DHQ: Digital Humanities Quarterly
2018
Volume 12 Number 1

Le texte numérique : enjeux herméneutiques [en]

Jean Guy Meunier <meunier_dot_jean-guy_at_uqam_dot_ca>, Université du Québec à Montréal

Abstract

La numérisation des textes est omniprésente dans les humanités numériques. Elle semble se présenter uniquement comme une modification du support matériel : du texte sur papier au texte numérique. Mais elle fait plus que cela. La numérisation touche aussi le texte en tant qu'objet sémiotique. Or, les multiples opérations de cette technologie mettent en œuvre des décisions interprétatives qui ne sont pas sans affecter le texte sémiotique, c'est-à-dire celui qui se donne à lire et à analyser. En ce sens, la numérisation des textes n'est pas neutre. Elle est un moment important d'une herméneutique matérielle.

1. Introduction

La numérisation des documents est une des nombreuses technologies informatiques qui transforment la culture et la science. Tous les supports matériels classiques de l'image au film, de la musique à la sculpture peuvent être transformés et déposés sur un support dit numérique. Mais plus que tout autre, le texte, dans son format traditionnel – manuscrit, codex, livre, revue, rapports, magazine, etc. –, est, depuis une vingtaine d'années, soumis à une numérisation massive. Et à cette masse de textes numérisés s'est ajoutée celle issue de la production directe de textes via des logiciels de traitement de textes. De nos jours, peu de textes échappent ainsi à la technologie de la numérisation[1]. 1

Si certains projets de ce « virage numérique » sont modestes, qu'ils soient créés ou gérés par des groupes de recherche, des institutions académiques, ou par des initiatives sociales ou commerciales locales, d'autres sont au contraire ambitieuses et de dimensions titanesques. Reprenant et prolongeant d'une certaine manière l'objectif premier et originel de la bibliothèque d'Alexandrie, dont le patrimoine textuel comptait plus de 500 000 rouleaux manuscrits, ces projets aspirent à la construction d'une bibliothèque numérique universelle, rendant accessible à tous, en tout temps et partout, une grande partie du patrimoine textuel[2]. 2

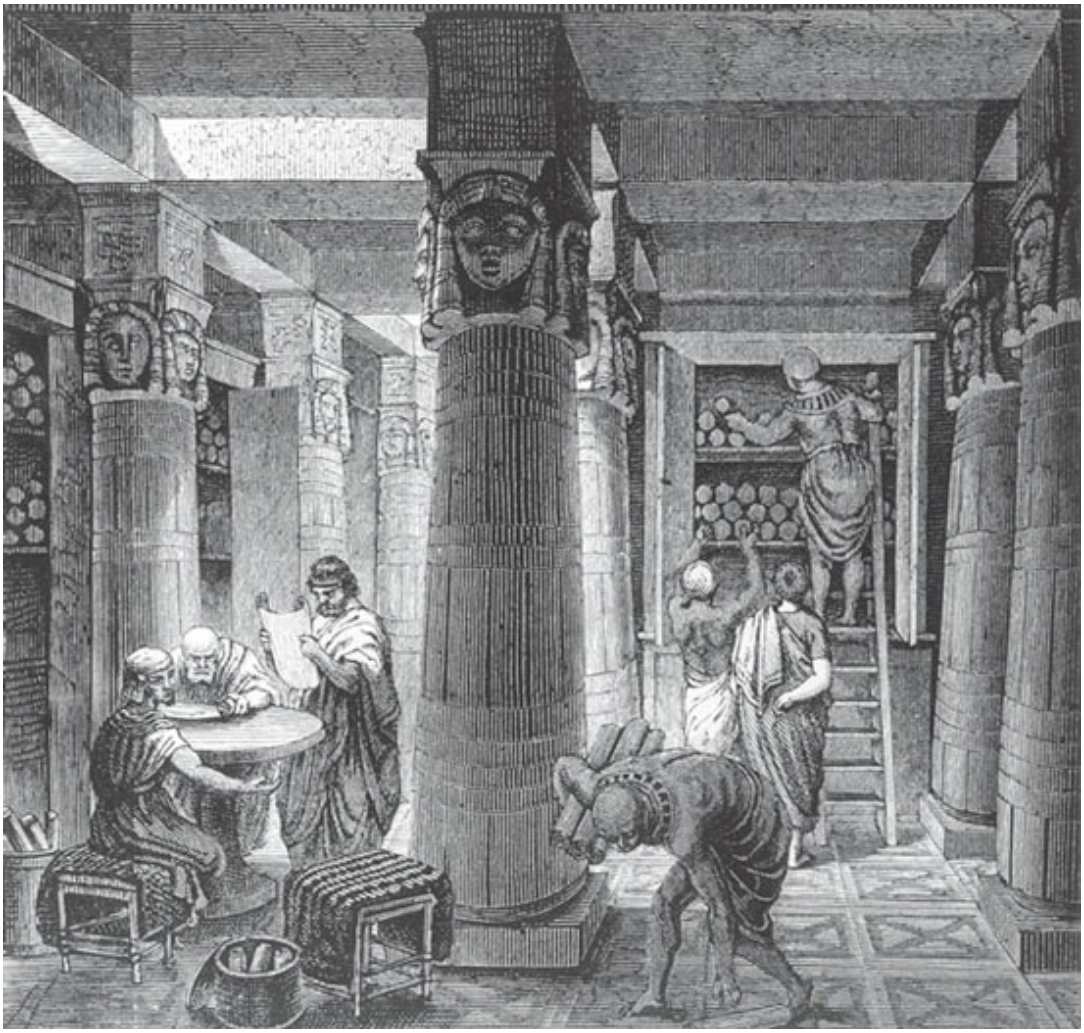

Figure 1. Évocation de la librairie d'Alexandrie, par O. von Corben (Tolzmann, Hessel et Peiss, 2001)

Grâce à de tels projets ainsi qu'aux nouvelles possibilités offertes par les technologies informatiques actuelles, tant d'archivage, de traitement que de télécommunications, on espère que des villages reculés auront désormais la possibilité de bonifier leur mince bibliothèque locale de plusieurs millions de copies de livres numériques. 3

Cette numérisation massive de documents textuels, par l'importance de son impact présent et futur sur la culture et la science, change radicalement le rapport au savoir et à sa communication[3]. La présente recherche veut analyser cette technologie de numérisation des textes afin d'identifier les opérations importantes qui y sont effectuées ainsi que les produits textuels qui en sont issus. Elle s'intéressera à certains des enjeux herméneutiques de la lecture et de l'analyse des textes numérisés. 4

2. La numérisation comme technologie et comme herméneutique

La technologie de numérisation des textes met en relation deux termes : la *numérisation* et le *texte*. Avant d'entamer notre analyse, nous devons préciser certains points de la sémantique de ces termes afin d'éviter toute ambiguïté. 5

Le terme *numérisation*, dans un contexte informatique, revêt d'abord une double signification. Dans une première acception, il renvoie à une technologie qui convertit un signal physique (sonore, lumineux, mécanique, etc.) en un signal dit numérique qu'un ordinateur peut traiter. De multiples technologies effectuent de la numérisation : par exemple, les systèmes d'alarme, les ouvertures automatiques de portes, l'imagerie médicale, etc. Lorsque cette technologie est appliquée à des documents textuels, la numérisation fait référence aux divers processus physiques (optiques, mécaniques, électroniques, etc.) que réalise un périphérique informatique appelé un *numériseur* ou un *scanneur*. 6

