres. Celles dont il s'agit ici ſont leurs fleurs irrégulières, compoſées de pluſieurs pétales, dont l'aſſemblage repréſente aſſez bien un caſque ouvèrt; c'eſt-à-dire, que la pétale ſupérieure fait le caſque du heaume, les deux latérales tiennent la place des deux oreillettes, & les inférieures repréſentent la mentonniére. Les éſpéces qu'on nomme tuë-loup, *Lycoctonum*, Αυκοκτόνον, ont leur caſque allongé en maniére de toque, ou de bonnet à la Polonoiſe. Les fruits qui ſuccèdent aux fleurs ſont compoſez de pluſieurs graines, qui s'ouvrent ſelon leur longueur, & renfèrment des ſemences anguleuſes, & chagrinées. Ses feuilles ſont arrondies & découpées plus ou moins profondément. Ce genre d'*aconit* comprend pluſieurs éſpéces; qu'on peut ranger ſous trois principales claſſes. La prémiére eſt de celle dont toute la fleur eſt bleue, ou violette, & la pétale ſupérieure de la fleur forme un caſque. On la nomme Napel, *Napellus*, à *Napo*, à cauſe que ſes racines ſont en navets. Le Napel eſt très-dangereux; mais on a trop éxagéré ſon venin. La ſeconde eſt de celle qui a ſes fleurs tout à fait ſemblables à celles du Napel, hormis qu'elles ſont jaunes. Elle s'appelle Anthora. *Anti-Thora*. C'eſt-à-dire, plante ſouveraine contre les mauvais éffets du Thora. Elle eſt auſſi venimeuſe que le Napel. Il eſt faux que l'Anthora croiſſe toujours auprès du Thora, ou du Napel. L'*aconit* de la troiſiéme claſſe ſe diſtingue des deux précédentes par la figure allongée de ſon caſque. Ses fleurs ſont pâles, ou jaunâtres. On l'a appellée tuë-loup, étrangle-loup, tuë-chien, à cauſe de ſes éffets. *Aconitum Lycoctonum*. Λυκοκτόνον, Κοκκτόνον. La prémiére & la dernière de ces trois ſortes d'*aconit* ſont très-cauſtiques, très-âcres, & cauſent des convulſions mortelles, ou des inflammations ſuivies d'une gangrène prochai-

FIG. 1 – image scannée des pages 105 et 106 du dictionnaire Trévoux

```
<?xml version="1.0" encoding="UTF-8"?>
<alto>
  <Description/>
  <Styles>
    <TextStyle ID="Arial_11." FONTFAMILY="Arial" FONTSIZE="11."/>
    <TextStyle ID="Times New Roman_10._466_29" FONTFAMILY="Times New Roman" FONTSIZE="10." FONTSPACING="29" FONTSCALING="466"/>
  </Styles>
  <Layout>
    <Page WIDTH="2480" HEIGHT="2480" ID="P0_DUT01_0082" PHYSICAL_IMG_NR="1">
      <PrintSpace>
        <TextBlock ID="B0" HPOS="64" WIDTH="124" VPOS="54" HEIGHT="74">
          <TextLine BASELINE="124" ID="B0_L0" HPOS="72" WIDTH="101" VPOS="58" HEIGHT="66" STYLEREFS="Times New Roman_22._616_-17">
            <String CONTENT="’⅞5" STYLE="bold italic" HPOS="72" VPOS="58" WIDTH="101"/>
          </TextLine>
        </TextBlock>
        <TextBlock ID="B3" HPOS="52" WIDTH="1174" VPOS="136" HEIGHT="4038">
          <TextLine BASELINE="189" ID="B3_L0" HPOS="98" WIDTH="1115" VPOS="144" HEIGHT="58" STYLEREFS="Times New Roman_11._-4">
            <String CONTENT="■blir" HPOS="98" VPOS="186" WIDTH="80"/>
            <SP HPOS="178" VPOS="161" WIDTH="10"/>
            <String CONTENT="un" HPOS="188" VPOS="161" WIDTH="43"/>
            <SP HPOS="231" VPOS="148" WIDTH="10"/>
            <String CONTENT="d" HPOS="241" VPOS="148" WIDTH="17"/>
            <SP HPOS="258" VPOS="148" WIDTH="10"/>
            <String CONTENT="Cqnftantinpple" HPOS="268" VPOS="148" WIDTH="273"/>
            <String CONTENT="’txanf" HPOS="1111" VPOS="156" WIDTH="89" SUBS_TYPE="HypPart1">
              <HYP CONTENT="-" HPOS="1200" VPOS="174" WIDTH="13"/>
            </String>
          </TextLine>
          <TextLine BASELINE="237" ID="B3_L1" HPOS="89" WIDTH="1123" VPOS="192" HEIGHT="61" STYLEREFS="Times New Roman_11._1">
          </TextLine>
        </TextBlock>
      </PrintSpace>
    </Page>
  </Layout>
</alto>
```

