Pour un modèle de dépôt de données adapté à la constitution de corpus de recherche

François Daoust[1], Jules Duchastel[2], Yves Marcoux[3], Élias Rizkallah[3]

[1]UQAM – Centre ATO – Québec – Canada

[2]UQAM – Chaire MCD – Québec – Canada

[3]UdeM – EBSI – GRDS – Québec – Canada

Abstract

The accessibility of research corpora in the public space is a necessity for the advancement of knowledge. This allows not only for the sharing of original texts for analysis, but alsofor the application of different analytical approaches and viewpoints. Many sociological and judicial considerations may explain why researchers resist giving access to their materials, but the concrete problem of modeling the textual data so that it can be annotated and shared is also a reason. We propose a framework for the development of data and referential metadata capable of supporting a researcher in the process of construction and deconstruction of a corpus. The model uses the Dublin Core and RDF for the metadata. As for the textual data, we are relying on the TEI model for the primary documents and their annotations, and on teiCorpus for the corpora made of documents already present in the referential. A new edition of the minutes of the Comité d'Instruction Publique, during the French Revolution, will serve as an example.

Résumé

La remise dans l'espace public des corpus de recherche est une pratique nécessaire pour l'avancement du travail scientifique. Cela permet non seulement de partager des textes sources à analyser, mais aussi d'en poursuivre l'analyse par diverses techniques et selon divers points de vue. Plusieurs facteurs sociologiques et juridiques peuvent expliquer la résistance des chercheurs à souscrire à une pratique de publication de leur corpus. Mais, il y a aussi un problème réel de modélisation des données textuelles apte à les rendre réellement utilisables dans un contexte de partage et d'annotation. Nous présentons ici une proposition de cadre de travail de type *référentiel de données et de métadonnées* supportant un modèle de données conçu spécifiquement pour le travail d'analyse sur corpus dans ses étapes de construction et de déconstruction. Ce modèle fait appel au *Dublin Core* et à *RDF* pour représenter les métadonnées. Les données textuelles font appel au schéma *TEI* pour les documents primaires et les documents d'annotation, ainsi qu'au schéma *teiCorpus* pour les corpus construits à partir des documents déjà versés dans le référentiel. Une édition nouvelle des procès-verbaux du Comité d'Instruction Publique, qui a tenu ses travaux durant la révolution française, servira d'exemple.

Mots-clés : référentiel de données et métadonnées, Dublin Core, RDF, TEI, ATO.

1. Contexte

En analyse de texte assistée par ordinateur, les chercheurs construisent des corpus raisonnés rassemblant des documents, ou extraits de documents, avec pour objectif de refléter, sous forme d'objets empiriques, des phénomènes discursifs à interpréter. Lorsqu'il s'agit de remettre ces pièces dans l'espace du débat scientifique, on doit à la fois rendre compte du corpus en tant qu'objet construit et des documents d'origine susceptibles d'être réutilisés par d'autres chercheurs qui voudront constituer leur propre corpus de recherche. L'absence d'un modèle opérationnel pour cette opération explique en partie le fait que les chercheurs hésitent à remettre leur corpus dans l'espace public. C'est là un frein réel à la recherche, car la

polémique scientifique exige qu'on puisse revenir sur les sources, réutiliser les enrichissements analytiques (annotations, éditions critiques, documents d'analyse, etc.), les compléter, les discuter, etc.

Depuis plusieurs années, les chercheurs impliqués dans l'utilisation de l'ordinateur à des fins d'analyse textuelle se réunissent et collaborent en vue de faire connaitre leurs outils, méthodes et pratiques d'analyse de texte assistée par ordinateur (ATO). C'est dans ce contexte que s'est constitué en 2005 le *Réseau pour l'échange de ressources et de méthodologies en analyse de texte assistée par ordinateur* réseau (ATONET). Après avoir convenu d'un format d'échange *XML-TEI* permettant de mettre dans l'espace public des corpus normalisés et documentés, nous avons identifié la nécessité de concevoir un modèle opérationnel de gestion de documents numériques qui soit en mesure d'accueillir les diverses couches d'annotations qui résultent du travail d'analyse textuelle. Ce cadre opérationnel s'articule autour de ce que nous appellerons un *référentiel de données et de métadonnées* soutenu par un *logiciel d'entrepôt de données* supportant un modèle de données conçu spécifiquement pour le travail d'analyse sur corpus. Ce référentiel, s'il s'apparente aux *dépôts de données institutionnels* que l'on retrouve dans le monde de l'édition électronique, par exemple pour la publication des thèses universitaires, prend ici sa particularité en ce qu'il vise à rendre compte du réseau intertextuel reliant les documents, les corpus et leurs traitements résultant en annotations, procéduriers et documents d'analyses.

2. Problématique

2.1. Introduction

Les questions théoriques en jeu ont trait à la construction-déconstruction des corpus en objets numériques riches pouvant rendre compte de la dimension documentaire selon divers niveaux de granularité tissés de relations multiples et diversifiées. La déconstruction consiste à décomposer les corpus, ou les collections de textes, en objets de nature plus atomique sans perte de richesse au niveau des relations qui les situent dans l'intertextualité et dans l'interdiscursivité. La (re)construction vise au contraire à rassembler ces pièces au sein de corpus raisonnés, objets de la démarche expérimentale du chercheur.

Plusieurs formalismes peuvent être utilisés pour représenter les diverses couches d'annotations sur un corpus, de même que les relations entre le corpus d'origine, les procédures de traitement, les résultats d'analyse et leurs interprétations. En particulier, on distingue l'annotation *in situ,* qui s'ajoute directement au document d'origine, et l'annotation *externe* qui est dégagée du corps du texte et peut même constituer un document indépendant. La notion d'organisation documentaire des corpus, et des documents qu'ils rassemblent, est donc aussi directement liée à la question de l'annotation analytique dans un contexte de débat public en constante évolution.

