Le balisage souple ou flottant : une piste pour l'informatisation des données encyclopédiques anciennes

Christophe Rey

Christophe Rey – Université de Provence (Equipe DELIC)

Le balisage *souple* ou *flottant* : une piste pour l'informatisation des données encyclopédiques anciennes

Longtemps restée dans l'ombre de la célèbre *Encyclopédie* ou *Dictionnaire raisonné* de Diderot et d'Alembert (1751-1780), dont elle constitue une édition profondément remaniée, l'*Encyclopédie Méthodique* (1782-1832) éditée par Charles-Joseph Panckoucke apparaît d'après plusieurs travaux (Cf. Ehrard (1991), Doig (1992), Douay (1994 et 1996), Teysseire (1991 et 1992), Rey (2004)) comme un ouvrage opérant une importante mutation épistémologique entre le siècle des Lumières et le XIX^{e} siècle.

Les travaux que j'ai pu conduire tout au long de mon doctorat confirment cet état de fait pour l'un des trente neuf dictionnaires de matière qui composent cette encyclopédie, le dictionnaire *Grammaire & Littérature* (1782-1832), et plus précisément pour ce qui concerne les théories sur les sons de la langue. Le corpus des articles traitant des sons dans le dictionnaire de Nicolas Beauzée (pour la partie Grammaire) et Jean-François Marmontel (pour la partie Littérature) illustre en effet cette maturation des connaissances depuis l'*Encyclopédie* de Diderot et d'Alembert. Les 236 entrées de dictionnaire qui composent notre corpus représentent donc un témoignage intéressant pour l'Histoire des théories linguistiques, témoignage qu'il est important de pouvoir livrer au plus grand nombre sous une forme électronique autorisant de multiples recherches. Nous nous proposons de fournir ici une illustration d'un balisage "souple" ou "flottant" mis en place pour l'informatisation de données de nature encyclopédique et ne répondant pas aux dispositions traditionnelles adoptées pour la rétroconversion des données anciennes.

1. Les solutions de rétroconversion existantes

Un bref regard sur l'horizon des travaux concernant la modélisation des données suffit à illustrer la multiplicité des solutions qui s'offraient à nous pour mener à bien l'informatisation de notre corpus. 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 envisagions de mener à bien notre projet.

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 courants scientifiques s'opposent.

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, 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.

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).

2. Le balisage Souple ou Flottant

Les possibilités respectives qu'offraient le *balisage minimal* et le *balisage analytique*, ne répondaient pas véritablement à nos attentes et nous ont poussé à privilégier l'utilisation d'un balisage intermédiaire : le *balisage souple* ou *flottant*. La création de ce balisage visait donc à pallier les difficultés sous-jacentes aux deux formes de balisage précédentes sur une structure aussi complexe que la structure encyclopédique des articles de notre corpus.

2.1 L'identification "large" des champs informationnels

La solution la plus séduisante vers laquelle nous aurions pu nous orienter, étant donné que nous ne voulions pas fournir une édition électronique de notre corpus offrant aussi peu de moyens d'interrogation des données que ceux proposés par le *balisage minimal*, aurait été celle du *balisage analytique*. Toutefois, la structure encyclopédique complexe et molle[1] des articles de notre corpus n'autorisait pas la mise en application d'un balisage analytique tel qu'il est appliqué par Wionet et Tutin. Un tel balisage semblait en effet ne pas être complètement adapté à notre étude, bien que nous puissions également délimiter la totalité des champs identifiés dans les autres entreprises de rétroconversion.

Le balisage de champs aussi complexes que la *définition,* la *contextualisation*, ou le *discours encyclopédique*, nous a permis de nous apercevoir que la difficulté essentielle n'était pas tant de repérer la totalité des champs informationnels – puisque ces derniers peuvent toujours l'être dans la mesure où dans les cas complexes la subjectivité du lexicographe doit intervenir – mais de baliser la totalité de la structure textuelle des articles. En d'autres termes,

[1] Nous opposons ici la notion de structure "molle" des articles de notre corpus, à la structure "rigide" des dictionnaires modernes, structure au sein de laquelle règne un formatage informationnel important.

dans une telle entreprise, pour baliser utilement, il faut savoir renoncer à faire rentrer la totalité de l'information dans des boîtes cloisonnées et bien distinctes.