FIG. 2 – extrait du fichier ALTO

```
<?xml version="1.0" encoding="iso-8859-1"?>
<!DOCTYPE TEI PUBLIC "-//TEI P5//DTD Main Document Type//EN" "Transcription-
dictionaires-anciens.dtd">
<TEI xmlns="http://www.tei-c.org/TEI_P5/">
 <teiHeader>
  <fileDesc><!-- ..... --></fileDesc>
  <encodingDesc>
   <tagsDecl>
    <rendition xml:id="sup">text is placed higher (superscript)</rendition>
    <tagUsage gi="hi">used to mark typographic specificities</tagUsage>
   </tagsDecl>
   <charDesc>
<!-- description of specific typographic glyphs -->
    <glyph xml:id="longS">
     <glyphName>LATIN SMALL LETTER LONG S</glyphName>
     <mapping type="PUA">U+017f</mapping>
     <desc>in common use in Roman types until the 18th century,
  used here for the sound 's'</desc>
     <graphic url="xx"/>
    </glyph>
    <!--  .... -->
   </charDesc>
  </encodingDesc>
 </teiHeader>
 <text>
  <body>
   <div>
    <pb n="image_DUT01_0082" xml:id="image_DUT01_0082_p0"/>
<fw type="head" place="top-centre">ACO.</fw><lb xml:id="l0"/>
<fw type="pageno" place="top-left">105</fw><lb xml:id="l1"/>
<p xml:id="par0">
  blir un à Con<g ref="#longST">st</g>antinople, qui après la mort du
  Saint fut tran-<lb xml:id="l2"/>féré en Bithynie par Jean
  <g ref="#longS">s</g>on <g ref="#longS">s</g>ucce<g ref="#ss">ss</g>eur.
  A Jean <g ref="#longS">s</g>uccéda Mar-<lb xml:id="l3"/>cellus,
  que Nicephore a cru être l'In<g ref="#longST">st</g>ituteur des
  <hi rend="#italic">Acoemétes</hi>.Sous<lb xml:id="l4"/>
  <!--  ..... -->
</p>
   </div>
  </body>
 </text>
</TEI>
```

FIG. 3 – extrait d'un document de vérité, codé en TEI

```
<?xml version="1.0" encoding="iso-8859-1"?>
<mets xmlns:xmlns="http://www.loc.gov/METS/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xlink="http://www.w3.org/1999/xlink">
 <fileSec>
  <fileGrp ID="3_ALTO" USE="alto generated automatically via OCR">
   <file ID="ALTO_DUT01_0082" MIMETYPE="text/xml">
   <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/tei/DUT01_0082_alto.xml"/>
   </file>
  </fileGrp>
  <fileGrp ID="1_TEI" USE="Documents de vérité en TEI level 3">
   <file ID="TEI_3" MIMETYPE="text/xml">
   <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/tei/DUT01_0082_tei_3.xml"/>
   </file>
  </fileGrp>
  <fileGrp ID="22_JPG" USE="Jpeg image">
   <file ID="JPG_DUT01_0082" MIMETYPE="image/jpeg">
   <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/images_inge/DUT01_0082.jpg"/>
   </file>
  </fileGrp>
  <fileGrp ID="21_TIF" USE="Scanned image (tiff)">
   <file ID="TIF_DUT01_0082" MIMETYPE="image/tiff">
   <FLocat LOCTYPE="URL" xlink:href="file:///home/falk/read/images_inge/DUT01_0082.tif"/>
   </file>
  </fileGrp>
 </fileSec>
<!--  ..... -->
</mets>
```