2.2. Un exemple complexe : les procès-verbaux du Comité d'instruction publique

Pour illustrer cette problématique, nous présentons un exemple que nous expliciterons davantage plus loin. Il s'agit d'une version électronique de l'édition nouvelle des *Procès Verbaux du Comité d'Instruction publique* des assemblées révolutionnaires en France dont les séances se sont tenues de 1791 à 1793 (Ayoub et Grenon 1997). Cette édition électronique se situe dans un projet plus vaste d'*Encyclopédie virtuelle des révolutions* dirigé par Josiane Boulad-Ayoub à l'UQAM.

En format papier, l'édition Ayoub-Grenon des procès-verbaux compte 6 354 pages. La particularité de cette édition, c'est qu'elle dévoile plusieurs *couches sédimentaires* de documents. En effet, les procès-verbaux avaient déjà fait l'objet d'une édition commentée par l'historien James Guillaume. Cette édition de 1889, s'étalant sur huit volumes, ajoute au procès-verbal de chacune des séances du Comité d'Instruction publique un ensemble d'annexes qui permettent d'éclairer les procès-verbaux. Des commentaires de liaison accompagnent ces annexes. On trouve aussi des centaines de notes ajoutées par Guillaume sur les procès-verbaux et les annexes. L'ouvrage contient également un index alphabétique et analytique des matières, un index des noms de lieux et de personnes. Enfin, on trouve des sommaires et des introductions.

L'édition de 1997 est aussi une édition augmentée qui ajoute un nouveau dispositif critique à l'édition de Guillaume. On y trouve donc de nouvelles introductions, de nouvelles notes et même de nouvelles annexes. Avec la mise en ligne de ces trois couches documentaires, impliquant une modélisation apte à rendre compte des multiples liens explicitement tressés entre les diverses pièces, il faut prévoir l'ajout de couches supplémentaires résultant du travail d'analyse futur des chercheurs. Il pourra s'agir de documents d'époque, mais aussi de nouveaux documents d'analyses et de multiples annotations sur les textes existants.

Si on peut, à la rigueur, considérer la collection comme un seul corpus, parce que constitué à des fins de publication sur papier, la déconstruction d'un tel *corpus* devient une nécessité pour permettre aux chercheurs de constituer leurs propres corpus raisonnés pouvant éventuellement, dans l'esprit de la recherche sur les révolutions, intégrer des pièces qui ne font pas partie de l'édition papier. Pour illustrer quelques-uns des problèmes soulevés par cette déconstruction, examinons quelques morceaux choisis autour de la trentième séance du *Comité de l'Instruction publique* qui s'est tenue le 25 janvier 1792.

Voici d'abord un extrait du procès-verbal lui-même.

TRENTIÈME SÉANCE

Du 25 janvier 1792

M. Vaublanc a relu le projet de décret sur les pompes triomphales. Le Comité en a adopté la nouvelle rédaction[202].

M. De Bry a lu une analyse du plan de M. Talleyrand[203].

M. Para offre au Comité trois ouvrages de sa composition : des Éléments de physique, des Principes du calcul et de la géométrie, un Cours complet de physique, le tout composant sept volumes. Le Comité arrête que le président écrira à M. Para pour lui dire que le Comité reçoit son offre avec reconnaissance[204.]

M. Lambert, ayant demandé à être autorisé à rendre à M. Métoyen le tableau en broderie qu'il avait présenté au Comité et qu'il redemandait, le Comité a approuvé que ce tableau fût rendu à la personne qui l'a présenté[205].

CONDORCET, *président* ;
ARBOGAST, LACÉPÈDE, *secrétaires*.

Le procès-verbal est accompagné d'une annexe rassemblée par l'historien Guillaume.

PIÈCES ANNEXES

Les procès-verbaux de l'Assemblée législative contiennent les indications suivantes au sujet du projet sur les récompenses militaires :

Du jeudi 26 janvier, au matin

Un membre a demandé qu'on indiquât une séance pour entendre le rapport du Comité de l'instruction publique sur les récompenses nationales à accorder aux armées qui auront combattu pour la liberté et la constitution.
Ce rapport a été ajourné à la séance de samedi[206] au soir[207].

Du samedi 28 janvier, au matin

Un membre fait la motion que le rapport du Comité de l'instruction publique sur les récompenses à décerner aux guerriers qui auront bien mérité de la patrie et qui devait être fait dans la séance de la veille, soit entendu dans celle-ci. Cette proposition est mise aux voix et adoptée.
Le rapporteur de ce Comité présente un rapport et un projet de décret sur les récompenses à accorder aux guerriers qui auront bien servi la patrie.
L'Assemblée ajourne à vendredi la seconde lecture[208] et ordonne l'impression du rapport et du projet de décret[209].

Les renvois de notes font référence à des entrées dans le fascicule des notes. Les titres et le paragraphe introductif sont de Guillaume. Voici un extrait du fascicule des notes pour l'annexe de Guillaume à la trentième séance.

206. Il faut sans doute lire *vendredi* au lieu de *samedi,* comme on le verra par l'extrait ci-après du procès-verbal de la séance du 28 janvier (qui était un samedi).
207. Procès-verbal de l'Assemblée, t. IV. p. 301.
208. La seconde lecture n'a pas eu lieu.
209. Procès-verbal de l'Assemblée, t. IV. p. 335 — Un exemplaire imprimé du rapport de Vaublanc se trouve aux Archives nationales, AD VI, 80. (Communication de M. A. Aulard.)

