Le balisage XML "ciblé": une nouvelle approche dans l'informatisation des corpus

Christophe Rey, Corinne Zaoui

Le balisage XML "ciblé": une nouvelle approche dans l'informatisation des corpus

Rey Christophe, Zaoui Corinne

Université de Provence, Equipe DELIC
Christophe.Rey@up.univ-aix.fr
zaoui@up.univ-aix.fr

RÉSUMÉ. *Nous proposons un balisage XML "ciblé" pour l'informatisation de dictionnaire anciens à la structure "floue". Ce dernier constitue un intermédiaire entre le balisage "minimal" et le balisage "analytique" déjà existants. Dans le cadre de cette communication, nous nous proposons de fournir un aperçu de son exploitation à travers l'élaboration d'une application C++, portable sur un grand nombre de corpus XML.*

ABSTRACT. *We propose a "targeted" XML markup to computerize old dictionaries whose structure is "vague". It stands for a markup halfway between the "minimal" markup and the "analytic" markup already available. As regards this paper, it is meant to supply a brief survey of the use of the "ciblé" mark up through the development of a C++ application which can be used on a large number of XML corpuses.*

MOTS-CLES: Encyclopédie Méthodique, Dictionnaires, Informatisation, Balisage "ciblé", XML, API DOM, C++.

KEYWORDS: Encyclopédie Méthodique, Dictionaries, Informatisation, "targeted" markup XML, API DOM, C++.

Actes de CIFT 2004, pages 121 à 133

Cette communication s'inscrit comme le résultat de travaux menés par C. Rey dans le cadre d'un doctorat de Linguistique française. Portant sur l'étude des sons de la langue au XVIIIe siècle et plus précisément sur le dictionnaire Grammaire & Littérature (1782-1786) de Nicolas Beauzée et Jean-François Marmontel tiré de l'Encyclopédie Méthodique (1782-1832), ce travail possède comme l'un de ses objectifs prioritaires la réalisation d'une version informatisée du corpus sélectionné.

Tout au long de cette communication nous souhaitons, afin de fournir une illustration originale de ce que peut être la « fouille de textes anciens», expliciter le cheminement théorique nécessaire à la réalisation d'un tel projet.

Nous allons ainsi au préalable nous intéresser aux besoins réels soulevés par l'informatisation d'un tel corpus, avant de fournir un aperçu des solutions les plus communément retenues pour ce genre d'entreprise et de présenter celles que nous proposons. La question fondamentale de l'exploitation de nos solutions sera enfin illustrée par la présentation de l'application mise en place pour la fouille des données informatisées.

1. Informatiser pour répondre à quels besoins?

L'informatisation d'un dictionnaire, quel que soit son type, constitue indubitablement une réponse à un certain nombre de besoins divers et variés qu'il est important de bien avoir à l'esprit étant donné l'incidence qu'ils possèdent dans le choix des méthodes d'informatisation retenues.

Notre entreprise a pour sa part été guidée par deux buts essentiels.

Le premier est en fait un besoin inhérent à toute entreprise d'informatisation, puisqu'il s'agit de disposer de l'ouvrage qui nous intéresse sous une forme lui assurant à la fois une pérennité, une plus grande accessibilité, et une meilleure exploitabilité.

Le second des buts que nous nous sommes fixés est celui de mettre à la disposition de tous (lexicographes, linguistes, amoureux du langage, ou autres) un corpus unique issu d'une encyclopédie relativement méconnue puisque restée dans l'ombre de son imposant prédécesseur l'Encyclopédie ou Dictionnaire raisonné: l'Encyclopédie Méthodique (1782-1832)[1].

A cheval sur deux siècles, l'Encyclopédie Méthodique (désormais EM) apparaît à la lueur des travaux actuels (Cf. Doig, 1992, Ehrard, 1991, Teysseire 1991 et 1992) comme un ouvrage clé de la mutation épistémologique opérée entre le siècle des Lumières et le XIXe siècle. Les travaux que C. Rey a pu lui-même conduire tout au long de son doctorat confirment cet état de fait pour le dictionnaire Grammaire & Littérature, et plus précisément pour ce qui concerne les théories sur les sons de la langue. Le corpus constitué s'érige en effet comme une illustration de

[1] Cf. Darnton, 1982.

cette maturation des connaissances depuis l'Encyclopédie de Diderot et D'Alembert (désormais DD[2]). Les 236 articles qui composent notre corpus représentent donc un témoignage intéressant pour l'Histoire des théories linguistiques, témoignage qu'il est donc important de pouvoir livrer au plus grand nombre sous une forme électronique autorisant de multiples recherches[3].