FIG. 4 – extrait d'un fichier METS, représentation des composants de l'objet numérique

```
<?xml version="1.0" encoding="iso-8859-1"?>
<mets xmlns:xmlns="http://www.loc.gov/METS/">
 <structMap TYPE="logical">
  <div ID="MAP_DUT01_0082" LABEL="Numérisation du dictionnaire Trévoux">
   <fptr FILEID="ALTO_DUT01_0082"/>
   <fptr FILEID="TIF_DUT01_0082"/>
   <fptr FILEID="JPG_DUT01_0082"/>
   <fptr FILEID="TEI_3"/>
   <div ID="MAP_DUT01_0082_P0" LABEL="pages 105-106"
       TYPE="representation of the scanned image of one (or two) pages">
    <fptr FILEID="TIF_DUT01_0082"/>
    <fptr FILEID="JPG_DUT01_0082"/>
    <fptr>
     <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="image_DUT01_0082_p0"/>
    </fptr>
    <fptr>
    <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="P0_DUT01_0082"/>
    </fptr>
    <div ID="MAP_DUT01_0082_L0" LABEL="line 0, pages 105-106"
        TYPE="representation of line 0 as recognized by the OCR">
     <fptr>
      <area FILEID="TIF_DUT01_0082" SHAPE="RECT" COORDS="72,58,101,66"/>
      <area FILEID="JPG_DUT01_0082" SHAPE="RECT" COORDS="72,58,101,66"/>
     </fptr>
     <fptr>
      <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="l0"/>
     </fptr>
     <fptr>
      <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="B0_L0"/>
     </fptr>
    </div>
     <fptr>
      <area FILEID="TIF_DUT01_0082" SHAPE="RECT" COORDS="518,45,320,77"/>
      <area FILEID="JPG_DUT01_0082" SHAPE="RECT" COORDS="518,45,320,77"/>
     </fptr>
     <fptr>
      <area FILEID="TEI_3" BETYPE="IDREF" BEGIN="l1"/>
     </fptr>
     <fptr>
      <area FILEID="ALTO_DUT01_0082" BETYPE="IDREF" BEGIN="B1_L0"/>
     </fptr>
    </div>
<!-- .... -->
   </div>
 </structMap>
</mets>
```


FIG. 5 – extrait d'un fichier METS, représentation d'une carte structurelle.

- d'un ou plusieurs documents alto représentant les résultats d'un ou des OCR : DUT01_0082_alto.xml (cf. fig 2)
- du documents (textuel) final représentant le dictionnaire au niveau des entrées lexicales.
- du fichier METS DUT01_0082_mets.xml (cf. fig 4, fig 5 et fig 6) qui va répertorier :
  - les objets physiques (fichiers) et leurs localisations
  - des liens structurels entre ces objets

Un exemple d'utilisation Une foi l'objet numérique mis en place on peut assez facilement (automatiquement, même à la demande) générer une présentation web tel que la suivante :

http://www-int.loria.fr/~falk/read/DUT01_0082.html

La page web a l'image scannée en fond d'image, quand on bouge la souris sur l'image on voit aparaître des cadres représentant les blocs-lignes reconnus par l'OCR. La partie supérieure de ces cadres contiendra le texte de la ligne tel qu'il a été reconnu par l'OCR, la ligne en dessous une ligne du document de vérité. L'alignement s'est fait automatiquement, il est normal qu'il y aie des décallages entre le texte OCR et le texte du document de vérité.

Ce n'est qu'une page web en XHTML+CSS, elle peut donc être visualisée par n'importe quel browser (conforme standard). Ceci n'est qu'un exemple, beaucoup d'autres applications pouvant être mises en œuvre sur ce modèle.