2.3. Les formalismes à l'appui

Avant de proposer un modèle de dépôt de données, nous présentons un certain nombre de formalismes à la base de ce modèle. Le choix de ces formalismes découle de la nature de notre démarche, qui se rapproche de celle de la *Free Bank* (Salmon-Alt, Romary, Pierrel, 2004), bien qu'elle se situe davantage dans une perspective d'analyse de discours que dans une perspective TAL d'annotation linguistique. Dans le cas de la *Text Encoding Initiative* (TEI), son choix remonte à des travaux précédents (Daoust, Marcoux, 2006).

XML est devenu le langage par excellence pour convenir d'une syntaxe concrète pour structurer les textes et les annotations sur les textes. XML, rappelons-le, est un langage général de balisage des documents électroniques qui permet de publier, conserver, annoter et transformer des textes selon un protocole indépendant des formats propriétaires. La conversion des données, des logiciels et des interfaces à la norme XML facilite grandement l'élaboration de chaines d'analyse textuelles réutilisables. La *Text Encoding Initiative* (TEI) a recours à XML pour proposer des façons de faire, des *schémas* permettant de nommer et d'organiser ces structurations. Il appartient ensuite à chaque communauté de choisir, parmi ces propositions, les formats les plus adaptés à ses données et à ses objectifs de recherche. C'est ainsi que les membres d'ATONET ont adopté une *proposition de représentation XML pour l'échange de corpus* annotés (Daoust, Marcoux, 2006). Cette proposition, conforme aux recommandations du TEI, s'inspire également des travaux du comité de l'*Organisation internationale de normalisation* dédié à la terminologie et autres ressources langagières (ISO

TC37/SC4). Cette première proposition de normalisation XML-TEI est complétée par une proposition plus élaborée basée sur le principe d'un découpage en *mots* qui permet de conserver la segmentation originale opérée par l'analyste du texte. Ce découpage permet ainsi de disposer d'un système référentiel pour l'annotation en place ou à distance au moyen de fichiers externes d'annotations. Il permet aussi de s'assurer que différents logiciels de lexicométrie puissent s'appuyer sur les mêmes unités lexicales, si l'analyste le souhaite.

L'adoption d'un format d'échange de corpus annotés ne suffit pas, cependant, à réaliser les conditions pour l'échange et le partage des ressources. Il faut encore documenter les conditions de production des documents et les règles de leur établissement en tant que ressources numériques. C'est là le domaine des métadonnées, c'est-à-dire des données sur les données. À ce niveau aussi, on retrouve divers formalismes, exprimés de plus en plus en XML, et qui permettent à une communauté de décrire ses ressources. C'est le cas notamment du *Dublin Core*, de l'*entête TEI* et de *RDF*, que nous avons retenus. L'*Encoded Archival Description* (EAD), norme pour la représentation d'instruments de recherche de fonds d'archives historiques, aurait pu être considérée au lieu de l'entête TEI, puisqu'elle permet elle aussi la représentation de métadonnées descriptives. Cependant, comme le choix de la TEI était déjà établi pour la représentation des corpus, l'intégration des métadonnées s'avérait beaucoup simple avec l'entête TEI.

Le *Dublin Core,* maintenu par le Dublin Core Metadata Initiative (DCMI), est un ensemble restreint et standard de matadonnées conçues pour la description de documents numériques. L'ensemble contient 15 champs de base tous facultatifs et répétables (ISO Standard 15836-2003). Ce noyau de base peut être complété par des champs supplémentaires (*raffinements*). Il existe aussi des groupes d'intérêt qui s'appuient sur le Dublin Core pour en proposer des usages spécialisés. C'est le cas du *Open Language Archives Community* (OLAC) qui propose des raffinements supplémentaires ou des schémas précisant le format des valeurs des éléments. Par exemple, OLAC, s'inspirant des termes de relation du format bibliographique *MARC*, propose une liste de rôles permettant de préciser l'apport des divers contributeurs dans la constitution d'une ressource numérique. Un des gros avantages du Dublin Core, c'est qu'il est directement supporté par le protocole de collecte des métadonnées qui est à la base du *Open Archive Initiative*, ce consortium qui propose un modèle permettant de fédérer les métadonnées des organismes qui consentent à les publier selon ce protocole. En s'appuyant sur ces protocoles standards, on est donc certain de pouvoir faire connaitre nos ressources. Il reste qu'on doit, en tant que communauté spécifique, convenir d'une politique d'utilisation de ces champs qui soit adaptée à nos objectifs.

RDF (*Resource Description Framework*) est un format extensible de métadonnées qui émane du *W3C*. Il permet d'exprimer des relations qualifiant une ressource numérique identifiée par un *URI* (*Uniform Resource Identifier*). Comme pour le Dublin Core, l'adhésion à des normes de marquage des métadonnées telles RDF, s'il encadre nos pratiques, ne nous dispense pas cependant de la nécessité de formaliser ces pratiques sous formes de politiques éditoriales.

Finalement, on a l'*entête TEI* qui, précédant les données textuelles elles-mêmes, les décrivent et les documentent *de l'intérieur*, si on peut dire, en ce sens que l'entête TEI est directement imbriquée dans le corpus considéré comme un seul document numérique. Le TEI propose aussi des éléments spécifiques (balises) qui permettent de pointer sur des documents, mais surtout des parties de documents, et de marquer les liens entre ces ressources.

Pour ce projet de dépôt de données adapté à la constitution de corpus de recherche, il faut réfléchir à la dualité nécessaire entre les formalismes de description des métadonnées (Dublin

Core *e*t RDF) d'une part, et, d'autre part, les formalismes de balisage de corpus de type TEI. La méthodologie que nous avons mise en place pour la recherche s'articule autour de deux dimensions interreliées. La première consiste à procéder à l'analyse de cas variés de corpus et de collections de textes. La deuxième dimension consiste en l'élaboration de chaines de traitement exploitant le modèle de données. Pour ce faire, nous avons identifié une plateforme logicielle ouverte qui offre toutes les possibilités de prototypage de nos modèles de données. Il s'agit de *Fedora* (*Flexible Extensible Digital Object Repository Architecture*), une plateforme Java développée dans un contexte universitaire (université Cornell et bibliothèque de l'université de Virginie). Basé sur un concept d'objets numériques pouvant être composés de plusieurs flux de données, Fedora est conçu sur le modèle du service WEB fournissant divers *diffuseurs* (*disseminator*) permettant de rassembler et de mettre en forme des objets numériques locaux ou distants en réponse à une requête sur les fiches de métadonnées et sur les relations entre objets.