2. Les solutions

Bien que toutes les personnes participant à la manifestation dans le cadre de laquelle nous proposons cette communication soient déjà sensibilisées à la lecture d'un document balisé et à des technologies telles que celles que nous allons détailler, le domaine de l'informatisation des dictionnaires anciens reste un domaine avec ses particularités. C'est pour cette raison que nous allons détailler ci-dessous les différentes solutions généralement retenues pour ce genre d'entreprise.

Un bref regard sur l'horizon des travaux concernant la modélisation des données suffit à illustrer la multiplicité des solutions qui s'offrent à nous pour mener à bien notre entreprise d'informatisation[4]. C'est en suivant la piste proposée par des lexicographes pionniers comme Terence Russon Wooldridge, qui depuis le début des années 1980 oeuvrent patiemment pour la "rétroconversion" des trésors lexicographiques de notre passé, mais en nous tournant vers des technologies plus récentes et plus standards, que nous envisageons de mener à bien notre projet.

2-1 Les balisages proposés

A l'origine de la parution des éditions électroniques d'ouvrages tels que le Dictionnaire de l'Académie Française (1694) (Wooldridge, 1994), ou le Dictionnaire Critique (1787) de Jean-François Féraud (Caron, Dagenais, Gonfroy, 1992), les travaux conduits depuis ces deux dernières décennies illustrent la place importante qu'occupe la rétroconversion des ouvrages de notre passé. Il n'est ainsi pas surprenant de constater qu'à l'intérieur même de ce mouvement d'informatisation plusieurs solutions s'opposent.

---

[2] Le choix de l'abréviation DD (Diderot-D'Alembert) est justifié par le fait que notre étude repose sur une comparaison entre l'*Encyclopédie Méthodique* et les trois branches de la première encyclopédie, à savoir le *Dictionnaire raisonné* (1751-1772), le *Supplément* à l'Encyclopédie (1776-1777), et la *Table Analytique* livrée par Panckoucke (1780).

[3] Nous détaillerons plus loin les fonctionnalités que nous souhaitons attribuer à notre version électronique.

[4] Cf. Véronis, J., & Ide, N. (1996). Encodage des dictionnaires électroniques: problèmes et propositions de la TEI. In D. Piotrowsky (Ed.), Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy (29, 30, 31 mai 1995) (pp. 239-261). Paris: Didier Erudition.

S'ils ont tous deux recours à un balisage consistant à poser des jalons ou repères dans le document subissant l'informatisation, les deux grands courants en présence s'opposent précisément sur le type de balisage à adopter, puisque de celui-ci vont immédiatement découler les possibilités d'interrogation et d'appropriation du contenu du répertoire.

Sans entrer dans des considérations trop détaillées, cette opposition traduit en fait deux orientations radicalement différentes dans l'approche du document ancien. La première, reposant sur un balisage minimal, est une approche minimaliste soucieuse de la préservation de l'intégralité du document, qui traduit la volonté de restitution formelle de l'ouvrage par une utilisation de "points d'accès" au document (Wooldridge, Leroy-Turcan, 1996), d'un nombre restreint de marquages[5], et laissant donc à l'utilisateur une entière liberté d'interprétation et d'investigation.

La deuxième approche, caractérisée par un balisage analytique, est à l'inverse articulée autour du repérage systématique de chacun des champs informationnels de l'article, grâce à l'analyse du spécialiste ayant procédé au découpage systématique de l'ouvrage, qui propose ainsi une interprétation du fonctionnement de ce dernier et offre donc une grille de lecture quasi-exhaustive à l'utilisateur.

Ces deux formes d'encodage ont trouvé chacune des adhérents, puisque plusieurs projets ont vu le jour. Le balisage minimal a ainsi été utilisé pour l'informatisation du Thrésor de la Langue Françoyse de Nicot (1606), par T.R.Wooldridge, ou pour le Dictionnaire de l'Académie Françoyse par I. Leroy-Turcan et T.R.Wooldridge, ainsi que le montre la Figure 1.

