

Recherches sur Diderot et sur l'Encyclopédie

Numéro 31-32 (Avril 2002)

L'Encyclopédie en ses nouveaux atours électroniques: vices et vertus du virtuel

Robert Morrissey

L'Encyclopédie électronique.

Avertissement

Le contenu de ce site relève de la législation française sur la propriété intellectuelle et est la propriété exclusive de l'éditeur.

Les œuvres figurant sur ce site peuvent être consultées et reproduites sur un support papier ou numérique sous réserve qu'elles soient strictement réservées à un usage soit personnel, soit scientifique ou pédagogique excluant toute exploitation commerciale. La reproduction devra obligatoirement mentionner l'éditeur, le nom de la revue, l'auteur et la référence du document.

Toute autre reproduction est interdite sauf accord préalable de l'éditeur, en dehors des cas prévus par la législation en vigueur en France.

revues.org

Revues.org est un portail de revues en sciences humaines et sociales développé par le Cléo, Centre pour l'édition électronique ouverte (CNRS, EHESS, UP, UAPV).

Référence électronique

Robert Morrissey, « L'Encyclopédie électronique. », *Recherches sur Diderot et sur l'Encyclopédie* [En ligne], 31-32 | Avril 2002, mis en ligne le 16 mars 2008. URL : <http://rde.revues.org/index3273.html>
DOI : en cours d'attribution

Éditeur : Société Diderot

<http://rde.revues.org>

<http://www.revues.org>

Document accessible en ligne sur : <http://rde.revues.org/index3273.html>

Ce document est le fac-similé de l'édition papier.

Tous droits réservés

Robert MORRISSEY

L'Encyclopédie électronique

Lors d'un colloque tenu à l'École des Hautes Études en Sciences Sociales en juin 1998 pour présenter pour la première fois au public des chercheurs l'*Encyclopédie électronique* de l'ARTFL¹, j'ai conclu mes remarques en citant un passage du célèbre article *ENCYCLOPÉDIE*, où Diderot s'exprime sur les difficultés et les limites de son entreprise monumentale :

Nous avons vû, à mesure que nous travaillions, la matière s'étendre, la nomenclature s'obscurcir, des substances ramenées sous une multitude de noms différens, les instrumens, les machines & les manœuvres se multiplier sans mesure, & les détours nombreux d'un labyrinthe inextricable se compliquer de plus en plus. Nous avons vû combien il en coûtoit pour s'assurer que les mêmes choses étoient les mêmes, & combien, pour s'assurer que d'autres qui paroissent très-différentes, n'étoient pas différentes. [...] Mais nous avons vû que de toutes les difficultés, une des plus considérables, c'étoit de le produire une fois, quelqu'informe qu'il fût, & qu'on ne nous raviroit pas l'honneur d'avoir surmonté cet obstacle. (t. V, p. 644).

Depuis la première installation de la base de données, nous avons eu quelques 80 000 interrogations de l'*Encyclopédie électronique* et nous commençons à avoir une idée des utilisations que l'on peut en faire et des limitations de cette première implantation². Notre projet initial prévoyait une première installation qui permettrait aux chercheurs de travailler avec l'*Encyclopédie électronique* sous sa première forme. Leurs commentaires et observations devaient nous aider à mieux comprendre les lacunes et les insuffisances de cette première version et à préparer une stratégie de correction et de réimplémentation de la base.

1. « *L'Encyclopédie. Du réseau au livre et du livre au réseau* » co-organisé par Philippe Roger et moi-même. Les actes du colloque ont été publiés sous ce même titre, Paris, Champion, 2001.

2. Le chiffre de 60 000 couvre la période entre juin 1999 et mai 2001. Ce chiffre ne tient pas compte des interrogations de l'installation de l'INaLF à Nancy.

Aujourd'hui, je voudrais évoquer rapidement quelques-uns des problèmes auxquels nous sommes confrontés ainsi que nos projets pour améliorer l'état des données.