Dans une seconde acception, la *numérisation* renvoie plutôt au traitement formel, c'est-à-dire aux processus algorithmiques qui opèrent dans un numériseur ou scanneur. En ce sens, la numérisation est un ensemble d'opérations de transformations qui, appliquées à des symboles ou signes linguistiques déposés sur un support physique (papier, microfiches, etc.), le transforment en un autre type de symboles ou signes qu'un programme peut traiter. Comprise ainsi, la numérisation produit un texte dit *numérisé*. Par exemple, si les symboles choisis sont des chiffres 0 et 1 (un encodage binaire,) alors le texte numérisé est un texte numérique à proprement parler (en anglais, *digital*), mais dans certains cas, les symboles choisis peuvent être des images formant des mots : comme dans le texte affiché sur un écran. Autrement dit, tout texte numérisé n'est pas nécessairement uniquement un texte numérique. 7

Cette double signification a son utilité. Elle permet au discours sur la numérisation d'être métonymique. De fait, ce terme naviguera souvent entre ces deux significations ; ce qui permettra de masquer la complexité tant du processus physique que du processus algorithmique mis en œuvre. Par exemple : si on dit qu'un service d'archives mène un projet de *numérisation*, on peut aussi bien penser que le projet mettra en place une technologie complexe de numérisation (les numériseurs) ou encore qu'il en appellera à des stratégies et processus algorithmiques pour réaliser la numérisation. 8

Un même type d'ambiguïté accompagne le terme *texte*. En effet, ce mot, selon les contextes, peut faire référence à un objet servant de support physique (papier, électronique) ou à un objet sémiotique, c'est-à-dire une entité linguistique complexe signifiante. Dans la première acception, le texte désigne le substrat physique d'un contenu sémiotique écrit[4], lequel se résume habituellement aux matériaux (papier, toile, carton, etc.) servant de support aux inscriptions scripturales (effectuées au moyen de crayons, d'encre, etc.), et constituant, une fois relié, un document textuel à part entière (qu'il s'agisse d'un codex, d'un livre[5], d'une brochure, d'un manuscrit, etc. (voir [Vandendorpe 2009]). Le texte constitue ainsi un objet physique dénombrable, rendant possibles et intelligibles des expressions telles que « *mettre un texte dans une* enveloppe » ou « *empiler dix textes sur un bureau* ». Dans le cadre informatique, le texte demeure aussi un objet physique, à ceci près toutefois, que le support physique change de nature et devient électromagnétique (disque dur, clef USB, etc.) ou lumineux d'un écran. Par ces transformations, la notion de texte se voit intégrée à de nouvelles pratiques discursives : on parle désormais de textes copiés par des imprimantes, reproduits par des scanneurs, traités par des logiciels, diffusés sur internet ou envoyés par des téléphones intelligents. Pour désigner le dernier cas, un nouveau synonyme, le « texto », est même apparu ; dans l'industrie, on dira dès lors que les messages « textes » sont plus économiques que les messages oraux. Dans tous ces cas d'énonciation et indépendamment de la forme matérielle qui lui est associée, du papier jusqu'au format électronique[6], le terme *texte* réfère au *support physique* d'un contenu signifiant écrit. 9

Parallèlement à cette conception du texte comme « contenant », une seconde acception, beaucoup plus classique, riche et complexe, renvoie au contenu proprement dit, c'est-à-dire à un ensemble organisé de signes linguistiques. Pris en ce sens, le texte est un objet sémiotique qui transcende sa matérialisation dans un support physique, ou du moins, il qui ne saurait s'y réduire. Conformément à cette acception, mille textes matériels imprimés peuvent tous être à propos d'un même texte sémiotique. Par exemple, une imprimante peut produire mille textes matériels du même texte de la constitution américaine. 10

Bien que dans certains cas, ces considérations sur les supports physiques soient importantes, c'est sur ce contenu du texte sémiotique que portent principalement les grands projets de numérisation. Pour ces derniers, le contenu de textes comme la Bible ou le Coran est en effet plus important que les supports physiques qui les ont portés au fil des siècles, de la peau de mouton au papyrus, en passant par le papier, le microfilm et, finalement, le support électromagnétique. 11

À la lumière de l'ambiguïté de ces deux notions de « numérisation » et de « texte », la problématique de la « numérisation de textes » s'avère beaucoup plus complexe qu'il n'y paraît de prime abord. En effet, par numérisation de textes, on peut autant renvoyer au processus physique qu'au processus algorithmique et à la manipulation des signes textuels. En conséquence, la compréhension de cette technologie de numérisation des textes sera constamment confrontée à cette ambiguïté. Aussi, pour explorer avec plus de précision ce qu'est cette technologie de numérisation des textes, nous aurons à répondre à deux questions épineuses : 1) Quelles sont les grands types d'opérations physiques et algorithmiques que la numérisation des textes met en œuvre ? 2) Quels effets ces opérations ont-elles sur 12

la lecture et l'analyse des textes ?

Les réponses à ces questions sont tout sauf simples. Et elles s'avèrent déterminantes pour la compréhension des pratiques d'une herméneutique matérielle[7] pour de la lecture et de l'analyse de textes. Il nous faut donc décrire plus en détail ces différentes opérations impliquées dans le processus de numérisation et présenter les impacts sur la dynamique herméneutique interprétative du texte numérisé. Nous espérons que les concepts mis en place éclaireront aussi les débats sur la lecture à l'ère numérique [Eberle Sinatra et Vitali-Rosati 2014] ou sur la lecture « distante » [Moretti 2013] et la lecture « électronique » [Baccino 2004]. 13

3. Les opérations et produits de la numérisation

À première vue, le processus de numérisation de texte semble très simple. Quelques secondes suffisent pour qu'un document textuel sur papier s'affiche sur un écran. La majorité des opérations physiques et algorithmiques mises en œuvre échappent à l'attention. Pourtant, le processus est complexe. En fait, il peut être décomposé en plusieurs phases, chacune constituée de plusieurs sous-opérations dont le produit est toujours un document constitué de signes qui se doivent d'être fidèles au texte source. En cela, la numérisation produit un document que nous appellerons un texte *numérisé*. Mais comme nous le découvrirons dans l'analyse qui suit, il existera plusieurs types de textes numérisés, chacun étant le résultat d'un type spécifique d'opération. 14

Dans notre description du processus de numérisation, nous symboliserons par une lettre indicée T_i le type de document textuel produit par chaque type d'opération. Ainsi, nous distinguerons (voir figure suivante) : 1) les textes sources T_c issus d'une opération de *collection*, 2) les textes T_p *sélectionnés* pour un *corpus*. Ces deux derniers de types de textes seront ceux qui seront soumis à la numérisation et qui produiront 3) les textes *électroniques* T_e, résultant d'une *transduction*, 4) les textes *proprement dits « numériques »* T_n, résultant d'un *encodage,* 5) les textes dynamiques T_d résultant d'une *reconnaissance optique de caractères*, 6) les textes *annotés* T_a, produit par une annotation, 7) les textes édités T_d, issus d'un travail d'*édition*. S'ajouteront à ces textes 8) des textes images Ti et 9) les textes à interpréter $T_{i.}$ Voici une représentation schématique de ces opérations et produits textuels. Nous expliciterons la spécificité de chaque type d'opérations et textes TiT_i produit. 15