Ces deux dimensions méthodologiques se combinent dans une approche prototypale. D'une part, nous disposons d'un ensemble de formalismes de représentation des données et métadonnées qui doivent subir l'épreuve des traitements informatiques. D'autre part, nous avons accès à plusieurs situations réelles de corpus raisonnés et de collections de documents numériques qui traduisent des pratiques discursives et analytiques. L'enjeu est de voir jusqu'à quel point nos formalismes seront aptes à représenter nos données et aussi à tester la performance de Fedora comme logiciel de dépôt de données susceptible de supporter un espace de travail fonctionnel pour la gestion de corpus numériques à des fins d'analyse.

3. Un exemple complexe : les procès-verbaux du Comité d'instruction publique

Dans cette section, nous donnons un aperçu du modèle proposé, par le truchement d'un exemple. Nous reprenons l'exemple des *Procès Verbaux du Comité d'Instruction publique* pour illustrer l'utilisation des divers formalismes présentés à la section précédente.

La modélisation de ce type de collection pose d'emblée le problème du niveau de granularité qui sera retenu pour constituer des *objets numériques* accompagnés de métadonnées aussi précises que possible. L'objet physique que constitue le document papier correspond généralement à l'unité documentaire qui fait l'objet d'une entrée dans le catalogue des bibliothèques. Il est techniquement possible de produire une édition électronique qui reproduirait ce niveau de granularité. On aurait alors un document fortement structuré en termes de balises TEI afin de rendre compte de la multiplicité des unités textuelles et de leurs relations. La constitution d'un objet d'une telle complexité est une tâche difficile qui a pour inconvénient de figer un état de la collection alors que la dynamique de la recherche voudra au contraire la reconfigurer sans cesse en la spécialisant et en l'augmentant tout à la fois.

Il serait assez logique de considérer le procès-verbal d'une séance comme un objet numérique dont la fiche *Dublin Core* pourra nous donner les auteurs, la date, la référence, etc. Les notes produites par Guillaume cent ans plus tard constituent un document distinct annotant le premier. On pourra aussi trouver des notes des auteurs de l'édition de 1997. Ces notes devraient aussi constituer un objet distinct avec une fiche *Dublin Core* distincte.

L'annexe au procès-verbal a été rassemblée par Guillaume cent ans après la tenue des séances du Comité. À ce titre, il s'agit d'un document distinct possédant sa propre fiche de métadonnées. Il en est de même des notes sur les annexes. Cependant, ces annexes sont elles-mêmes des objets composites. Ici, par exemple, on a deux extraits des procès-verbaux de

l'Assemblée législative correspondant à deux sessions distinctes. Guillaume introduit et situe ces extraits dans le corps même de l'annexe, en plus d'y ajouter des renvois à des notes publiées à part dans l'édition 1997 des procès-verbaux. Dans la mesure où il s'agit ici d'extraits de procès-verbaux, dont l'intégral pourrait se retrouver dans l'édition électronique des procès-verbaux de l'Assemblée législative, on ne voit pas l'avantage de déconstruire l'annexe en unités plus atomiques. On pourra donc considérer ici un modèle composite avec une fiche de métadonnées indiquant qu'il s'agit d'extraits de procès-verbaux introduits par James Guillaume. Voici un exemple, en format libre, de fiche *Dublin Core* pour cette annexe.

Fiche *Dublin Core* de l'annexe (en format libre)

dc:title | Annexe au procès-verbal de la trentième séance des procès-verbaux du comité d'instruction publique de l'Assemblée législative : version électronique.
dc:description | Extrait du procès-verbal de l'Assemblée législative, France, 26 et 28 janvier 1792.
dc:creator | Ayoub, Josiane.
dc:contributor | Assemblée législative de France; Guillaume, James
dc:publisher | Université du Québec, Projet d'encyclopédie virtuelle des révolutions
dc:date | 2007-10-01
dc:identifier | http://corpus.ato.uqam.ca/corpus/evr/cip-legis-pv30-annexe.xml
dc:format | text/xml
dc:source | France. Assemblée nationale législative (1791-1792). Comité d'instruction publique. Procès-verbaux du Comité d'instruction publique de l'Assemblée législative, publiés et annotés par J. Guillaume. - Edition nouvelle présentée, mise à jour et augmentée par J. Ayoub et M. Grenon. Volume 2, Fascicule 1 (Séances, annexes et appendices), pp 354-355. Paris : L'Harmattan, 1997. ISBN: 2738457916.
dc:language | fr
dc:coverage | France 1792-01-26 1792-01-28
dc:rights | http://creativecommons.org/licenses/by-nc-sa/2.5/ca/
dc:relation |http://corpus.ato.uqam.ca/corpus/evr/cip-legis-pv30-notes.xml ; http://corpus.ato.uqam.ca/corpus/evr/cip-legis-pv30.xml

Les noms des divers champs de la fiche sont introduits par dc:title, dc:description, etc. En format XML, la fiche est plus précise en ce qu'elle permet d'indiquer des formats qui décrivent davantage la sémantique et la syntaxe des entrées. Même si le champ dc:relation permet d'indiquer que des ressources supplémentaires peuvent être pertinentes, la nature de ces relations est mieux décrite par des relations RDF. Voici un exemple de définition RDF reliant l'annexe au procès-verbal.