TIMBRE. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. *Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon.*[...] Timbrer. v. a. Terme de blason, Accompagner
d'un timbre. *Timbrer une armoirie*.
Timbrer. v. a. Terme de Pratique, Escrire au haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient. *Timbrer des pieces*.
On dit aussi, *Timbrer du papier, timbrer du parchemin*, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.[6]

<page n="563"><col n="1">[...]<p><lc>TIMBRE</lc>. s. m. Sorte de cloche ronde qui n'a point de battant en dedans, & qui est frappée en dehors par un marteau. Le timbre d'une horloge. timbre d'un reveille-matin. le timbre de cette horloge est tres-bon.<p>
[...]
<sc>Timbrer</sc>. v.a. Terme de blason, Accompagner d'un timbre. Timbrer une armoirie.
<p><sc>Timbrer</sc>. v.a. Terme de Pratique, Escrire en haut d'un Acte, la nature de cet acte, sa date & le sommaire de ce qu'il contient. Timbrer des pieces.

[5] L'édition, la page, la colonne, les alinéas, les caractères (grandes et petites capitales, italique, gras) ou la vedette peuvent ainsi être balisés.
[6] Les codages <page n="563">, <col n="1">, <p>, <lc></lc>, , <sc></sc>, servent respectivement à marquer la page dans laquelle se situe l'article balisé, les colonnes occupées par l'article dans la page, présence d'un paragraphe, les grandes capitales (Low capitales), le caractère italique de l'information codée, et les petites capitales (Small capitales).

<p>On dit aussi, Timbrer du papier, timbrer du parchemin, pour dire, Imprimer la marque du Roy sur du papier, sur du parchemin, pour faire qu'il puisse servir aux actes de Justice.<p>

Figure 1. ***Exemple de balisage minimal (Dictionnaire de l'Académie Française 1694).***[7]

Le balisage analytique, a pour sa part été retenu par Isabelle Leroy-Turcan pour son projet d'informatisation du Dictionnaire Etymologique ou Origine de la Langue Françoise (1694) de Gilles Ménage. Il a également été retenu par Chantal Wionet et Agnès Tutin pour leur projet d'informatisation du Dictionnaire Universel de Furetière revu par Basnage de Bauval (1702) (Wionet, Tutin, 1998 et 2001), ainsi que l'illustre la Figure 2 ci-dessous, représentant un article balisé de manière analytique:

```
<Entry>
  <Form Type=LEMMA><Orth Rend=CAPS>DAGUET</Orth>. </Form>
  <GramGrp><Pos Type=S></Pos><Gen Type=M></Gen></Gramgrp>
  <Sense><CDomain><Lbl>Terme de</Lbl><Domain> Venerie</Domain>.</CDomain>
    <Def>Jeune cerf, qui est à sa premiere tête;
    qui pousse son premier bois.</Def></Sense>
  <Re> <Form Type=HOMOGRAPH><Orthre Rend=SCAPS>Daguet</Orthre>.</Form>
    <GramGrp><Pos Type=ADV>adv. </Pos></GramGrp>
    <Sense><Def>Sourdement, en cachette. </Def>
    <Eg><Q>Il s'en est allé, il a tiré ses
    chausses <Oref Rend=IT>daguet</Oref>. </Q></Eg>
    <CUsg><Lbl>Cela est</Lbl><Usg>bas et
    populaire.</Usg></Cusg></Sense></Re>
</Entry>
```

DAGUET. Terme de Venerie. Jeune cerf, qui est à sa premiere tête; qui pousse son premier bois.

Daguet. adv. Sourdement; en cachette. Il s'en est allé, il a tiré ses chausses daguet. Cela est bas et populaire.[8]

---

[7] Exemple tiré de l'article de T.R. Wooldridge, "L'informatisation du *Dictionnaire de l'Académie française* (*DAF*)", dans Actes du colloque-atelier international DictA1998 organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA), Université de Limoges, 19-20 novembre 1998.