16

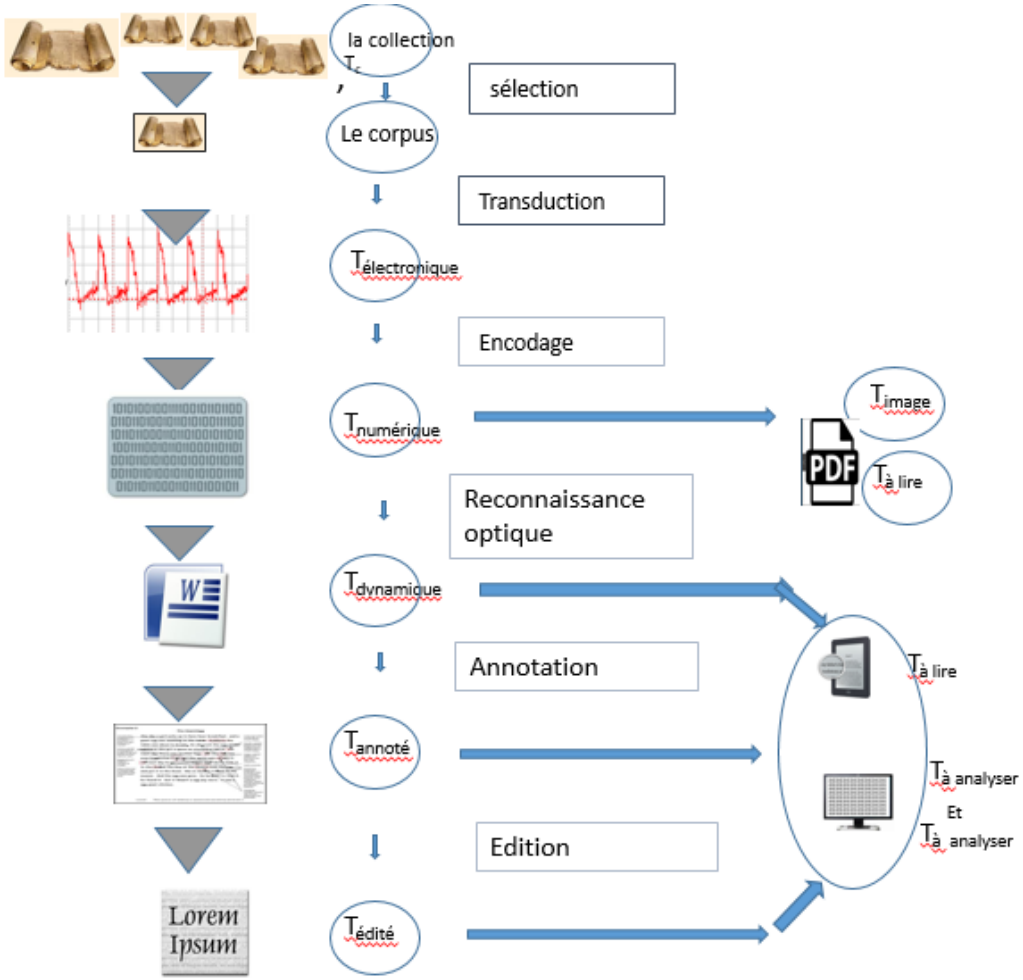

Figure 2.

Vu la multiplicité des objectifs et complexité des opérations en jeu dans les divers projets de numérisation de texte, cette liste n'est aucunement exhaustive. Il serait possible d'identifier de nombreux autres types d'opération et documents textuels, mais ceux retenus ici seront suffisamment pertinents pour révéler, ou à tout le moins illustrer, les enjeux herméneutiques de la numérisation des documents textuels sur leur lecture et leur analyse. 17

3.1 La collection des textes T_c et le corpus textuel T_p

Le premier moment de ce processus de numérisation peut lui sembler externe. Pourtant, il s'agit d'une étape qui lui est tout aussi essentielle qu'intrinsèque. En effet, dès le point de départ, un projet de numérisation doit distinguer au moins deux ensembles de documents textuels. 18

Un premier ensemble est constitué des documents textuels T(c) c'est-à-dire de documents qui forment une *collection* de textes sources pertinents et disponibles, que le projet vise à soumettre à la numérisation. Dans certains projets, on peut vouloir soumettre le plus grand ensemble possible de textes. Mieux vaut plus que moins, dira-t-on. Après, on pourra toujours revenir et choisir parmi les textes numérisés ceux qui intéressent des projets particuliers. On numérise alors à la chaîne tout ce qui semble pertinent. Par exemple, des bibliothèques pourraient décider de convertir en version électronique tous leurs documents, sans exception ou discrimination. Des revues scientifiques autrefois sur format papier pourraient envisager de convertir en format électronique la collection entière de leurs anciens volumes. 19

Mais dans plusieurs cas, surtout dans des projets de recherche, un choix sera effectué. Car il peut s'avérer impossible ou encore non pertinent de tout numériser. Ainsi, un deuxième ensemble de documents textuels T_p sera construit et il constituera le corpus. Par exemple, la collection T_c des œuvres écrites de Jean -Paul Sartre pourrait contenir 20

uniquement les œuvres publiées, délaissant la correspondance, les manuscrits, les cahiers de notes, etc. Et un projet de recherche pourrait ne retenir comme corpus T_p que les textes qui sont de nature philosophique.

21 On définit habituellement un corpus de textes T_p comme un sous-ensemble de la collection de textes à numériser ou déjà numérisés qui sont réunis en regard des objectifs d'une recherche ([Habert et al. 1997] ; [Rastier 2001] ; [Mayaffre 2002]). Dans certains domaines, le corpus peut s'avérer coextensif à la collection des textes T_c. Cependant, dans la grande majorité des cas, il n'en constitue qu'un sous-ensemble.

22 Pour construire tant une collection qu'un corpus, deux critères particuliers peuvent être considérés, l'un interne, l'autre externe. Le critère interne est lié aux hypothèses d'utilisation du corpus ou des manipulations qui lui seront appliquées. Aucun corpus n'est neutre et toute sélection de textes est déterminée ou normée par une pratique ou une théorie. Par exemple, des bibliothèques auront souvent à choisir des documents textuels à numériser pour construire leur propre collection électronique. Cette collection sera formée, entre autres, en regard d'une politique d'archivage ou encore des besoins supposés de leur clientèle. Pour constituer leur corpus, les littéraires pourront privilégier les textes qui sont susceptibles de susciter une critique ou une analyse ; les archivistes choisiront ceux qui peuvent constituer un témoignage prototypique d'un événement institutionnel ou social ; les linguistes sélectionneront ceux qui présentent des régularités d'une langue, tandis que les philosophes retiendront les textes les plus pertinents sur le plan théorique ou conceptuel. Bref, tout projet de numérisation débute par la sélection de textes et donc de la constitution, soit d'une collection, soit d'un corpus en regard des objectifs poursuivis. Ces collections et corpus ne sont pas sans liens avec ce que [Genette 1979], [Rastier 2001], [Jeanneret 2014], [Souchier et Jeanneret 1999], entre autres, appellent un ensemble architextuel, c'est-à-dire un ensemble de textes qui présentent une certaine unité sémantique (par genre, thème, etc.) et qui influencent le sens des autres textes avec lesquels ils sont réunis.