Document *RDF* lié à l'annexe (en format XML)

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:evr="http://www.evr.org/termes/"> <rdf:RDF>
 <rdf:Description rdf:about="http://corpus.ato.uqam.ca/corpus/evr/cip-legis-
pv30-notes.xml">
 <annexe rdf:resource="http://corpus.ato.uqam.ca/corpus/evr/cip-legis-
pv30.xml"/>
 <dc:creator rdf:resource="Ayoub, Josiane">
 </rdf:Description>
 </rdf:RDF>
```

Dans cet exemple de définition RDF, la relation a*nnexe* est définie dans l'espace de noms du projet EVR. On fait aussi appel à l'espace de nom du Dublin Core *xmlns:dc* pour montrer que les entrées du Dublin Core peuvent aussi s'exprimer dans un document RDF.

Imaginons qu'un chercheur veuille constituer un corpus regroupant le procès-verbal de la trentième session du CIP, les annexes ajoutées par James Guillaume ainsi que ses notes. La liste de ces documents devrait lui être fournie suite à une requête au moteur de recherche des métadonnées, à la manière d'une recherche dans le catalogue électronique d'une bibliothèque. Cochant les documents qu'il juge pertinents, le chercheur voudra que le système de dépôt de données assemble ces pièces sous forme d'un corpus analysable par ses outils d'analyse textuelle. Normalement, ce corpus devrait être produit en format TEI quitte à être soumis par la suite à un filtre de conversion vers le *format propriétaire* d'un logiciel particulier, comme celui développé par le réseau ATONET.

Voyons d'abord la transcription en *TEI minimal* du texte de l'annexe et du document de notes. Dans ce format, on utilise un balisage non hiérarchique faisant appel à des balises sans contenu qui servent d'éléments frontières (*milestone*) découpant le corpus en zones. Dans cet exemple, ces zones traduisent le marquage initial du document sous forme de noms de style dans le traitement de texte. Des balises vides sont aussi utilisées pour rendre compte des frontières physiques de l'édition papier en termes de page (<*pb/*>) et de ligne (<*lb/*>). Les paragraphes sont encadrés par <*p*> <*/p*> et les mots par <*w*> <*/w*>. Ce découpage en mots, accompagné d'identifiants uniques, a pour objectif d'indiquer aux logiciels de textométrie les unités (token) qui devront servir aux comptages. Mais, ce découpage servira aussi de points d'ancrage pour référer à des parties du document par des pointeurs externes, par exemple pour accrocher le contenu d'une note au mot qui contient l'appel de la note. Comme pour tout document TEI, le contenu du texte est précédé d'une entête qui reprend des éléments de la fiche Dublin Core en plus de fournir divers renseignements sur le codage du texte. On notera que la transcription TEI de l'annexe ne fait pas mention des notes numérotées puisque celles-ci sont considérées comme une annotation externe. Les commentaires de liaison de Guillaume sont cependant marquées par le *milestone* portant l'attribut *unit="partie"* et qui permet de distinguer minimalement les parties du texte.

Annexe à la session 30 du CIP (en format TEI minimal)

```
<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-
c.org/ns/1.0">
 <teiHeader>
  <fileDesc>
   <titleStmt> <title>Annexe au procès-verbal de la trentième séance des
procès-verbaux du comité d'instruction publique de l'Assemblée législative :
version électronique</title>
    <principal>Ayoub, Josiane</principal>
    <notesStmt> <note>Extrait du procès verbal de l'Assemblée législative,
France, 26 et 28 janvier 1792</note>       <note>Des notes sur le texte
figurent dans un fichier séparé.</note> </notesStmt>
  </titleStmt>
  <publicationStmt>
    <publisher>Université du Québec, Projet d'encyclopédie virtuelle des
révolutions</publisher>
    <pubPlace>Québec, Canada</pubPlace> <date>2007-10-01</date>

<availability>http://creativecommons.org/licenses/byncsa/2.5/ca/</availability>
  </publicationStmt>
```

```
  <sourceDesc> <bibl> France. Assemblée nationale législative (1791-1792).
Comité d'instruction publique. Procès-verbaux du Comité d'instruction publique
de l'Assemblée législative, publiés et annotés par J. Guillaume. - Edition
nouvelle présentée, mise à jour et augmentée par J. Ayoub et M. Grenon. Volume
2, Fascicule 1 (Séances, annexes et appendices), p 74. Paris : L'Harmattan,
1997. ISBN: 2738457916</bibl> </sourceDesc>
  </fileDesc>
  <profileDesc> <langUsage> <language ident="fr">Français</language>
</langUsage> </profileDesc>
  <encodingDesc> <refsDecl> <p>Les balises «milestone n="valeur-de propriété"
unit="nom-de-propriété"» concernent les mots qui suivent la balise jusqu'à
l'apparition d'un nouveau milestone de même «unit». Les références de
pagination utilisent les balises pb (début de page), lb(début de ligne) et w
(word).</p> </refsDecl> </encodingDesc>
 </teiHeader>
 <text>
  <body>
<pb n="cip-pv030a0/74"/>
<p><!-- *{ref:: 2-7384-5791-6, p.74} --> <lb n="1"/><milestone unit="partie"
n="NoteInterne"/><w xml:id="w2">Les</w> <w xml:id="w3">procès-verbaux</w> <w
xml:id="w4">de</w> <w xml:id="w5">l'</w><w xml:id="w6">Assemblée</w> <w
xml:id="w7">législative</w> <w xml:id="w8">contiennent</w> <w
xml:id="w9">les</w> <w xml:id="w10">indications</w> <w
xml:id="w11">suivantes</w> <lb n="2"/><w xml:id="w13">au</w> <w
xml:id="w14">sujet</w> <w xml:id="w15">du</w> <w xml:id="w16">projet</w> <w
xml:id="w17">sur</w> <w xml:id="w18">les</w> <w xml:id="w19">récompenses</w> <w
xml:id="w20">militaires</w> <w xml:id="w21">:</w> </p>