[8] Afin de ne pas rentrer dans des détails trop complexes nous pouvons commenter ce balisage en indiquant le repérage des champs informationnels principaux tels que <Form> indiquant si l'entrée ou la sous-entrée est un lemme ou un homographe, la marque de domaine <Domain>, l'information grammaticale <GramGrp>, la définition <Def>, et des champs secondaires comme les remarques sur la typographie des entrées, des sous-entrées, ou des références,

Figure. 2. *Exemple de balisage analytique.*

Au delà de cette opposition fondamentale sur la pose des repères dans le corps des articles, les deux courants lexicographiques évoqués peuvent également être opposés en fonction du langage de balisage utilisé et des technologies associées.

Ainsi, l'approche minimaliste possède historiquement la particularité de s'être reposée sur l'utilisation d'un langage de balisage non-normatif et d'outils propriétaires, c'est à dire d'outils développés au sein d'équipes de recherches ou de laboratoires, et ne faisant pas l'objet de spécifications par un organisme de standardisation. Les plus connus de ces logiciels sont les logiciels TACT (Text Analysis Computing Tools) et WordCruncher, respectivement utilisés par T. R. Wooldridge pour les dictionnaires de la série Estienne-Nicot, puis pour la première édition du Dictionnaire de l'Académie Française, en ce qui concerne le premier, et pour l'informatisation du Dictionnaire Critique, entreprise par Philippe Caron, Louise Dagenais et Gérard Gonfroy, pour le second.

Face à ce pôle principal, l'approche du balisage analytique s'est plus orientée vers l'utilisation d'outils standards, à travers notamment le choix du Standard Generalized Markup Language (Langage normalisé de balisage généralisé, SGML), langage de balisage reconnu comme norme internationale[9]. Un tel choix vise clairement à inclure les travaux d'informatisation des dictionnaires dans une communauté de documents «standards».

2.2 Le balisage retenu

En ce qui concerne à présent notre propre entreprise d'informatisation, les choix que nous avons opérés ont été guidés à la fois par notre analyse des avantages et inconvénients des deux courants présentés ci-dessus, mais aussi par l'ampleur de la tache.

Le choix d'un balisage reposant sur un langage normatif ne faisait pour nous aucun doute (cf. REY, 1999) dans la mesure où nous sommes convaincus de la nécessité d'intégrer les dictionnaires anciens informatisés dans la galaxie des documents électroniques standardisés. Néanmoins, à la différence des travaux de Chantal Wionet et Agnès Tutin qui reposent sur le langage SGML, nous avons préféré avoir recours à l'Extensible Markup Language (Langage de balisage extensible, XML), langage qui s'est aujourd'hui incontestablement imposé comme

---

<Orth Rend=CAPS>, <Orthre Rend=SCAPS>, <Oref Rend=IT>, les marques d'usage <Usg>, les particules introduisant un champ, nommées libellés <Lbl>, etc.

[9] Norme ISO 8879.

l'un des principaux outils de codage des données électroniques (Attar et Chattel, 2000)[10].

Le choix entre le balisage minimal et le balisage analytique ne souffrait à priori lui non plus d'aucun doute en raison de la possibilité de concilier le respect de l'intégrité des documents rétro-convertis et la richesse de leur balisage informationnel. Toutefois, la nature des articles du corpus nous a fait préférer une solution intermédiaire.

En effet, la nature encyclopédique des articles de notre échantillon imposait une trop grande part de subjectivité dans la délimitation et le découpage de certains champs informationnels identifiés. La solution que nous avons privilégiée est donc celle d'un balisage ciblé, un balisage intermédiaire entre le balisage minimal et le balisage analytique.

En vue de pallier la trop grande complexité de la mise en oeuvre d'un balisage de type analytique sur des articles tels que ceux de notre corpus, c'est-à-dire des articles ne possédant pas un patron organisationnel aussi rigide que celui des dictionnaires modernes mais une structure plutôt « molle », le balisage ciblé s'attache seulement au repérage des champs informationnels les plus facilement identifiables, tels que le lemme, l'information grammaticale, la marque de domaine, les citations d'auteurs, les citations d'œuvres, etc.[11] La structure "molle" du répertoire à informatiser est illustrée ci-dessous par une série d'exemples attestant l'irrégulière distribution de l'information étymologique dans le corps des articles:

ACCENT, s.m. **Ce mot vient d'*accentum*, supin du verbe *accinere* qui vient de *ad* & *canere*** :
(N.) GUTTURAL, E, adj. Appartenant à la gorge ou au gosier. *Vaisseau guttural. Glande gutturale. Articulations, Consonnes gutturales.*
Ce mot, **tiré immédiatement du latin *Gutturalis,*** qui a le même sens, **vient du nom *Guttur* (Gorge, Gosier)**.
LABIAL, E, adj. *Gram*., qui appartient aux lèvres. **Ce mot vient du latin *labia* (les lèvres).**

Notre balisage ne s'intéresse pas à des champs comme l'énoncé définitoire ou la zone d'exemplification dont la délimitation, dans des ouvrages de type encyclopédique, est soumise à une importante part d'arbitraire. Ce balisage se rapproche donc en un sens du balisage minimal, mais diverge radicalement de celui-ci par le fait qu'il repose sur un langage faisant office de standard de codage.

[10] Le langage XML est présent dans plusieurs travaux d'informatisation de dictionnaires, et s'est notamment imposé comme le langage de codage de la version informatisée du *Trésor de la Langue Française* : http://atilf.inalf.fr/tlfv3.htm (Cf. Dendien, Pierrel, 2003).
[11] Notons que ce type de balisage peut s'appliquer à tous les types de données possédant une structure relativement aléatoire.

2.3 Explication de nos balises

Intéressons-nous à présent aux principaux jeux de balises retenus dans notre entreprise d'informatisation, en rappelant qu'ils constituent autant de clés d'accès au document informatisé.

Trois grands types de balises sont en fait à distinguer, 1) les balises repérant les champs dictionnairiques traditionnels, 2) les balises servant au repérage des informations relatives à notre analyse du corpus, et 3) les balises de style.

2.3.1 Les balises traditionnelles

Le balisage de la microstructure de nos articles se trouve illustré à travers l'agencement d'un certain nombre d'éléments dont la nature est explicitée par le choix même du nom des balises. La figure 3 ci-dessous dresse une liste des principales balises retenues:

```
<ARTICLE>
<STATUT TYPE="NOUVEAU"/>
<ENTREE TYPE="EP"><FORME>(N. ) PALATAL, E</FORME>,
<INFORMATION_GRAMMATICALE TYPE="ADJECTIF">
<PARTIE_DU_DISCOURS TYPE="ADJECTIF">
adj. </PARTIE_DU_DISCOURS>
</INFORMATION_GRAMMATICALE>
</ENTREE>
<CORPS>Apartenant au palais de la bouche. <PAIRE_MINIMALE">Les articulations palatales sont des
articulations linguales sifflantes, dont le sifflement s'exécute dans l'intérieur de la bouche, entre le milieu
de la langue & le palais. Il y en a deux en francois, j & ch, telles qu'on les entend au commencement des
mots Japon, chapon. </PAIRE_MINIMALE>Voyez <REFERENCE
TYPE="VEDETTE">ARTICULATION</REFERENCE>. Ce mot est formé du mot <LANGUE
TYPE="LATIN">palatum </LANGUE>(palais de la bouche), & n'est pourtant employé dans ce sens que
par les grammairiens. Les anatomistes disent palatin : en quoi ils dérogent mal à propos à l'analogie des
adjectifs homogènes dental, lingual, guttural ; & occasionnent d'ailleurs une équivoque, à cause de palatin
tiré de <LANGUE TYPE="LATIN">palatium </LANGUE>(palais du prince). Quelques grammairiens se
servent du mot de Palatial au lieu de palatal. En cela ils pêchent doublement : I°. contre l'usage reçu,
puisque l'<REFERENCE TYPE="OUVRAGE">Académie</REFERENCE> , le <REFERENCE
TYPE="OUVRAGE">Trévoux</REFERENCE> , & nos meilleurs vocabulistes & grammairiens ont tous
adopté Palatal ; 2°. contre l'analogie, puisque palatial ne pourroit venir que du latin <LANGUE
TYPE="LATIN">palatium </LANGUE>, & qu'on le trouve effectivement en ce sens dans le
<REFERENCE TYPE="OUVRAGE">Trévoux</REFERENCE> .
<SIGNATURE>(M.BEAUZÉE.)</SIGNATURE>
</CORPS>
</ARTICLE>
```

Figure. 3. *Balisage XML ciblé de l'article PALATAL,E.*

Les différents éléments retenus dans ce type de balisage peuvent être associés à un certain nombre d'attributs comme le montre la balise <LANGUE TYPE="LATIN"> à laquelle a été associé un attribut permettant de spécifier qu'il

s'agit d'un passage textuel en latin. Cette balise possède bien entendu autant d'attributs que l'on peut relever de langue étrangère dans le corpus.