Le balisage Souple apporte donc selon nous une réponse aux difficultés que rencontre généralement le lexicographe qui désire un balisage riche de son corpus.

Ce balisage repose sur un repérage large des champs informationnels de l'article dictionnairique et abandonne le marquage de certains champs présents dans le balisage Analytique.

En accordant une place majeure au découpage de l'article en deux grands blocs, à savoir le bloc <ENTREE></ENTREE> et le bloc <CORPS></CORPS>, notre balisage admet le fait que certaines portions de texte ne soient pas formellement identifiées comme relevant de tel ou tel champ informationnel.

Les principaux champs identifiés sont les entrées principales et les sous-entrées, la forme faisant l'objet de l'article[2], l'information grammaticale, l'information étymologique, la marque de domaine, la définition, l'exemplification et le discours encyclopédique. Il s'agit là de la plupart des champs également identifiés par Wionet et Tutin, mais nous abandonnons toutefois le marquage de champs comme les locutions ou les collocations, les indications rhétorico-sémantiques du type "figurément, par extension", et le "sous-article locutif", traitant de collocations, introduit par des marqueurs du type "On dit proverbialement d'[…]", "On dit figurément d' […]". Ces champs peuvent tout à fait être pris en compte par notre balisage mais nous avons choisi de ne pas les mettre en évidence pour conserver un repérage informationnel large et donc un balisage moins lourd.

La figure ci-dessous nous donne une illustration d'un article balisé grâce au balisage *Souple* ou *Flottant* :

```
<ARTICLE><ENTREE TYPE="EP"><FORME>(N. )
ALPHA</FORME>.<INFORMATION_GRAMMATICALE TYPE="SUBSTANTIF
MASCULIN"><PARTIE_DU_DISCOURS TYPE="SUBSTANTIF">s.
</PARTIE_DU_DISCOURS><GENRE
TYPE="MASCULIN">m.</GENRE></INFORMATION_GRAMMATICALE></ENTREE><CORPS
><DEFINITION>C'est le nom <LANGUE TYPE="GREC">Αλφα </LANGUE>de la première lettre
des grecs.</DEFINITION> <DISCOURS_ENCYCLOPEDIQUE>Ils ont eux-mêmes emprunté ce nom
des hébreux ou des phéniciens, en prenant d'eux les caractères littéraux. <REFERENCE
TYPE="PERSONNE">Eusèbe</REFERENCE> (<REFERENCE TYPE="OUVRAGE">Præp. evang.
X. 6.</REFERENCE> ), en fait la remarque, & le prouve par un raisonnement bien simple: <LANGUE
TYPE="LATIN">Id ex grœcd singulorum elementorum appellatione quivis intelligit : quid enim Aleph
ab Alpha magnopere differt ? quid autem vel Beta à Beth, vel à Gamma Gimel, aut Delta à Delt, aut He
ab E, aut Zaïn à Zeta, ceteraque deinceps his similia ? </LANGUE>Une observation qui confirme cette
origine, c'est que le mot <LANGUE TYPE="GREC">Αλφα </LANGUE>, chez les grecs, est
simplement le nom de leur première lettre comme première lettre ; qu'en conséquence il est dans cette
langue un radical primitif, d'où l'on a dérivé <LANGUE TYPE="GREC">αλφάνω,
αλφέω</LANGUE>ou <LANGUE TYPE="GREC">αλφω</LANGUE> (je trouve, j'invente le premier
& au même rang que tient <LANGUE TYPE="GREC">αλφω</LANGUE> parmi les lettres),
<LANGUE TYPE="GREC">αλφηστής </LANGUE> (inventeur, premier auteur) : au lieu que le nom
hébreu de la premiere lettre hébraïque vient du verbe <LANGUE TYPE="HEBREU">
הרא</LANGUE><LANGUE TYPE="LATIN">(alaph) </LANGUE>apprendre, enseigner, mot qui
signifie aussi enseignement, doctrine, & par extension prince ? chef, parce que le prince & le chef doit
conduire le peuple & lui enseigner les bonnes lois ; de là vient que les hébreux ont nommé de même leur
première lettre, pour indiquer qu'elle est à la tête des autres, qu'elle en est le
chef.</DISCOURS_ENCYCLOPEDIQUE> <SIGNATURE TYPE="BEAUZEE"> (M.BEAUZÉE.)
</SIGNATURE></CORPS></ARTICLE>
```

Figure. 1 : Balisage *Souple* ou *Flottant* de l'article ALPHA.