<p><lb n="3"/><milestone unit="partie" n="SéanceDate"/><w xml:id="w23">Du</w>
<w xml:id="w24">jeudi</w> <w xml:id="w25">26</w> <w xml:id="w26">janvier</w><w
xml:id="w27">,</w> <w xml:id="w28">au</w> <w xml:id="w29">matin</w> </p>

<p><lb n="4"/><milestone unit="partie" n="Texte"/><w xml:id="w31">Un</w> <w
xml:id="w32">membre</w> <w xml:id="w33">a</w> <w xml:id="w34">demandé</w> <w
xml:id="w35">qu'</w><w xml:id="w36">on</w> <w xml:id="w37">indiquât</w> <w
xml:id="w38">une</w> <w xml:id="w39">séance</w> <w xml:id="w40">pour</w> <w
xml:id="w41">entendre</w> <w xml:id="w42">le</w> <w xml:id="w43">rapport</w> <w
xml:id="w44">du</w> <w xml:id="w45">Comité</w>
<lb n="5"/><w xml:id="w47">de</w> <w xml:id="w48">l'</w><w
xml:id="w49">instruction</w> <w xml:id="w50">publique</w> <w
xml:id="w51">sur</w> <w xml:id="w52">les</w> <w xml:id="w53">récompenses</w> <w
xml:id="w54">nationales</w> <w xml:id="w55">à</w> <w xml:id="w56">accorder</w>
<w xml:id="w57">aux</w> <w xml:id="w58">armées</w>
<lb n="6"/><w xml:id="w60">qui</w> <w xml:id="w61">auront</w> <w
xml:id="w62">combattu</w> <w xml:id="w63">pour</w> <w xml:id="w64">la</w> <w
xml:id="w65">liberté</w> <w xml:id="w66">et</w> <w xml:id="w67">la</w> <w
xml:id="w68">constitution</w><w xml:id="w69">.</w> </p>

<p><lb n="7"/><w xml:id="w71">Ce</w> <w xml:id="w72">rapport</w> <w
xml:id="w73">a</w> <w xml:id="w74">été</w> <w xml:id="w75">ajourné</w> <w
xml:id="w76">à</w> <w xml:id="w77">la</w> <w xml:id="w78">séance</w> <w
xml:id="w79">de</w> <w xml:id="w80">samedi</w> <w xml:id="w81">au</w> <w
xml:id="w82">soir</w><w xml:id="w83">.</w> </p> <!-- ... -->
  </body>
 </text>
</TEI>
```

Dans notre modèle, les notes sur l'annexe constituent un document autonome dont le lien logique avec le texte de l'annexe est inscrit dans les métadonnées sous la forme d'une relation RDF. Cependant, à l'intérieur même du document de notes, on devra retrouver des structures

de pointage spécifiques permettant de lier le texte de chacune des notes avec le mot auquel il se raccroche. Mis à part ce mécanisme de pointage, le codage du document de notes suit le même modèle que l'annexe annotée. L'entête TEI contient la partie documentaire alors que le texte lui-même utilise des balises de type *milestone* pour rendre compte des frontières en parties logiques et physiques du texte. Le mécanisme de pointage se retrouve à la fin du document et consiste en éléments <*link/*> reliant, pour chacune des notes, le texte de la note, exprimé comme un empan (*range*) entre les premier et dernier mots. Ces deux liens sont encadrés par une balise <*LinkGrp*> qui donne des indications sur la sémantique d'utilisation de ces liens.

Notes de Guillaume sur l'annexe (en format TEI minimal)

```
<?xml version="1.0" encoding="utf-8"?> <TEI xmlns="http://www.tei-
c.org/ns/1.0">
<teiHeader>
<fileDesc>
<titleStmt><title>Notes sur l'annexe au procès-verbal de la trentième séance
des procès-verbaux du comité d'instruction publique de l'Assemblée législative
: version électronique</title></titleStmt>
<publicationStmt> <p>Université du Québec, Projet d'encyclopédie virtuelle des
révolutions</p></publicationStmt>
<sourceDesc> <bibl>France. Assemblée nationale législative (1791-1792). Comité
d'instruction publique. Procès-verbaux du Comité d'instruction publique de
l'Assemblée législative, publiés et annotés par J. Guillaume. - Edition
nouvelle présentée, mise à jour et augmentée par J. Ayoub et M. Grenon. Volume
2, Fascicule 2 (Notes et index), pp 354-355. Paris : L'Harmattan, 1997. ISBN:
2738457916</bibl> </sourceDesc>
</fileDesc>
<encodingDesc> <refsDecl>
<p>Les balises «milestone n="valeur-de propriété" unit="nom-de-propriété»
concernent les mots qui suivent la balise jusqu'à l'apparition d'un nouveau
milestone de même «unit». Les références de pagination utilisent les balises pb
(début de page), lb(début de ligne) et w (word).</p>

<p>milestone partie symbol "SéanceNo" "NoteNo" "NoteTxt"</p>
</refsDecl> </encodingDesc>
</teiHeader>
<text>
<body>
<pb n="cip-pv030a0ng/354"/>
<p><!-- *{ref:: 2-7384-5791-6, p.354-355} --><lb n="1"/><milestone
unit="partie" n="SéanceNo"/><w xml:id="w2">30_e</w> <w xml:id="w3">séance</w>
</p>