La saisie

Il n'est peut-être pas inutile de rappeler certains éléments concernant la saisie et les dimensions du corpus. A partir d'un exemplaire microfiché de la première édition de Paris, nous avons effectué une première saisie que nous avons soumise à un ensemble d'opérations de repérage automatique. Pour distinguer toute une gamme d'« objets » tels que les articles, les classifications, les renvois, etc., nous avons élaboré ce que l'on pourrait appeler une sorte de « grammaire » des articles, basée surtout sur des critères de positionnement, de typographie, et de ponctuation. Ainsi le travail d'identification de ces objets a été effectué automatiquement et sans vérification manuelle par un logiciel incorporant cette grammaire.

Quelques chiffres aident à donner une idée de l'envergure du projet. *L'Encyclopédie* contient environ 20,5 millions d'occurrences de mots et 391 000 mots différents. Sur l'ensemble des dix-sept tomes d'articles et des onze tomes de planches, nos logiciels ont identifié quelque 75 590 parties distinctes du texte (principalement les articles et les légendes). Nous avons classé cet ensemble en 44 981 articles principaux, 28 366 sous-articles et 2 576 légendes. En plus de ces catégories, il reste 667 « objets » divers tels que les pages de titre, le « Discours préliminaire », les préfaces, les avertissements, les sous-ensembles de planches, etc. Il y a approximativement 64 000 renvois entre articles (dont un nombre non-négligeable à des articles inexistantes) et 12 000 renvois de planches.

Nous estimons avoir pu identifier par procédés automatiques autour de 99 % des articles et sous-articles, un peu moins en ce qui concerne les indications d'auteurs, de classification et de fonction grammaticale. A cause de la diversité des formes de renvois, le niveau d'identification est légèrement inférieur, de l'ordre d'entre 90 à 95 %, et pour les planches, il est encore quelque peu plus bas. Pour le moment, nous n'avons toujours rien ajouté au texte. Ainsi si (ce qui est souvent le cas) les encyclopédistes ont négligé d'indiquer la fonction grammaticale d'un terme (par exemple, « verbe »), l'ordinateur ne pourra pas l'identifier par sa fonction grammaticale.

Sachant qu'il y aurait des erreurs de saisie et d'identification, nous avons décidé dès le début d'incorporer dans la base une image numérisée de chaque page. La transcription du texte permet aux utilisateurs d'effectuer des recherches multiples et de balayer le texte de plusieurs manières, mais il fallait aussi l'image numérisée — c'est-à-dire un facsimilé de chaque page — parce que nous savions que la version « dactylo-

graphiée » comporterait certaines lacunes et insuffisances. Une fois cette décision prise, nous nous sommes heurtés à un problème de qualité des images : celle des images tirées de l'édition microfichée (édition de Paris) était insuffisante pour pouvoir les exploiter. Nous avons pris le parti de trouver une meilleure source d'images et nous nous sommes tournés vers la micro-édition en fac-similé de Pergamon Press³. Nous avons procédé à une comparaison détaillée des deux éditions et nous avons conclu que le fac-similé en micro-édition était aussi une édition de Paris. Malgré les très légères différences entre l'édition sur microfiche à partir de laquelle nous réalisons la saisie et la micro-édition de Pergamon Press à laquelle nous avons recours pour les images, nous avons cru que nos utilisateurs seraient mieux servis par la qualité visuelle largement supérieure des images. Nous avons fourni sur notre site-web une description détaillée de nos critères de choix et de nos conclusions⁴.

La décision d'offrir l'image des pages nous a permis de faire l'économie de la saisie de certains éléments non-textuels tels que les tables, les partitions musicales et d'autres images qui émaillent le texte des articles. Pour consulter ces éléments, les utilisateurs n'ont qu'à remonter de la version textuelle à celle des images.