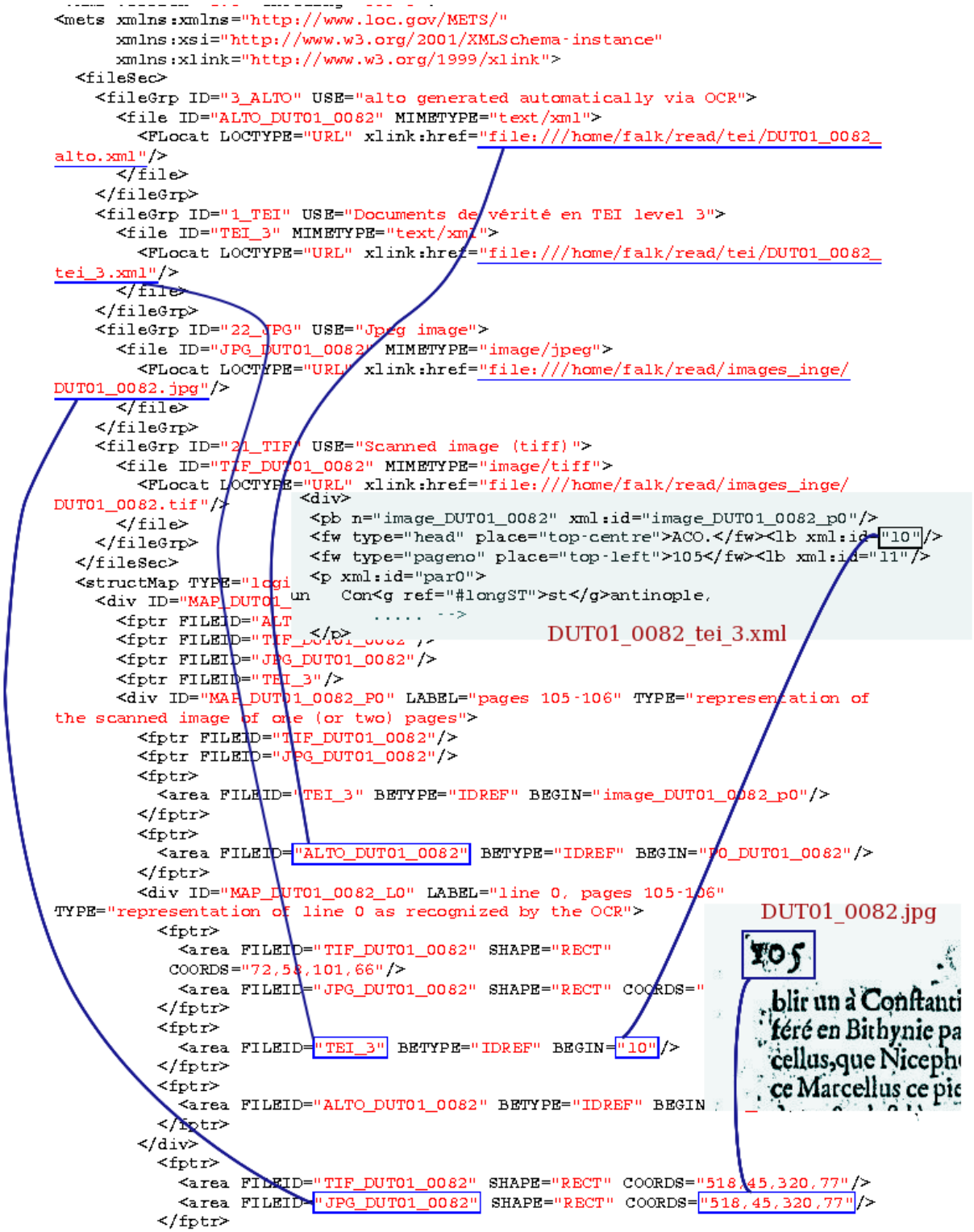

FIG. 6 – Composition d'un fichier METS

Références

[1] *The Text Encoding Initiative* Manuels, tutoriels, outils pour le codage en TEI.
http://www.tei-c.org.

[2] *Digital texts with XML and the TEI*, Text Encoding Initiative
http://www.tei-c.org/Talks/OUCS/2004-02/One/teixml-one.pdf.

[3] *Analyzed Layout and Text Object* Références et détails techniques sur le format ALTO.
http://www.ccs-gmbh.com/alto/

[4] *Metadata Encoding and Transmission Standard* Site officiel de METS, avec les spécifications, présentations et exemples d'usages.
http://www.loc.gov/standards/mets/

[5] Un article sur METS dans le Bulletin d'information francophone sur l'EAD, n°19 mars 2005.
http://www.archivesdefrance.culture.gouv.fr/fr/publications/dafbulead19.html#ancre4

[6] *Métadonnées pour les nuls, épisode 3 : METS* Un blog francophone sur METS.
http://blogokat.canalblog.com/archives/2005/06/20/589285.html

[7] *Meta Data Engine* Le site du projet européen METAe.
http://meta-e.aib.uni-linz.ac.at/

[8] *National Digital Newspaper Program* Un projet de numérisation qui utilise METS et ALTO.
http://www.loc.gov/ndnp/

[9] *Les cartulaires numérisés d'Ile de France* Un autre exemple de METS et TEI mises en œuvre par l'École de Chartes.
http://elec.enc.sorbonne.fr/cartulaires