[2] Nous évoquons ici le terme de "forme" car celui-ci permet de résoudre le problème lié au fait que les mots faisant l'objet de l'article ne sont pas toujours des lemmes.

L'exemple proposé montre qu'à l'intérieur de la structuration en blocs, <ENTREE></ENTREE> et <CORPS></CORPS>, les différents champs informationnels identifiés et balisés flottent à l'intérieur d'une structure textuelle qui elle ne l'est pas. Les éléments <LANGUE> et <REFERENCE> flottent ainsi à l'intérieur des données textuelles présentes dans le jeu de balises <DISCOURS_ENCYCLOPEDIQUE> </DISCOURS_ENCYCLOPEDIQUE>.

Cette souplesse dans le balisage n'oblige pas la personne qui pose les jalons à faire correspondre chaque portion de texte de l'article à un champ informationnel particulier. L'exemple suivant est encore plus significatif:

<ARTICLE>
<STATUT TYPE="DIFFERENT"/><ENTREE
TYPE="EP"><FORME>LONGUE</FORME>,
<INFORMATION_GRAMMATICALE TYPE="ADJECTIF
FEMININ"><PARTIE_DU_DISCOURS
TYPE="ADJECTIF">adj.</PARTIE_DU_DISCOURS><GENRE
TYPE="FEMININ">f.</GENRE></INFORMATION_GRAMMATICALE></ENT
REE><CORPS>en terme de <DOMAINE
TYPE="GRAMMAIRE">Grammaire</DOMAINE>.<DEFINITION>On appelle
Longue une syllabe relativement à une autre, que l'on appelle brève, & dont la durée
est de moitié plus courte.</DEFINITION> **Voyez <REFERENCE
TYPE="VEDETTE">BRÈVE</REFERENCE>.**
<DISCOURS_ENCYCLOPEDIQUE>La Longueur & la brieveté n'appartiennent
jamais qu'à la voyelle, ou plus tôt à la voix qui est l'âme de la syllabe; les
articulations sont essentiellement instantanées & indivisibles. On met un trait droit
couché au dessus d'une voyelle, pour marquer qu'elle est longue, comme on y met un
c couché, pour marquer qu'elle est brève. Ainsi, on écriroit tempora, pour marquer
que la première syllabe est longue, & les deux dernières
brèves.</DISCOURS_ENCYCLOPEDIQUE> <SIGNATURE
TYPE="BEAUZEE">(M.BEAUZÉE.)</SIGNATURE></CORPS></ARTICLE>

Figure. 2 : Balisage *Souple* ou *Flottant* de l'article LONGUE.

La portion de texte "BRÈVE" a été balisée comme constituant une référence à un autre article, mais l'expression "Voyez BRÈVE" n'appartient à aucun des grands champs informationnels identifiés dans la grammaire de notre document[3]. Elle flotte donc au sein du jeu de balises <CORPS></CORPS>.

2.2 Dissociation du balisage logique et du balisage physique

L'une des autres caractéristiques de notre balisage réside dans le fait que, dans la lignée du *balisage analytique* de Chantal Wionet et Agnès Tutin qui s'appuyait sur une norme de codage, le langage SGML (Standard Generalised Markup Language), ce dernier s'appuie sur le langage XML (eXtensible Markup Language), langage faisant également office de norme de codage internationale.