Les difficultés et les erreurs

Il va sans dire qu'étant donné la nature automatique de tout le travail d'identification et de correction, la base comporte des erreurs. On pourrait classer les problèmes ainsi : les problèmes de saisie au niveau du texte ; les problèmes de saisie et de repérage dans les éléments identificateurs ; les problèmes d'affichage ou de représentation des données sur l'écran ; et les problèmes avec les liens entre texte et images en fac-similé. Sans prétendre à une analyse complète de ces problèmes, je voudrais aujourd'hui en évoquer des éléments importants avant d'esquisser très brièvement notre plan pour améliorer la qualité de la base.

3. New York, 1969.

4. Pour l'ensemble des documents sur les limitations et défauts de l'*Encyclopédie* électronique figurant sur notre site, les utilisateurs devraient consulter sous la rubrique « More about the Encyclopédie Project » [<http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/more.html>], la partie intitulée « Caveat : Functionality and Limitations » [<http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/caveat.html>]. Là, entre autres choses, on trouvera la comparaison des deux éditions : « general comparison » [<http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/ENCYC.editions.html>] ainsi qu'une discussion des éléments que nous n'avons pas pu identifier : « limitations of the identification procedures » [<http://www.lib.uchicago.edu/efts/ARTFL/projects/encyc/caveat.html#identifications>].

1. Les problèmes de saisie et de repérage au niveau du texte

Il y a deux classes d'erreurs de saisie : les erreurs balisées lors de la saisie et celles qui n'ont pas été balisées. Lors de la saisie, quand les opérateurs n'arrivaient pas à bien distinguer une lettre, ils indiquaient leur incertitude par un signe [*< ? >*]. Comme ils travaillaient à partir de photocopies tirées de microfiches, beaucoup des hésitations étaient dues à la mauvaise qualité d'une image sur microfiche. En fait, il s'agit plus de caractères non-reconnus que d'erreurs à proprement parler. Parmi les pages avec plus de vingt erreurs, on trouve :

60 : v. 3 p. 859
 50 : v. 2 p. 660
 38 : v. 3 p. 197
 33 : v. 17 p. 215
 32 : v. 3 p. 335
 30 : v. 2 p. 649
 27 : v. 12 p. 243
 26 : v. 12 p. 880
 24 : v. 3 p. 196
 22 : v. 7 p. 265
 22 : v. 3 p. 201
 21 : v. 3 p. 799
 21 : v. 3 p. 591
 20 : v. 3 p. 92
 20 : v. 3 p. 195
 20 : v. 17 p. 119

En tout, sur à peu près 18 000 pages de texte saisies, nous avons relevé 83 pages ayant plus de 10 erreurs balisées ; 387 pages comportant de 5 à 10 erreurs balisées ; 1 605 pages avec entre 2 et 4 erreurs et 2 303 avec une erreur balisée.

En plus des erreurs balisées, il y en avait beaucoup qui n'ont pas été identifiées par les opérateurs. Parmi les plus fréquentes, on constatait des confusions récurrentes (parfois balisées, parfois non), par exemple entre 's' allongé [*f*] et 'f' ; entre 'c' et 'e' ; entre 't' et 'r' ou 'i'. Parfois les opérateurs ont inséré un blanc ou milieu d'un mot. Cette table donne une idée de l'ampleur du problème des fautes de frappe :

<i>mot</i>	<i>erreur</i>	<i>fréq.</i>	<i>correct</i>	<i>fréq.</i>	<i>taux d'erreur</i>
ainsi	ainfi	306	ainsi :	26 485	1 %
difficile	dissicile :	114	difficile :	5 835	1.9 %
enfants	ensans :	97	enfants :	4 758	2 %
enfin	ensin :	504	enfin :	9 500	5 %
afin	asin :	306	afin :	5 189	5.6 %

Ainsi le taux d'erreur reste relativement faible. Dans les cas où il n'y avait aucune ambiguïté possible, comme ceux indiqués ci-dessus, nous avons pu corriger automatiquement. Mais dès qu'il y avait possibilité d'ambiguïté (par exemple, confusion entre « tendre » et « rendre »), nous n'avons pas pu le faire automatiquement et ces erreurs restent dans la base telle qu'elle est actuellement. Il faut aussi signaler que, pour des raisons esthétiques, nous avons pris le parti de masquer les signes d'erreur [`< ? >`] mis par les opérateurs. En revanche, ils sont conservés dans la base et pourront être utilisés au moment d'entreprendre un travail systématique de correction.