2.3.2 Les balises relatives à notre corpus

Aux jeux de balises illustrant les champs les plus classiques et les plus facilement délimitables d'un article de dictionnaire, nous avons associés des jeux de balises en relation directe avec notre analyse des théories sur les sons de la langue dans l'EM. Ainsi avons-nous décidé de rajouter, ainsi que l'illustre l'exemple de la Figure. 3, un attribut à l'élément LEMME pour systématiquement spécifier le degré de disparité informationnel constaté entre les articles de l'EM et leur éventuel prédécesseur de la DD. Volontairement généralistes, les étiquettes utilisées comme attributs spécifient des disparités plus ou moins importantes et très finement décrites dans notre analyse. Un tel repérage se traduit par les balises <STATUT TYPE= "NOUVEAU">, <STATUT TYPE= "IDENTIQUE">, <STATUT TYPE= "PEU_DIFFERENT">, <STATUT TYPE= "DIFFERENT">, <STATUT TYPE= "TRES_DIFFERENT">.

Toujours dans le but d'illustrer la maturation des connaissances sur les sons entre la DD et l'EM, nous avons eu recours au jeu de balises <AJOUT></AJOUT>, non présent dans la Figure. 3, servant à spécifier lors de la reprise dans l'EM d'articles déjà présents dans la DD, les commentaires propres à la seconde encyclopédie et donc susceptibles d'apporter des connaissances supplémentaires.

Dans la perspective de mettre en évidence les connaissances les plus exactes possible des grammairiens-philosophes du XVIIIe siècle sur les sons de la langue, il nous a par ailleurs semblé primordial de baliser les rares manifestations d'exemples faisant office de paires minimales. Ceci s'est manifesté par la création de l'élément <PAIRE_MINIMALE>[12] présent dans la Figure 3.

Peu nombreuses, ces balises relatives à l'analyse que nous avons faite de l'évolution des sons entre la DD et l'EM, constituent en quelque sorte une strate supplémentaire venant se greffer à notre balisage "ciblé", mais dont la présence est à exclure dans la perspective de l'application de ce balisage "ciblé" à d'autres répertoires.

2.3.3 Les balises de style

Le langage XML impose la dissociation du balisage logique et du balisage physique d'un document et repose donc sur l'utilisation de feuilles de styles, du type Extensible Stylesheet Language (XSL) ou du type Cascading Style Sheet (CSS), où sont consignées toutes les informations concernant la mise en forme. En ce qui concerne plus particulièrement notre entreprise, nous nous sommes principalement attelés à la description logique du document et non à son aspect physique, dans la mesure où l'essentiel de son exploitation se fera sur son fond et non sur sa forme.

[12] Cet élément note de manière large les passages contenant un exemple de paires minimales.

3. La fouille des articles

Après la présentation des différents jeux de balises retenus, intéressons-nous à présent à la question cruciale de l'exploitation des données encodées.

Nous décrivons ci-dessous la mise en place d'une application destinée à la fouille de notre document balisé, et dont la particularité essentielle est de ne pas se limiter à notre propre corpus mais à d'autres corpus de taille moyenne.

3.1 Le stockage

Au préalable, attardons-nous sur la question du stockage des données à interroger et tranchons soit en faveur du stockage sous forme de base de données, soit en faveur du stockage sous forme de fichiers.

Malgré la gratuité de certaines bases, la solution consistant à utiliser les bases de données intégrant du XML ne semble pas véritablement adaptée au volume peu important des données des corpus considérés. C'est notamment pour cela que nous avons privilégié le recours à un stockage sous forme de fichiers dont l'analyse syntaxique est fournie par une Interface de Programmation (API, Application Programming Interface). En l'occurrence nous avons retenu la représentation hiérarchique sous forme d'arbre fournie par l'API DOM (Document Object Model)[13], API complètement indépendante de tout langage et de toute plate-forme informatique.

3.2 Le langage de requêtes

En étroite relation avec la question du stockage des données, vient ensuite celle du choix d'un langage de programmation, prenant en considération le DOM, nécessaire à la réalisation de l'application permettant l'interrogation de ces données.