Ainsi que le prône la philosophie du langage XML, nous avons clairement dissocié le balisage logique des articles, qui consiste à identifier les différents champs informationnels de la microstructure dictionnairique, du balisage physique de ces derniers, qui lui consiste à

[3] Connue sous le nom de DTD (Définition de Type de Document), une grammaire de document au sens où nous l'entendons fournit la liste des balises et les règles d'agencement de ces dernières utilisées lors d'une entreprise de balisage.

décrire tous les aspects de mise en forme des données rétroconverties. En ce sens, notre entreprise diverge encore de celle de Chantal Wionet et Agnès Tutin qui dans leur *balisage analytique* ne dissocient pas de manière stricte la forme et le contenu des articles.

La dissociation des deux niveaux de codage a été particulièrement aisée pour notre entreprise puisqu'à la différence de ce qu'on peut constater pour les dictionnaires les plus anciens, les articles de tout le dictionnaire *Grammaire & Littérature* de l'*Encyclopédie Méthodique*, marquent une certaine régularité typographique pour certains champs.

Ainsi, la forme faisant l'objet de l'article apparaît toujours en lettres capitales majuscules lorsqu'il s'agit d'une entrée principale. Les sous-entrées sont également en lettres capitales majuscules mais avec une casse plus petite. Notons toutefois que certaines portions de texte présentes dans le <CORPS> de l'article peuvent apparaître en italiques. Dans la perspective de conserver la dissociation entre le balisage logique et le balisage physique, nous avons choisi de ne pas tenir compte de ce fait puisque ce phénomène ne justifie pas à lui tout seul de mêler les deux types de balisage.

Toutes les régularités typographiques notables ne sont pas pour autant occultées puisque nous avons la possibilité de les spécifier au sein d'une feuille de style qu'on associe au document XML.

Par sa flexibilité qui ne force pas le document rétronconverti à répondre à un patron organisationnel au sein duquel chaque portion de texte correspond à un champ informationnel particulier, le *balisage souple* ou *flottant* constitue selon nous une solution intéressante pour l'informatisation d'ouvrages de nature encyclopédique. Bien plus qu'une alternative au *balisage minimal* et au *balisage analytique*, ce dernier pourrait peut-être même constituer une chance réelle de pouvoir informatiser des ouvrages dont la complexité organisationnelle pourrait rebuter les initiatives les plus courageuses.

Bibliographie

Sources primaires

DIDEROT, Denis, ALEMBERT, Jean Le Rond d'., 1751-1766, *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de Lettres*, Stuttgart, F. Frommann Verlag – G. Holzboog, 1990.

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 Editeurs de l'Encyclopédie, 1782-1832, A Paris (chez Panckoucke), Liège (chez Plomteux). 210 vol.

Sources secondaires

CARON, Philippe, DAGENAIS, Louise, GONFROY, Gérard (1992). "Le programme d'informatisation du *Dictionaire critique de la langue française* de l'abbé Jean-François

Féraud (1787)", *Historical Dictionary Databases* (éd. T.R. Wooldridge). *CCH Working Papers*, 2: 87-103.

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

DOUAY, Françoise. 1996, "Le paradoxe et son cortège, de l'Encyclopédie à l'Encyclopédie méthodique", R. Landheer et P. J. Smith (eds) *Le paradoxe en linguistique et en littérature*, Genève, Droz, 221-237.

DOUAY, Françoise, 1994, "Les figures de rhétorique: actualité, reconstruction, remploi", in *Langue française* n°101: *Les Figures de rhétorique et leur actualité en linguistique*, Paris, éditions Larousse, pp. 13-26.

EHRARD, Jean, 1991, "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.

TEYSSEIRE, Daniel, 1992**,** "Les idéologues et l'idéologie dans l'*Encyclopédie Méthodique* Premier inventaire", *Europäische Sprachwissenschaft um 1800, Methodologische und historiographische Beiträge zum Umkreis der "idéologie"*. Band 3, Nodus Publikationen, 165-179.

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

REY, Christophe, 2004, *Analyse et informatisation des articles traitant de l'étude des sons dans le dictionnaire Grammaire & Littérature de Nicolas Beauzée et Jean-François Marmontel, issu de l'Encyclopédie Méthodique.* Thèse de doctorat. Aix-en-Provence.

WOOLDRIDGE, Terence Russon, 1994, "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.