Il y a une autre catégorie d'erreurs qui proviennent non pas de la saisie, mais de notre manière de l'analyser, c'est-à-dire d'une faiblesse dans la « grammaire » que nous avons élaborée pour traiter les données. Deux cas sont particulièrement flagrants. Pierre Lafon en a souligné les inconvénients dans sa communication. Quand il a fallu fixer les paramètres permettant à nos logiciels de distinguer un mot suivant « qu' » comme un mot distinct, dans un cas particulier où le mot suivant commence par un « h », nous avons introduit une erreur typographique qui supprime la lettre « h ». Ainsi dans la phrase tirée de l'article CHANT, on lit dans la version électronique : « C'est donc au propre qu'il faut prendre ce qu'omere, Hésiode, &c. ont dit au commencement de leurs poèmes. » Dans le cas d'Homère, on constate 92 occurrences de cette erreur sur 775 occurrences d'Homère, soit un taux d'erreur de 12 %. Dans le cas de la configuration « qu'homme(s) », on trouve 35 occurrences du « h » tronqué sur 21 127 occurrences du mot, soit 1,65 %. Je tiens à souligner que cette erreur typographique se trouve dans notre logiciel de repérage et non pas dans les données elles-mêmes. La bonne nouvelle est que cette information existe dans nos fichiers ; la mauvaise est qu'il faut attendre une réinstallation avant de pouvoir repêcher ces données.

Le deuxième cas concerne le trait d'union. Il est particulièrement important parce qu'il constitue la source des erreurs les plus répandues dans l'*Encyclopédie* électronique. Dans la saisie originelle, il y a des mots à traits d'union et des mots coupés pour des raisons typographiques à la fin d'une ligne, d'une colonne ou d'une page. La difficulté surgit quand il y a un mot à trait d'union qui se trouve à cheval sur deux lignes. Dans ce cas, il faut savoir distinguer entre les vrais mots à trait d'union et ceux qui ont été divisés artificiellement. Lors de l'installation de la version actuelle de la base, nous avons procédé à un travail de désambiguïsation visant à ressouder les mots coupés pour des raisons typographiques. Mais, comme nous voulions respecter absolument la pagination et les divisions en colonnes, nous avons établi une distinction entre trois types de traits d'union : celui que l'on trouve dans des mots à trait d'union et que nous représentons dans la base par un double trait [--], celui qui divise les mots se trouvant à la fin

d'une colonne [`<cb->`] (cb= column break), et celui qui coupe les mots à la fin d'une page [`<pb->`] (pb= page break).

Par rapport aux traits d'union, les problèmes qui se manifestent dans l'installation actuelle de la base sont multiples. Le logiciel de désambiguïsation n'étant pas d'une très grande finesse, il y a des mots qui ne devraient pas être coupés qui le sont restés et, parfois, des mots à traits d'union qui ont été abusivement rejoints. Pour faciliter le travail de correction ultérieure, nous avons distingué entre les traits d'unions « réels » identifiés avec une très grande certitude et affichés à l'écran par un double trait [--], par exemple « gardes--notes », et les traits d'union que notre désambiguïsation n'arrivait pas à résoudre avec certitude ; ceux-ci sont affichés dans l'installation actuelle ainsi : [`<->`] (1 545 cas dans la base). Dans la version actuelle, tous les mots divisés par un trait d'union ont été indexés comme deux mots différents. Ainsi « gardes--notes » est indexé en deux mots : « gardes » et « notes ». Mais cela veut dire qu'un mot comme « adresses » coupé en « adres `<pb->` ses » ou en « adres `<->` ses », a été aussi indexé en deux mots : « adres » et « ses ». Dans la base, il y a 3 716 traits d'unions arrivant en fin de page, 3 755 en fin de colonne. La mauvaise nouvelle est que pour corriger ces erreurs, il faudrait désambiguïser chaque cas, et pour cela il faudra attendre une réinstallation des données et un meilleur logiciel de désambiguïsation (ou bien un travail manuel). La bonne nouvelle est que nous avons conservé l'information originelle et qu'une réinstallation est tout à fait possible.