23 Un critère externe, de nature matérielle ou sociale, entre également en jeu. Sur le plan matériel, la constitution d'une collection ou d'un corpus peut demander une attention particulière à l'état physique des textes. Par exemple, la numérisation de textes anciens nécessite une analyse préalable de leur état de conservation, de leur dégradation ou de leur capacité à subir des manutentions mécaniques. De même, la numérisation de journaux et des revues ne peut se faire sans prendre préalablement en considération leur format, leur quantité, leur qualité d'impression, etc. La structure même des documents demande également à être analysée, et ce pour plusieurs raisons. Tout d'abord, les variantes structurelles entre les différents types de textes (articles de revue, lettres, pièces de théâtre, manuscrits, etc.), qu'ils soient inclus dans un même corpus ou non, nécessitent l'utilisation de procédés de numérisation adaptés pour chacun d'eux. Aussi, pour chacun, différentes transformations structurelles sont possibles. Par exemple, au sein d'un même document, on trouvera souvent des variantes dans la pagination, la mise en page, dans les types de polices, dans la justification, dans la disposition des notes. Autant d'éléments dont un processus sérieux de numérisation devra éventuellement s'occuper. Par exemple, si un texte contient des *marginalia* ou des commentaires, on devra se demander si ceux-ci doivent être conservés dans une version électronique.

24 Sur le plan social, certains projets auront peut-être à considérer les autres projets avec lesquels ils pourraient entrer en relation. Par exemple, un projet de numérisation des œuvres d'un auteur peut s'inscrire dans le cadre ou répondre aux objectifs de diverses politiques et activités organisationnelles, des centres de recherche aux bibliothèques et libraires. Dans de tels cas, les projets de numérisation peuvent avoir avantage à se conformer à différents standards industriels (XML, SGML) ou académiques (TEI). Par exemple, un projet de numérisation de la correspondance d'un écrivain français lauréat d'un prix Nobel de littérature aura peut-être à se conformer aux divers paramètres du corpus numérisé du *Trésor de la langue française* (TLF) dans lequel il pourrait se retrouver éventuellement.

25 Bref, tous ces critères matériels et ceux propres à un projet particulier jouent un rôle déterminant dans la construction d'un corpus. En tant qu'ils construisent une classe particulière de textes, ils établissent des relations entre les textes. Relations qui ne sont pas sans faire émerger du sens nouveau dans chacun d'eux. Une collection et un corpus ont donc un effet direct sur la lecture et l'analyse de textes. L'interprétation qui s'ensuivra sera différente de celle qui serait proposée d'un texte isolé sans lien avec une collection ou un corpus. On peut imaginer comment un corpus contenant *Le Capital* de Marx et *La Doctrine sociale de l'Église catholique* influencerait la lecture et l'analyse de l'autre. Dans cette perspective, une collection et un corpus instancient l'interdiscursif dont parlait Foucault.

3. 2 La transduction : le texte électronique (T_e)

26 Habituellement réalisé au moyen d'un dispositif appelé *numériseur* ou *scanner*, le processus de la seconde phase de numérisation consiste à modifier le document textuel issu de la collection ou du corpus habituellement déposé sur un support matériel papier, quelquefois encore sur microfiches ou même sur photos (textes anciens sumériens, grecs romains, arabes, cunéiformes sur divers supports physiques papyrus, pierre, etc.[8]), pour le convertir et le déposer sur un support *électronique*. Parce que réfléchissant la lumière, ces divers supports émettent des flux de photons, que des capteurs électroniques transforment (par transduction) en des signaux électriques discrets. Cette technologie matérielle est complexe. Elle repose sur un ensemble de sous-technologies aussi complexes, qu'ils soient physiques (rouleau, vitre, etc.), optiques (lentilles, filtres, miroir, etc.), mécaniques (roues, rubans, etc.), électriques (ampoules, moteurs, etc. ou électroniques (puces, semi-conducteurs, transistors, condensateurs, etc.). Prises conjointement, toutes ces technologies forment une « machine », c'est-à-dire un mécanisme physique intégré dont le produit final est une configuration de variations normées de voltage électrique. Ces configurations représentent le document source sous une forme électronique que nous appellerons ici, le document textuel électronique T_e. Il est un texte numérisé. Il faut bien noter cependant, que ce document textuel T_e n'est pas encore transformé en document textuel encodé par des chiffres 0 et 1, il n'est qu'une suite d'engrammes physiques (des « bits » physiques) à voltage varié, alternant et inscrit sur un support électronique (semi-conducteur) souvent miniaturisé (disque dur, mémoire, clef USB, etc.). Ces suites d'inscriptions électroniques ne sont pas directement « lisibles » par un humain[9].

27 À cette étape, le support électronique d'un document textuel n'est toutefois pas toujours isomorphe au document textuel source. Très souvent, la copie électronique ne retient pas tout ce qui se trouvait sur l'original. Plusieurs propriétés physiques informationnelles du document d'origine ne sont pas intégralement captées et reproduites. Ces pertes et bruits produits dans l'opération de transduction peuvent avoir plusieurs causes. Par exemple, la nature incandescente de la source lumineuse peut contribuer à restreindre le spectre lumineux. Relativement à ce même spectre, celui-ci peut n'être qu'imparfaitement capté ou converti par les capteurs, voire déformé par la lentille du numériseur. La vibration mécanique (*aliasing*[10]) du scanneur peut également interférer au niveau de l'enregistrement du voltage. Un échantillonnage statistique peut avoir été appliqué pour permettre une compression des informations. Ainsi, dans cette conversion d'un signal physique analogique et continu en un signal électronique atomique et discret, une perte non négligeable d'informations est susceptible de se produire.

28 En raison de cette complexité de l'opération physique, certaines précautions doivent être prises afin de s'assurer de la conformité du processus de numérisation aux objectifs initiaux : par exemple, on devra être attentif à la manutention physique des documents originaux[11], à la qualité et à la complexité du numériseur[12], à la couleur de numérisation à privilégier[13] ainsi qu'aux logiciels d'assistance, formats d'enregistrement, à l'espace disque et au support matériel à utiliser[14].

29 Dans tout ce processus de conversion du document textuel source en document textuel électronique TeT_e, certaines propriétés ou caractéristiques des signaux originaux sont omises ou laissées pour compte. Ou encore, certains ajouts peuvent produire du bruit. D'un point de vue herméneutique, si cet ajout ou cette perte d'information est négligeable pour des documents textuels simples ou d'utilisation courante (par exemple dans le cas d'un texte dactylographié), il en va tout autrement lorsque les textes originaux sont anciens et dégradés. Une numérisation de papyrus ou de vieux codex est particulièrement sensible à ce type de traitement : qu'il s'agisse d'un manuscrit médiéval ou d'un parchemin retrouvé dans une ancienne mosquée, l'omission d'une marque ou d'un signe particulier peut donner lieu à des interprétations radicalement différentes. Pour cette raison et afin d'éviter que des informations cruciales du texte original échappent ou soient ajoutées à la transduction du texte matériel en texte électronique, il est coutume de solliciter l'aide et l'expertise d'exégètes, de philologues et de paléographes à cette étape du processus de numérisation. À la lumière de ces différentes considérations, à mi-chemin entre la transduction électronique et l'interprétation textuelle, force est de convenir que le concept d'« herméneutique matérielle » prend ici un tout nouveau sens. La production d'un document textuel électronique T_e implique toujours des décisions herméneutiques relatives à la représentativité de texte électronique. Elle en appelle à une multitude d'actes interprétatifs.