Deux langages ont en fait retenu notre attention, nous autorisant à développer deux applications distinctes. La première repose sur le langage PHP, et la seconde, celle que nous détaillons plus précisément dans cette communication[14], sur l'association du langage C++ et des librairies QT[15].

Bien que reposant sur la même démarche, à savoir la récupération des balises grâce au DOM, l'application C++/QT est destinée à proposer aux utilisateurs une

---

[13] Pour plus d'information se référer au site suivant: http://www.w3.org/DOM/

[14] Notons que le choix de ce langage, plutôt qu'un autre, s'explique en partie par notre volonté d'utiliser des outils les plus à la pointe possible de la dynamique de standardisation des documents échangés.

[15] Kalle Dalheimer, Matthias, 1999, *Programmer avec Qt*, O'Reilly.

interrogation du corpus "en local"[16], alors que l'approche PHP leur offre la possibilité d'une consultation sur Internet.

A l'aide de l'association C++/QT et du DOM, il est possible, comme nous allons le voir à présent, de récupérer la totalité d'un document XML et d'interroger sa structure selon les critères de notre choix. Le processus à adopter est décrit ci-dessous.

On commence par charger le document en mémoire. Sachant que tout document XML possède une structure arborescente, le DOM permet de parcourir les différentes branches du document et de récupérer son contenu aussi bien au niveau de sa structure physique que de son contenu. Enfin, le code C++ permet de récupérer l'ensemble des balises dans le document XML.

L'application élaborée permet alors une exploitation souple et aisée du balisage. Grâce au parcours du document XML on récupère dans des listes déroulantes l'ensemble des éléments et des attributs. Le choix d'un élément permet alors l'affichage des divers attributs qui lui sont associés et la liste et le nombre des articles qui sont concernés par cette sélection.

Dans certains cas, si les attributs d'un élément possèdent plusieurs formes graphiques, l'utilisateur peut opérer une sélection parmi elles.

Sans avoir procédé à un balisage très strict de la microstructure des articles, il nous est alors très facile et très rapide de nous livrer à une exploitation fine et poussée du corpus. Il est ainsi possible d'extraire de ce dernier la liste des ouvrages auxquels il est fait référence, et notamment ceux mentionnés dans l'article PALATAL, E (Cf. Figure 3). Il nous est de même aisé de vérifier que ce même article PALATAL, E fait partie des articles comportant des passages en latin et signés par le grammairien Nicolas Beauzée.

Le fait d'utiliser des interfaces basées sur des listes déroulantes chargées à partir de la récupération des balises et de leurs attributs dans le document XML, permet l'utilisation de l'application sur d'autres corpus balisés avec ce langage.

Conclusion

Bien plus qu'une simple illustration de la fouille de données informatisées, le projet dont nous avons tracé ici les grandes lignes, constitue une approche novatrice en matière d'informatisation des dictionnaires. Le balisage ciblé constitue une alternative intéressante entre le balisage minimal offrant des perspectives limitées d'exploitation du document et le balisage analytique dont la complexité s'avère trop importante sur des articles dictionnairiques de type encyclopédique.

---

[16] Nous proposons à l'utilisateur de télécharger sur sa propre machine l'application nécessaire à la fouille du document balisé.

En faisant reposer notre application sur l'API DOM qui décrit la structure des fichiers XML sous forme d'arbre, nous assurons sa portabilité sur d'autres corpus de taille moyenne.

Bibliographie

Sources primaires

Diderot D., Alembert, Jean Le Rond d'. (1751-1772). *Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers*.

Mouchon P., Diderot, D., Alembert, Jean Le Rond d'. (1780). *Table analytique et raisonnée des matieres contenues dans les XXXIII volumes in-folio du dictionnaire des sciences, des arts et des métiers, et dans son supplément*. Paris, Panckoucke.

Panckoucke, Ch-J. (1782-1832). *Encyclopédie méthodique ou par ordre de matières par une société de gens de lettres, de savants et d'artistes; précédée d'un Vocabulaire universel, servant de Table pour tout l'Ouvrage, ornée des Portraits de MM. Diderot et d'Alembert, premiers Éditeurs de l'Encyclopédie.* Paris, Panckoucke.