2. Les problèmes de saisie et de repérage dans les éléments identificateurs

A cette catégorie appartient un ensemble de références telles que titres d'articles, classes de connaissance, renvois, noms ou codes d'auteurs. En ce qui concerne les titres des articles et des sous-articles, nous pensons en avoir repéré à peu près 99 %. En revanche, dans les domaines des classes de connaissance et des renvois, nous croyons le taux d'erreur plus élevé, de 5 à 10 %. Quant à l'identification des noms d'auteurs indiqués dans l'*Encyclopédie* même (souvent par une lettre, H pour Toussaint, I pour Daubenton, etc.), pour mieux comprendre l'ampleur du problème, nous avons pris un cas, celui de Diderot, qui signait, comme on le sait, avec un astérisque. Nous l'avons choisi parce que les travaux exhaustifs de J. Proust, R. Schwab, T. Rex et J. Lough nous permettaient d'avoir une très grande certitude sur les articles dont Diderot a été l'auteur⁵. Dans le tome V des *Œuvres complètes*, J. Lough et J. Proust fournissent une liste de

5. L'Inventaire de l'*Encyclopédie* élaboré par Richard Schwab, Walter Rex et John Lough paru dans *Studies on Voltaire and the Eighteenth Century*, v. 80, 83, 85, 91, 92, 93, 223, 1971-1984 ; *Œuvres complètes*, édition critique et annotée, publiée sous la direction de Herbert Dieckmann, Jean Fabre, Jacques Proust, Jean Varloot, Paris, Hermann, 1975—

5 633 articles⁶. Nos logiciels ont repéré 5 634 articles que Diderot aurait écrits. Mais en comparant les titres d'articles sur les deux listes, nous avons trouvé 37 articles attribués à Diderot dans l'*Encyclopédie* électronique qui ne figurent pas sur la liste de J. Lough et J. Proust et, inversement, 47 articles sur la liste de Proust qui ne figurent pas parmi les articles que l'*Encyclopédie* électronique indique comme étant de Diderot. Ainsi, il y a 84 articles signés par Diderot, soit 1,5 % du total, qui n'ont pas été correctement identifiés par nos logiciels [Voir Appendice A].

Une partie de ces erreurs provient d'une confusion que fait notre logiciel de repérage entre les astérisques indiquant les notes — il y a 288 notes ainsi marquées dans l'*Encyclopédie* — et quelques-uns des astérisques indiquant que Diderot est l'auteur de l'article. Dans ce cas, la saisie est bonne, mais nous avons mis un mauvais paramètre dans notre « grammaire » des articles. Il sera facile de corriger le tir lors d'une réinstallation de la base.

Mais il y a une autre question, plus épineuse, celle-ci, qui concerne les attributions établies par des chercheurs. Dans le cas de Diderot, l'édition de J. Proust fournit une liste de 100 « articles non-signés qu'on peut attribuer avec certitude à Diderot » mais qui ne sont pas marqués d'un astérisque, et une deuxième liste de 553 « articles non signés qu'on pourrait peut-être attribuer à Diderot⁷ ». Il me semble que l'on devrait avoir pour but la possibilité d'intégrer les résultats de la recherche des 30 dernières années. Dans le cas de Diderot, nous pouvons afficher ces listes sur nos sites. Mais je pense que notre première priorité devrait être d'arriver à un texte plus fidèle au texte originel, ce qui implique un travail de correction de grande envergure.

3. Les problèmes d'affichage ou de représentation des données sur l'écran

Notre problème principal dans la représentation des données concerne le grec (mais aussi les autres alphabets non-latin, l'hébreu, par exemple). Nous avons saisi le grec d'une manière systématique, mais nous n'avons pas trouvé de convention permettant de représenter le grec d'une manière satisfaisante. Ce travail reste à faire.