3. 3 L’encodage : le texte numérique (T_n)

Le document textuel électronique T_e n’est cependant pas encore un texte numérique au sens propre du terme. Pour le devenir, chaque variation électrique du document textuel électronique doit être encodée en une suite de symboles appartenant à un code numérique binaire[15], c’est-à-dire composé uniquement des symboles 0 et 1. Seule cette forme de texte encodé est, à proprement parler, le texte *numérique* (T_n) ou en anglais le *digital text*. 30

Figure 3. Un texte numérique

Enregistré sur des supports électroniques auxquels on peut ajouter de l’annotation de formatage tels JPEG (*Joint Photographic Expert Group*) ou TIFF (*Tagged Image File Format*)[16], cet encodage binaire sert surtout à des fins de traitements algorithmiques ultérieurs, d’analyse, mais surtout de distribution et d’archivage. Par ailleurs, ce statut de texte numérique est fondamental pour une manipulation informatique. En effet, il s’agit du seul format qui peut être pris en charge par un ordinateur. Certains des documents textuels inscrits dans l’ordinateur et que nous présenterons ci-après auront ce format. 31

Il existe donc une différence importante entre le document textuel électronique (T_e), inscrit sur les supports électroniques et le document textuel numérique (T_n). Le document textuel numérique est en effet issu d’une technologie où des algorithmes complexes (la compilation) transforment les signaux physiques en une suite (et même de couches) de symboles 1 et 0. Cette suite de symboles est le langage natif d’un ordinateur. Comme [Desclés 1996, 103–145] l’a souvent montré, le processus de compilation qui produit ces couches de symboles est une sémiose formelle qui à partir d’une configuration originelle proche du texte électronique lui applique une série de transformations pour l’amener à format final qui sera ultimement celui que l’ordinateur manipulera à diverses fins : mise en mémoire, archivage, distribution, reconnaissance optique de caractères, affichage, annotation, édition, etc. Dans ces transformations, il peut s’insérer survenir encore un fois des ajouts et des pertes d’information qui ultimement peuvent produire un document textuel numérique présentant des différences avec le document textuel électronique et avec le document textuel source. Et donc, un texte numérique altéré en viendra à affecter la lecture et l’analyse. Ce n’est pas sans raison que les exégètes voudront le plus souvent consulter le texte source. 32

3. 4 L’affichage : le texte image (T_i)

Ce dernier document textuel numérique n’est normalement pas « lisible » (en tant que chiffres) par des humains. Il est 33

constitué uniquement de séquence de symboles propre au langage machine de base. En effet, il est possible pour un humain de reconnaître des suites de symboles 1 et 0, mais il sera exceptionnellement rare que des humains[17] puissent traduire rapidement de telles suites de ces symboles numériques dans des symboles appartenant à une langue naturelle ou mathématique. Pour que l'interprétation devienne possible, une autre transformation est nécessaire : il faut convertir le document textuel numérique de T_n en un format qui fait apparaître des symboles directement interprétables par des humains.

Une manière simple de procéder consistera à traduire des configurations de symboles binaires par l'activation d'une imprimante ou de par l'activation de point lumineux (pixels) sur un écran électronique (moniteur) ou par l'activation d'un projecteur sur une toile réfléchissante. Ainsi, est produit un document textuel image T_i. Ce document textuel est évidemment « lisible » c'est-à-dire qu'il peut être parcouru sur le plan de son contenu sémiotique. Il permet la lecture au sens cognitif de ce terme. 34

Ce document textuel image T_i est en un sens comme une photographie du document textuel original. De fait, il contient le texte original. En effet, apparaissent aussi toutes les autres marques que le document source présentait (ratures, corrections, tache, trous, etc.). Certains sont importants mais d'autres sont du bruit. 35

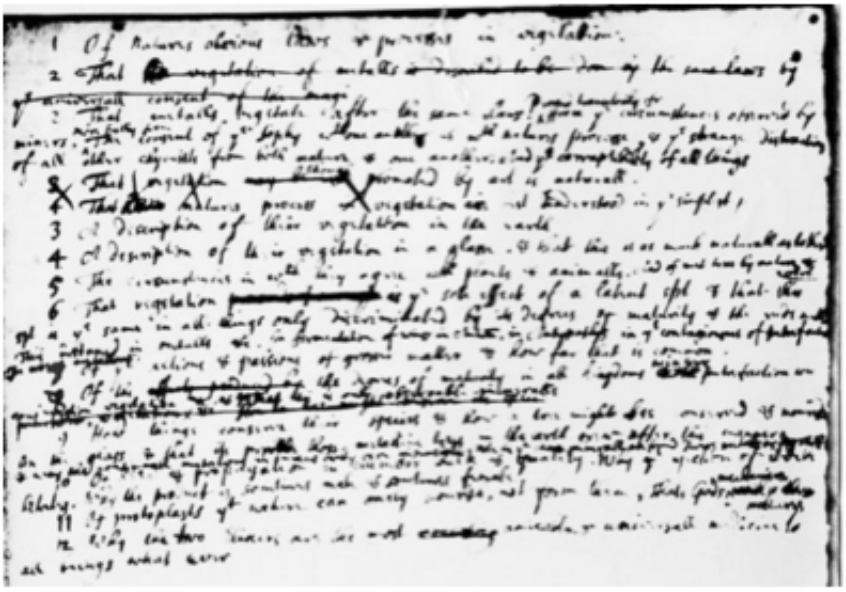

Figure 4. Texte image du manuscrit de *l'Alchimie* de Newton[18]

Malgré son contenu hybride, ce texte-image est précieux ; dans plusieurs domaines de recherche, il devra être conservé et facilement accessible et disponible pour les chercheurs. Faute d'avoir accès au document matériel originel, le chercheur pourra s'en servir comme socle de validation. Reste que ce document image n'est pas le document source. Dans un projet paléographique, un petit trou dans le manuscrit original peut apparaitre comme une marque sémio-linguistique importante (exemple : dans les documents textes anciens). 36

Encore une fois, comme dans les transformations précédentes, il existera divers types d'interventions qui toucheront ce texte-image ; par exemple : la compression, la résolution en pixels, le filtrage, etc.[19]. Ceci peut modifier substantiellement le contenu des différents types de textes impliqués. À chacune de ces diverses opérations des décisions affecteront l'interprétation du texte. 37

Par ailleurs, le texte-image pose des problèmes particuliers à la lecture, surtout si celle-ci porte sur le texte image-écran. Le parcours du texte impose des contraintes perceptuelles et cognitives qui ont été mises en évidence par les recherches. Le texte-image sur écran perd de nombreuses balises qu'offrait le codex. Des repaires physiques disparaissent. Il surcharge la mémoire. L'annotation, le commentaire, le marquage ne sont pas toujours accessibles. Ce sont autant d'éléments qui affectent la lecture, l'interprétation et la compréhension du contenu sémiotique. Cela dit, il 38

est important de noter que le texte-image n'épuise pas toutes les variantes des textes numérisés.