Robinet, J.B., Diderot, D., Alembert, Jean Le Rond d'. (1776-1777). *Supplément à l'Encyclopédie ou dictionnaire des sciences, des arts et des métiers / par une société des gens de lettres ; mis en ordre et publié par M.***.*

Sources secondaires

Attar, Pierre, Chatel, Bruno, "Etat des recommandations XML dans le domaine documentaire"*, Cahiers GUTenberg* n° 37-38, décembre 2000, 53-85.

Darnton, Robert, , *L'Aventure de l'Encyclopédie*. *1775-1800*. *Un best-seller au siècle des Lumières*, Paris, Perrin, 445 p. Ill. Traduction de Marie-Alyx Revellat. Préface d'Emmanuel Le Roy Ladurie, 1982 (1979).

Dendien, Jacques, PIERREL, Jean-Marie, "Le trésor de la Langue Française informatisé: un exemple d'informatisation d'un dictionnaire de langue de référence", *TAL*. Volume 44 - n°2/2003, 28 p.

Doig, Kathleen .H, "L'Encyclopédie méthodique et l'organisation des connaissances", *Recherches sur Diderot et sur l'Encyclopédie*, *12 (1992)*, pp. 59-69.

Ehrard, Jean, "De Diderot à Panckoucke : Deux pratiques de l'alphabet", *L'Encyclopédisme: actes du Colloque de Caen*, *12-16 janvier 1987*. - Paris, 1991, pp. 234-252.

Kalle Dalheimer, Matthias, *Programmer avec Qt*, O'Reilly, 1999.

Leroy-Turcan, Isabelle, Wooldridge, Terence. Russon, "L'informatisation du Dictionnaire de l'Académie française", *Actes du colloque DictA1998, Table ronde sur l'informatisation des dictionnaires anciens*, Limoges, 19-20 novembre 1998.

Leroy-Turcan, Isabelle, "Modalités de mise en oeuvre de l'informatisation de la première édition du Dictionnaire de l'Académie française (1694)", *Actes des Journées Dictionnaires électroniques des XVIe- XVIIe s*, Clermont-Ferrand, 14-15 juin 1996.

Teysseire, Daniel, "A propos de l'Encyclopédie Méthodique", *Recherches sur Diderot et sur l'Encyclopédie*, 11(1991), pp. 142-149, 1991.

Rey, Christophe, *Informatisation des dictionnaires anciens: l'exemple du métalangage grammatical dans le Dictionnaire françois de César-Pierre Richelet*. Mémoire de Diplôme d'Etudes Approfondies réalisé sous la direction de Véronis Jean et sous la co-direction de Wionet Chantal, 1999.

Véronis, J., & Ide, N. (1996). Encodage des dictionnaires électroniques: problèmes et propositions de la TEI. In D. Piotrowsky (Ed.), Lexicographie et informatique - Autour de l'informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy (29, 30, 31 mai 1995) (pp. 239-261). Paris: Didier Erudition.

Wionet Chantal, Tutin Agnès, *Pour informatiser le Dictionnaire universel de Basnage (1702) et de Trévoux (1704) Approche théorique et pratique*. Honoré Champion.

Wionet Chantal, Tutin Agnès, "Informatisation du Dictionnaire Universel de Furetière revu par Basnage de Bauval (1702) : premier bilan", *Actes du colloque-atelier international DictA1998 organisé par le Groupe d'Études sur l'Histoire de la Langue Française (GEHLF) et la Société Internationale d'Études Historiques et Linguistiques des Dictionnaires Anciens (SIEHLDA)*, Université de Limoges, 19-20 novembre 1998, 2001.

Wooldridge, T.R., "Projet d'informatisation du Dictionnaire de l'Académie (1694-1935)", *Actes du Colloque international Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne*, Institut de France, novembre 1994; (ed. B. Quemada & J. Pruvost), Paris, Champion: 309-20, 1994.

Wooldridge, T.R., Leroy-Turcan Isabelle, « Les mots-clefs métalinguistiques comme outil d'interrogation structurante des dictionnaires anciens », *Lexicomatique et dictionnairiques* (éd. A. Clas, P. Thoiron & H. Béjoint), Beyrouth: FMA & Montréal: AUPELF-UREF, 1996, pp. 307-16.