<!-- ... --> <p><lb n="4"/><milestone unit="partie" n="NoteNo"/><w
xml:id="w43">207</w><w xml:id="w44">.</w> <milestone unit="partie"
n="NoteTxt"/><w xml:id="w45">Procès-verbal</w> <w xml:id="w46">de</w> <w
xml:id="w47">l'</w><w xml:id="w48">Assemblée</w><w xml:id="w49">,</w> <w
xml:id="w50">t</w><w xml:id="w51">.</w> <w xml:id="w52">IV</w><w
xml:id="w53">.</w> <w xml:id="w54">p</w><w xml:id="w55">.</w> <w
xml:id="w56">301</w><w xml:id="w57">.</w> </p> <!-- ... -->

<linkGrp type="note-appel" targFunc="NoteTexte NoteAppel">
<!-- ... --> <link xml:id="n207" targets="#range(#w43,#w57) cip-
pv030a.xml#w83"/> <!-- ... -->
<\linkGrp>
</body>
</text>
</TEI>
```

Ces deux documents pourront faire partie du corpus qui sera constitué suite à la requête du chercheur dans la base de métadonnées. La figure 1, à la fin de l'article, donne une représentation schématique du rapport entre les métadonnées et les textes, textes individuels d'une part et corpus composites d'autre part. Dans ce schéma, on retrouve un bloc de métadonnées par document. Le Dublin Core (DC) renvoie à la description du contenu du document tandis que le RDF indique les relations entre documents. Une de ces relations consiste à indiquer dans quels corpus se retrouvent un document. Les documents eux-mêmes sont en format TEI. Les corpus utilisent le schéma teiCorpus qui permet de fédérer des documents TEI individuels. Un mécanisme d'inclusion permet de faire référence au document individuel sans le dupliquer physiquement dans la base de données.

Il existe plusieurs façons de représenter un corpus annoté en TEI. Prenons seulement le cas des notes. On peut procéder à la fusion du document contenant les notes avec le document annoté en insérant un élement <*note*> en lieu et place de l'appel de note. Il est aussi possible de conserver intégralement le document contenant les notes comme un texte autonome dans le corpus. C'est la voie que nous choisirons ici en s'appuyant sur le TEI qui définit un document de type corpus comme un *texte composite* rassemblant des textes individuels possédant leur propre entête. L'élément <*teiCorpus*> sera donc utilisé pour rassembler les éléments <*TEI*> employés pour baliser les documents individuels. Le mécanisme des pointeurs utilisé pour relier le texte des notes avec les mots annotés pourra ainsi être conservé. On notera cependant que les identificateurs d'éléments *xml:id* ont dû être normalisés pour s'assurer de leur unicité dans le corpus.

Corpus rassemblant les documents afférents à la session 30 du CIP

```
<?xml version="1.0" encoding="utf-8"?> <teiCorpus xmlns="http://www.tei-
c.org/ns/1.0">
 <teiHeader type="corpus">
  fileDesc> <titleStmt> <title>Procès-verbal de la trentième séance des procès-
verbaux du comité d'instruction publique de l'Assemblée législative avec notes
et annexes: version électronique</title> </titleStmt> <!-- ... --> </fileDesc>
  <encodingDesc>  <refsDecl> ...  </refsDecl>  </encodingDesc>
 </teiHeader>
 <TEI>
  <teiHeader type="text">
   fileDesc> <titleStmt> <title>Procès-verbal de la trentième séance des
procès-verbaux du comité d'instruction publique de l'Assemblée législative :
version électronique</title> </titleStmt>  </fileDesc>
  </teiHeader>
  <text/>
 </TEI>
 <TEI>
  <teiHeader type="text">
   <fileDesc> <titleStmt><title>Annexe au procès-verbal de la trentième séance
des procès-verbaux du comité d'instruction publique de l'Assemblée
législative : version électronique</title> </titleStmt> <!-- ... -->
</fileDesc>
  </teiHeader>
  <text>
<!-- ... --> <p><lb n="7"/><w xml:id="cip-pv030a-w71">Ce</w> <w xml:id="cip-
pv030a-w72">rapport</w> <w xml:id="cip-pv030a-w73">a</w> <w xml:id="cip-pv030a-
w74">été</w> <w xml:id="cip-pv030a-w75">ajourné</w> <w xml:id="cip-pv030a-
w76">à</w> <w xml:id="cip-pv030a-w77">la</w> <w xml:id="cip-pv030a-
w78">séance</w> <w xml:id="cip-pv030a-w79">de</w> <w xml:id="cip-pv030a-
w80">samedi</w> <w xml:id="cip-pv030a-w81">au</w> <w xml:id="cip-pv030a-
w82">soir</w><w xml:id="cip-pv030a-w83">.</w> </p> <!-- ... -->
```

```
  </text>
 </TEI>
 <TEI>
  <teiHeader type="text">
   <fileDesc> <titleStmt><title>Notes sur l'annexe au procès-verbal de la
trentième séance des procès-verbaux du comité d'instruction publique de
l'Assemblée législative : version électronique</title> </titleStmt> <!-- ... --
>   </fileDesc>
  </teiHeader>
  <text>
<!-- ... --> <p><lb n="4"/><milestone unit="partie" n="NoteNo"/><w xml:id="cip-
pv030an-w43">207</w><w xml:id="cip-pv030an-w44">.</w> <milestone unit="partie"
n="NoteTxt"/><w xml:id="cip-pv030an-w45">Procès-verbal</w> <w xml:id="cip-
pv030an-w46">de</w> <w xml:id="cip-pv030an-w47">l'</w><w xml:id="cip-pv030an-
w48">Assemblée</w><w xml:id="wcip-pv030an-49">,</w> <w xml:id="wcip-pv030an-
50">t</w><w xml:id="wcip-pv030an-51">.</w> <w xml:id="cip-pv030an-w52">IV</w><w
xml:id="cip-pv030an-w53">.</w> <w xml:id="cip-pv030an-w54">p</w><w xml:id="cip-
pv030an-w55">.</w> <w xml:id="cip-pv030an-w56">301</w><w xml:id="cip-pv030an-
w57">.</w> </p><!-- ... -->