Bref, si nous résumons ces premières étapes, nous devons constater qu'il y a des enjeux herméneutiques distincts, mais importants. Chacune des étapes peut comporter de décisions qui modifient le document textuel, soit en ajoutant soit en éliminant quelque chose. Toutes ces décisions qui touchent la matérialité du texte peuvent ultimement affecter l'interprétation des textes. On imagine ce que toutes ces modifications pourraient signifier si le document textuel source était issu de la collection biblique des rouleaux manuscrits de Qumran ! Et qu'il faut lire ceux-ci sur en format PDF sur l'écran d'un téléphone intelligent ! 39

3. 5 La reconnaissance *linguistique* : le texte dynamique (T_d)

À ce stade de la chaîne de traitement, le texte-image (T_i) n'est qu'un ensemble de transcriptions de configurations de taches d'encre sur un support papier ou d'activation de pixels lumineux sur un écran. Certaines des configurations de tâches d'encre ou de pixels sont reconnues par les humains comme des symboles linguistiques, mais un ordinateur ne peut manipuler ces symboles comme des signes linguistiques. Le texte-image est figé, statique. Or, pour de nombreuses finalités de lecture, et surtout d'analyse et d'édition, de diffusion, l'ordinateur doit manipuler de manière dynamique ces symboles comme des signes linguistiques. Le texte-image doit donc être transformé en texte *dynamique* (T_d). Il y a deux manières de procéder pour ce faire : l'une manuelle, l'autre automatique. 40

En ce qui a trait à l'approche manuelle, il arrive que certains textes-images soient si complexes, bruités et idiosyncrasiques qu'aucun algorithme ne peut réussir à reconnaître des configurations de signes linguistiques. On peut penser ici aux manuscrits écrits à la main, en langue ancienne ou ceux contenant des symboles particuliers, comme les notes sténographiées de Husserl ou les textes de *l'Alchimie* de Newton. Dans de tels cas, il faut alors recopier au clavier[20], signe par signe, le texte-image. Cette procédure manuelle permet d'assigner directement un code numérique standardisé spécifique[21] (code ASCII : *American Standard Code for Information Interchange*) à chacun des signes linguistiques. L'ordinateur, par l'intermédiaire de logiciels de traitement de texte, peut ainsi manipuler directement des suites de codes 0, 1 comme des signes linguistiques et afficher des configurations de pixels ou d'encre correspondant à ces signes linguistiques[22]. 41

La procédure automatique repose quant à elle sur la reconnaissance optique de caractères (ROC)[23]. Celle-ci identifie dans la multitude des configurations de codes binaires (0, 1), du texte *numérique* T_n ou même du *texte-image* T_i – celles qui forment des signes linguistiques dynamiques : des lettres, ponctuation chiffres, espaces, etc.) et filtrant, si nécessaire des marques résultant de la texture du papier, des taches ou de tout autre source non pertinente du point de vue linguistique. Ces algorithmes complexes, basés sur des modèles mathématiques de reconnaissance ou de classification de formes, permettent dans les configurations de pixels des textes-images, notamment par le truchement de différentes opérations de translation, de rotation, de compression et de réduction ou d'agrandissement d'échelle, d'identifier des signes linguistiques et d'éliminer les effets dus au bruit ou aux imperfections. Ces opérations de reconnaissance utilisent parfois des dictionnaires ou des outils linguistiques. Il est évident que le document textuel dynamique T_d est distinct du texte-image, tout comme du texte numérique avec lequel il est souvent confondu. 42

Les signes linguistiques reconnus et affichés sur écran ou imprimés sur papier correspondent à des standards, par exemple, ASCII. Le texte peut alors être enregistré sous un format manipulable par des logiciels spécialisés dans le traitement de signes linguistiques. Les formats d'enregistrement les plus connus et utilisés sont TXT[24], RTF, ou ceux utilisés par des logiciels de traitement de textes[25] comme *Word* et *OpenOffice*. En général, le choix d'un format particulier dépend essentiellement des objectifs poursuivis. Mais il est évident que ce choix doit prendre en considération la durabilité, de ces standards pour l'archivage, et leur a compatibilité avec les divers outils informatiques de lecture et d'analyse. 43

Ainsi, à ce stade du processus le texte numérique dynamique[26] (T_d) peut certes être lu comme le texte-image, mais surtout il peut être, corrigé, souligné, commenté et ultimement traité par une variété d'algorithmes. Il reste cependant que ce texte dynamique est lui-même présent dans l'ordinateur comme un texte électronique manipulable dans sa 44

version numérique par des programmes.

45 Ici encore, cette phase de la chaîne de traitement présente ses propres enjeux herméneutiques. En effet, la transformation manuelle ou automatique à l'origine de la création du texte dynamique influencera la lecture et l'analyse. Tout comme dans le cas des textes sources électroniques, numériques, images celui-ci subira des transformations importantes. Par exemple, de multiples informations textuelles, tels le soulignement, le surlignage, les polices de caractères, la mise en italique, en gras ou en page, la pagination, les notes et commentaires peuvent être conservées ou éliminées. Des erreurs de reconnaissance dues notamment au bruit (une tache, une ombre, une interférence) peuvent s'y glisser[27].

46 De telles modifications affecteront éventuellement l'analyse du texte. C'est surtout d'ailleurs ce type de texte qui servira comme point de départ des multiples stratégies d'analyse du contenu textuel : annotations, lexicométrie, classification, visualisation, édition, etc. Ces opérations d'analyse exigent un texte dynamique. Encore ici, la différence entre une herméneutique classique et matérielle prend tout son sens.

3. 6 L'annotation : le texte annoté (T_a)

47 Dans sa forme la plus élémentaire, un texte numérisé dynamique n'est qu'une suite de caractères séparés par des espaces. Mais au sens sémiotique, un texte est plus qu'une suite de symboles. En effet, comme le soulignent régulièrement plusieurs linguistiques et sémioticiens ([Halliday et Hasan 1976] ; [Rastier 2001] ; [Adam 1999]) un texte est un objet sémiotique complexe. Il est un ensemble structuré de plusieurs niveaux de signes qui, conjointement, sont créateurs de sens. L'organisation de ces signes est complexe, car ceux-ci sont de divers types. Certains par exemple, appartiennent à l'organisation éditoriale du texte (la ligne, le paragraphe, le titre, le sous-titre, la note de bas de page et les numéros de page), alors que d'autres sont de nature linguistique (les mots, les phrases, les formes morphologiques ou grammaticales). Si les marques d'édition sont évidentes dans les textes numérisés, les marqueurs liés aux formes et aux contenus linguistiques y sont toutefois beaucoup plus discrets. Aucun texte numérique dynamique ne montre la différence grammaticale du mot « lit » dans la séquence alphabétique : « Il lit ce livre au lit. »

48 Aussi, si le projet de numérisation implique certaines manipulations sémiotiques du texte dynamique (T_d), il peut devenir nécessaire d'ajouter des informations spécifiques aux multiples types ou formes de signes présents dans le texte dynamique. Cela sera effectué par le biais d'annotations qui représentent sur le plan informatique des métadonnées, c'est-à-dire des étiquettes ou des marqueurs qui nomment la catégorie de l'information et qui sont ajoutés au texte dynamique.

49 Les types d'annotations varient selon les objectifs du projet de numérisation, qu'il s'agisse de production d'une édition papier ou électronique, d'archivage, d'aide à la recherche sur Internet, de construction d'un Web sémantique, de fouille ou d'analyse de données textuelles spécialisées. Plusieurs stratégies (manuelles ou automatiques) et classes d'annotations ont été proposées par le passé. Si certaines formes d'annotation portent sur le traitement informatique de type documentaire (indexation, archivage, notamment)[28], d'autres portent plutôt sur les formes linguistiques et textuelles ou relèvent de la sémioticité du texte[29]. En ce qui a trait à la lecture et à l'analyse, finalités premières des projets de numérisation de textes scientifiques et académiques, les formes sémiotiques d'annotation semblent être les plus pertinentes. Malheureusement, compte tenu de leur complexité et des difficultés qu'elles impliquent, la traduction informatique adéquate de ces formes d'annotations reste à faire. Nous nous contenterons ici d'en décrire quelques-unes.