4. Les problèmes des liens entre texte et images en fac-similé

Dans un certain nombre de cas, quand on essaie de remonter du texte à l'image en fac-similé d'une page, le logiciel répond qu'il ne trouve pas l'image. Ce problème provient du fait que parfois la pagination n'est pas en

6. *Œuvres complètes*, t. V, p. 133-206.

7. *Œuvres complètes*, p. 207-220.

chiffres arabes mais en chiffres romains et notre logiciel ne sait pas établir le lien ; c'est le cas, par exemple, du « Discours préliminaire ». Lors d'une réinstallation de la base, ce problème pourra être corrigé.

Et maintenant ?

Dans l'ensemble, nous sommes dans les normes que nous avons annoncées au moment du lancement de l'*Encyclopédie électronique*. Nous avançons dans la connaissance et la documentation des erreurs et des limites de l'installation actuelle de la base⁸. Nous sommes sur le point d'entreprendre le travail d'amélioration et, comme je l'ai indiqué plus haut, nous pensons que la meilleure stratégie consiste à commencer par mettre le texte dans un état aussi exact, aussi correct que possible par rapport à l'original. Pour cela, nous avons décidé de nous associer avec une maison d'édition, les Éditions Honoré Champion, qui s'est engagée à nous rendre un corpus corrigé qui ne comporterait pas plus d'une erreur tous les 15 000 signes. L'*Encyclopédie électronique* corrigée sera mise à la disposition des chercheurs sur l'internet, mais elle sera aussi accessible sous forme de CD-ROM produit par les éditions Champion. Entre-temps nous essayons d'améliorer l'information sur les erreurs et de les afficher sur notre site-web.

Robert MORRISSEY
Université de Chicago
ARTFL Projet
 (USA)

Annexes

I. Liste des articles qui ne figurent pas sur la liste de J. Proust mais que l'*Encyclopédie électronique* a attribués par erreur à Diderot :

ABDICATION
 ABRAXAS
 ABRÉVIATION
 ACHEENNE
 ALTERATION
 ANECDOTES

8. Nous sommes particulièrement reconnaissants à Marie Leca-Tsiomis et Martine Groult de leurs observations non seulement sur l'état actuel de l'*Encyclopédie électronique*, mais aussi sur les manières de l'améliorer.

ANTIPARASTASE
ARCANUM DUPLICATUM
ARÉOPAGE
ARGO
BOULE
BOUSSOLE
BOUTS
FERMIERS
FESTINS ROYAUX
FÊTE
FÊTES DES GRANDES VILLES DU ROYAUME
FORMULE
FROTTEMENT
GARDES-FRANÇOISES
GÉNÉRIQUE
GÉNIE
GRÊLE
HÉBRAÏQUE
HORLOGERIE
LAIT
LANGUE
MINIATURE
MOUCHES LUISANTES
PRYTANE
VÉNERIE
TRANSFUGE
VINGTIÈME, imposition.

II. Liste des articles qui figurent sur la liste de J. Proust, mais que l'*Encyclopédie* électronique n'attribue pas à Diderot :

ABREGÉ
ABRICOTS
ACATHABOLE
ACCÈS
ACIÉRIE
ADONIS
AGRICULTURE
AIGLE
ALAMBIC
AMMONIAQUE
AMMONITES
AMNISIADÉS
AMOGABARE
AMOL
AMOME
AMORCE
ANCRE/ENCRE

ANCRÉ
ANTÉDILUVIENNE
ARCANUM JOVIS
BARDANE
BESOIN
CAILLE
CASTOR ET POLLUX
CHABOT
CHANTEPLEURE (2)
CHARBON DE BOIX
CHAT (2)
COELIUS
COLONNE ANTOMINE
COMTES DE LYON
COUP, PETITS COUPS
COUP
CROUTE DE GARENCE
D
DAUPHIN
ÉCHAPPEMENT
FER-BLANC
FONTAINE
FORCES (4)
FOUET
FOURMI