<linkGrp type="note-appel" targFunc="NoteTexte NoteAppel">
  <link xml:id="n206" targets="range(#cip-pv030an-w7,#cip-pv030an-w41) #cip-
pv030a-w80"/>
  <link xml:id="n207" targets="#range(#cip-pv030an-w43,#cip-pv030an-w57) #cip-
pv030a-w83"/> <!-- ... -->
<\linkGrp>
  </text>
 </TEI>
</teiCorpus>
```

Ce modèle de données privilégie l'annotation externe non seulement pour l'apparat critique, mais aussi pour tout balisage analytique s'appuyant sur le document répertorié dans le référentiel de données. Aussi, comme pour la *FreeBank*, le découpage en mots simples (balise TEI *w*) est utilisé comme trame de base pour référer aux unités à annoter. Le chercheur qui voudra poursuivre l'analyse pourra, selon sa perspective de recherche, importer ou pas les documents d'annotation portant sur la sélection de documents sources pertinente à la constitution de son corpus de recherche.

4. Conclusion

Notre travail de modélisation est encore en cours et sera confronté à diverses collections de textes. Le montage du prototype Fedora est également en cours. Une implantation opérationnelle exigera le développement des modules permettant de procéder à la mise en forme de corpus suite à une requête sur les métadonnées. Sous réserve de l'obtention de fonds à cet effet, il est donc possible d'envisager la mise sur pied de véritables dépôts de données destinés à la gestion de corpus à des fins d'analyse textuelle. Ce serait un grand pas pour constituer un véritable espace public permettant l'accès à des corpus en format numérique.

Références

ATONET. http://www.atonet.net

Ayoub J. et Grenon M. (1997). *Édition nouvelle, présentée, mise à jour et augmentée des procès-verbaux du comité d'instruction publique*. L'Harmattan, Paris, Montréal, ISBN 2738457916.

Ayoub J. (2007). *Encyclopédie virtuelle des révolutions* : http://corpus.ato.uqam.ca/forum/evr/.

Burnard L. and Bauman S. (2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*.

http://www.tei-c.org/release/doc/tei-p5-doc/html/.

Daoust F. et Marcoux Y. (2006). Logiciels d'analyse textuelle : vers un format XML-TEI pour l'échange de corpus annotés, in *Les Cahiers de la MSH Ledoux no. 3*, *Actes des JADT-2006*, vol. 1, pp.327-340, Presses universitaires de Franche-Comté, 2006. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2006/PDF/029.pdf.

Dublin Core Metadata Initiative. Site web : http://dublincore.org/

Encoded Archival Description. Version 2002 Official Site : http://www.loc.gov/ead/

Fedora. Site web : http://www.fedora.info/.

ISO TC37/SC4. Site web : http://tc37sc4.org/

MARC Standards. Site web : http://www.loc.gov/marc/

OLAC. Site web : http://www.language-archives.org/

Open Archive Initiative. Site web : www.openarchives.org.

Salmon-Alt, L. Romary, J.-M. Pierrel (2004). Un modèle générique d'organisation de corpus en ligne : application à la FreeBank, *Traitement Automatique des Langues*, Vol.45, n°3, pp. 145-169, 2004. ISSN 1248-9433

W3C (2000). *Harvesting RDF Statements from XLinks*. Note http://www.w3.org/TR/xlink/rdf. Editors: Ron Daniel Jr. (Metacode Technologies Inc.).

W3C (2001). *XML Linking Language (XLink) Version 1.0*. Recommandation http://www.w3.org/TR/xlink/. Editors: Steve DeRose, Brown University Scholarly Technology Group, Eve Maler, Sun Microsystems, David Orchard, Jamcracker.

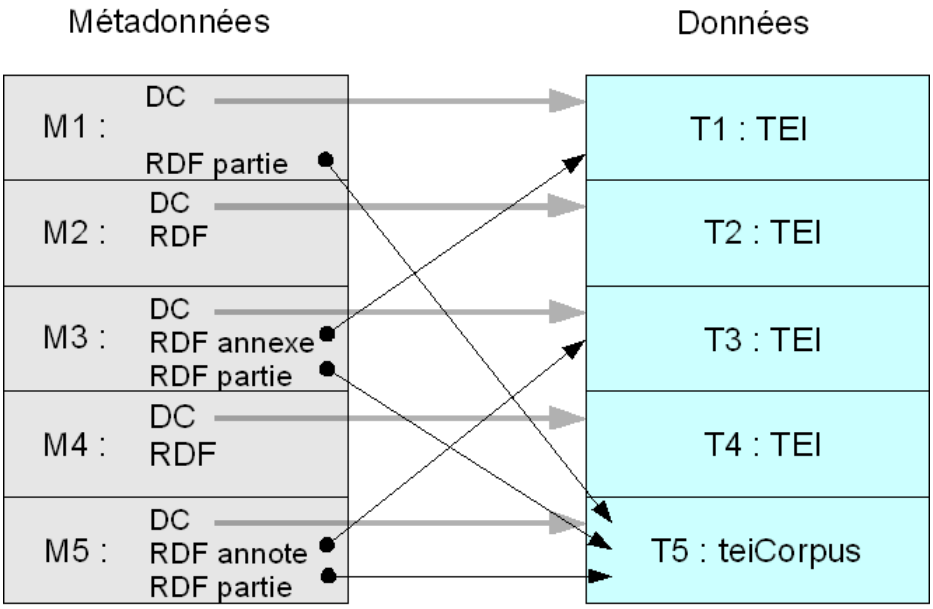

Figure 1 : Représentation des textes en objets numériques