50 Un premier type d'annotation relève de ce que [Genette 1987] appelle le péritexte. Ce terme renvoie à l'ensemble des signes qui, sous la responsabilité de l'auteur jouent un rôle externe, mais immédiat relativement au contenu du texte. Par exemple, sont dits membres du péritexte tous les mots ou passages référant à l'un des éléments ou dimensions textuels suivants : le titre, l'auteur, la date de publication, la référence, la pagination,les chapitres et sections, les épigraphes, la dédicace, la table des matières, les index et la couverture. Ce type d'annotations s'avère essentiel à la manipulation informatique du texte numérique. Par exemple, les marqueurs indiquant le numéro des pages ou des sections et des titres seront d'une importance cruciale pour le rappel, le résumé, la classification comme d'un point de

vue rhétorique ou argumentatif.

Les annotations intratextuelles portent quant à elles sur le contenu interne du texte. Celles-ci peuvent toucher différentes dimensions textuelles. Ainsi, on pourra vouloir indiquer le statut linguistique des signes – notamment : leur catégorie syntaxique (p. ex., « porte » comme *verbe* ou comme *nom*), leur catégorie sémantique (p. ex., « porte » comme *objet physique*, comme *action*, etc.) leur catégorie pragmatique, logique, rhétorique, discursive, etc. Certains voudront marquer le genre du texte, les attitudes, les sentiments, les jugements de valeur, les citations, etc. Certains voudront peut-être ajouter de l'information contextuelle, sociale, etc. 51

De plus, l'annotation peut avoir pour but de distinguer les signes non linguistiques présents dans le texte, mais qui participent de manière importante au contenu du texte sans pour autant constituer du « texte », par exemple les tableaux, les schémas, les cartes, les photos, les images et ainsi de suite. Finalement, on inclura aussi les commentaires, variantes, remarques, précisions, etc. – c'est-à-dire, les *marginalia* qui participent à leur manière au contenu sémiotique du texte. En vérité, la liste des types d'annotations intratextuelles est presque infinie. 52

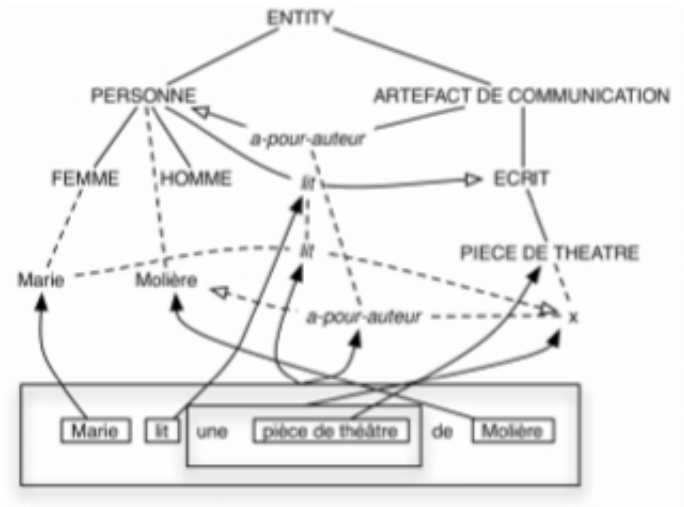

Figure 5. Exemple d'annotation de « Marie lit une pièce de théâtre de Molière »[30]

Une troisième catégorie d'annotation, relevant de ce que Genette appelle « épitexte » et que Foucault et Kristeva désignent par « intertexte », renvoie à des textes externes, liés de manière intermédiaire au contenu textuel principal. Certains éléments « épitextuels » peuvent être produits par l'auteur du texte principal, par exemple la correspondance, le journal intime et les interviews, alors que d'autres portent spécifiquement sur lui, par exemple les critiques et les analyses. Certains seront privés d'autres publiques. Enfin, d'autres formes d'annotations relevant du même « cadre discursif » que le texte annoté peuvent jouer un rôle important dans la compréhension de son contenu sémiotique, par exemple les textes historiques ou techniques. Bien que ces différents textes n'aient pas de rapport direct avec le texte principal, une analyse textuelle rigoureuse se doit toutefois de les prendre en compte. 53

Plusieurs projets de numérisation ont recours à ces divers types d'annotations, bien qu'à des degrés divers. Les pratiques peuvent faire preuve d'une grande variabilité de détail et de complexité. Compte tenu de cette diversité et des complications qu'elle est susceptible d'entraîner, une normalisation des marquages s'est imposée pour de nombreux projets de numérisation, normalisation permettant d'assurer à la fois une certaine cohésion interprétative et une communicabilité informatique. 54

Au niveau proprement informatique, cette normalisation des pratiques d'annotations s'est traduite par le développement de plusieurs standards, certains étant plus utilisés que d'autres. Le SGML (*Standard Generalized Markup Language*), fortement utilisé dans les projets de numérisation antérieurs aux années 2000, représente l'un des formats les plus connus. Sommairement, ce format est basé sur la distinction entre la structure dite « logique » du document et son contenu (titre, chapitre, sections, paragraphe, etc.). Utilisé par plusieurs industries documentaires, le SGML a toutefois 55

été relativement boudé par la communauté académique, notamment en raison de sa lourdeur, de son coût d'exécution élevé et de son manque de précision pour plusieurs types de tâches.

Un second type de marquage plus simple, XML (*Extensible Markup Language*), a rapidement remplacé SGML dans le monde industriel et académique. Permettant une plus grande variété d'annotations que ce dernier, il a notamment contribué au développement de plusieurs formes de marquage secondaire. Ainsi, l'émergence du Web a mené au développement des balises structurantes HTML, et de balises de forme CSS, lesquelles permettent non seulement la mise en forme de documents à des fins de publication en ligne, mais aussi l'insertion de liens *hypertextuels* entre le texte original et des textes connexes[31]. Les variantes RDFS (*Resource Description Framework Schema*) et OWL (*Ontology Web Language*) permettent pour leur part d'organiser l'information sémantique d'un texte par l'intermédiaire d'ontologies[32]. 56

En tant qu'ensembles terminologiques et conceptuels structurés, recouvrant la dimension sémantique d'un champ de connaissances, les ontologies peuvent servir à organiser les informations sémantiques contenues dans des textes, notamment en vue de faciliter leur intégration web. Dans la perspective où une bonne partie des textes numérisés est susceptible d'être affichée dans des sites internet, plusieurs spécialistes du domaine proposent d'intégrer de ces ontologies aux pratiques d'annotation habituelles[33]. 57

Dans l'ensemble, les différentes normes décrites ci-dessus ont certes contribué à uniformiser la mise en ligne des collections textuelles. Désormais, les formats XML et HTML ainsi que leurs variantes sont essentiels à tout projet de numérisation et de mise en ligne. Toutefois, ce type de balises ne saurait entièrement convenir aux projets académiques, dans la mesure où ceux-ci nécessitent bien souvent des formats de balisage plus fins. 58

Le format proposé par le consortium international de la *Text Encoding Initiative* (TEI), sorte de compromis entre les formats généraux SGML et celui des ontologies, semble mieux à même de répondre aux demandes du monde académique. Permettant une annotation textuelle sophistiquée, il est adaptable et n'empêche aucunement son insertion sur le Web. Parallèlement au développement et à la popularisation de la TEI, des types de plus en plus complexes d'annotations ont été conçus et introduits. Tout en permettant une plus grande précision, ces types d'annotations bénéficient également d'un certain statut consensuel, facilitant la collaboration et les échanges au sein de la communauté académique. Toutefois, ces formats tendent trop souvent à être lourds et de réalisation coûteuse. 59

Par le truchement de ces différents formats et techniques d'annotation, cette dernière phase du processus de numérisation produit ainsi un nouveau texte : le texte annoté (T_a). De plus comme il faut préciser le type d'annotations, il faut indexer ce texte : soit T_{ai} où *i* indique le type d'annotation effectuée. 60

Sur le plan interprétatif, les techniques d'annotation soulèvent leurs propres enjeux herméneutiques. Encore plus que pour les autres phases du processus de numérisation, ces opérations peuvent orienter profondément l'interprétation du contenu sémiotique du texte. Et dans ce contexte transformationnel, l'herméneutique matérielle est directement interpellée. 61

Un premier enjeu est la multiplicité des perspectives possibles. À l'inverse des autres opérations qui peuvent occasionner une réduction du texte, les annotations lui ajoutent au contraire une quantité non négligeable d'informations. En outre, ces nouvelles informations reposent souvent sur une diversité de cadres théoriques qui ne sont pas universellement partagés[34] : s'il est facile de s'entendre sur le fait qu'une suite linguistique particulière est un *verbe* ou un *titre*. Il n'en va pas de même lorsque différentes interprétations linguistiques, discursives, énonciatives ou socio-psychologiques sont sollicitées. En conséquence, il est difficile de proposer des types universels ou tout au moins généraux d'annotations, d'où le caractère fortement subjectif de tout projet d'annotation : à chaque utilisateur ou groupe d'utilisateur son système d'annotation. Face à cette situation, certains projets ont rejeté toute standardisation des annotations[35], lui préférant plutôt des marquages hybrides ou *ad hoc*. Ce type d'approche semble par ailleurs celle qui est devenue la plus acceptable et la plus pratiquée. 62

Malgré ces difficultés et la tendance générale qui en découle, la possibilité de découvrir et d'établir des normes 63

d'annotation minimales, applicables aux textes présentant certaines similarités (par exemple, d'ordre littéraire, philosophique ou technique), demeure néanmoins réaliste. Par exemple, un poème pourrait permettre une annotation sensible au vers, au verset, à la métrique, aux lignes ou aux strophes, sans pour autant que ces différentes annotations soient liées entre elles. Également, une annotation littéraire pourrait se résumer à identifier les personnages d'une pièce de théâtre ou à préciser certains types d'actes de langage. Une annotation sémiologique ou linguistique pourrait se contenter de distinctions entre actants, actions ou épreuve, de même qu'une annotation philologique pourrait se limiter à préciser des variantes dans des manuscrits. En ce sens, il semblerait donc possible d'effectuer certaines annotations générales communes, malgré le caractère « spécifique au domaine » d'un bon nombre d'entre elles. En fait, le cœur du problème de l'annotation est que celle-ci constitue une forme déguisée d'interprétation ou, pour reprendre l'expression de [Pincemin 2007, 12], l'expression technique, mais déterminante de plusieurs choix théoriques participant de l'interprétation d'un texte.

Un deuxième enjeu herméneutique est la multiplication des documents textuels annotés T_a. En effet, contrairement à une démarche herméneutique classique qui ne porte que sur un texte source canonique et ses diverses transcriptions ou éditions, l'annotation multiplie les types de textes presque à l'infini. De plus, ces annotations sont souvent transparentes au lecteur et à l'analyste : alors que dans le texte papier, auquel l'interprète peut ajouter des *marginalia*, les annotations demeurent visibles, il en va tout autrement dans le cadre numérique. En effet, même si les catégories d'annotations sont diversifiées ou précises, elles ne sont pas toujours visibles à la lecture immédiate (écran ou imprimé). Dans certains cas, leur origine peut même être inconnue, ce qui complique d'autant plus la tâche des interprètes. 64

Ainsi, dans le cadre numérique, l'annotation informatique multiplie les types de textes. Un texte annoté syntaxiquement est différent d'un texte non marqué sémantiquement. Ainsi, presque chaque texte numérique T_a se multiplie en n textes différents par l'ajout divers types d'annotations, complexifiant d'autant la démarche herméneutique. 65

3. 7 L'édition : le texte édité (T_s)

Les deux derniers textes, le texte-image et le texte dynamique, peuvent donner accès au contenu sémiotique textuel ; ils sont lisibles. Mais seuls les textes dynamiques et annotés permettent une manipulation computationnelle et analytique. Par contre, ces derniers documents ne sont habituellement pas la forme ultime que visent les projets de numérisation. On désire offrir aux différents lecteurs un document textuel édité T_d qui contiendra les multiples qualités résultant d'un travail éditorial propre à un document numérisé affichable sur écran ou ultimement imprimable sous une forme ou une autre. L'édition dite « électronique » pourra se plier à diverses normes ou pratiques selon les usages qu'on en fera (par exemple : les livres, les revues, l'accès libre, l'interopérabilité, le catalogage, la, pérennisation, les tablettes – liseuses, le multimodal, etc.). Bref, ces éditions électroniques ne contiennent plus uniquement des ensembles de « lignes de textes ». Elles créent des « textes en ligne ». 66

Ce travail d'édition permettra divers types d'accès au contenu textuel. Comme [Virbel 1993] l'avait bien vu, l'édition électronique permet une créativité importante dans les formes de présentation d'un texte et par conséquent au contenu sémiotique du texte. Nous n'en présentons ici que des échantillons. 67

Un premier est de type décompositionnel ; le texte édité déconstruit les formes classiques de la présentation du codex ou du livre connu. Par exemple, si dans textes édités pour des sites web (voir les sites web consacrés à Shakespeare[36], Kierkegaard[37], Russell[38], Wittgenstein[39], Claude Bernard[40], etc.), on retrouve des lignes de textes similaires (mais flexibles) à celles trouvées dans l'édition papier, on trouve aussi des textes décomposés en de multiples sous-textes qui deviennent tabulaires, réticulaires, empilés, gigognes, juxtaposés, hypertextualisés, navigables, etc. Dans ses formes fragmentées, le parcours du texte n'est plus uniquement linéaire, mais multidirectionnel ; il invite à parcours intra-, péri-, para- et architextuel. Ainsi, le texte édité ouvre à un contenu sémiotique hybride. 68

Un autre type est compositionnel. Ici le texte édité devient agrégation de segments de textes autonomes, qui, par exemple dans *Wikipédia*, peuvent provenir d'auteurs et de sources diverses et être l'objet de changement constant. 69

Comme le dit [Gabler 2010, 50], le texte édité est *supersegmenté.* Ces segments, que les programmeurs appellent des « *snippets* », permettent des recompositions infinies de nouveaux textes qui à leur tour peuvent être ajustés afin de répondre aux divers types de lecteurs. Ces *snippets* peuvent même être convertis en « textos » pour être diffusés dans les réseaux sociaux par l'intermédiaire de téléphones intelligents. Cela invite évidemment à une multitude de parcours de lecture. Il va sans dire le contenu sémiotique des textes devient alors de plus en plus hybride.

70 Une des dimensions importantes du travail éditorial classique plus particulièrement de l'édition experte, académique et critique est le sceau d'autorité et de validité qu'il appose un texte sémiotique. En effet, les éditeurs jouent un rôle de garant de la qualité d'un texte par la correction, l'évaluation, la disposition, l'ajout d'appareillage critique, etc. Or, l'édition électronique des textes, ce travail ne se retrouve pas toujours de manière évidente. Certes, on le voit dans l'édition de l'*Index Thomisticus* dont l'éditeur est un spécialiste : le jésuite Busa. La confiance, cependant, n'est pas la même pour les textes en libre accès, ou encore ceux de Wikipédia.

71 Bref, comme, le texte numérique édité (Td) modifie à sa manière la dynamique herméneutique. Les nouvelles formes d'édition comme le dit Gabler, elle invite au dépassement des frontières qui délimitaient l'édition classique :

> The digital medium has the potential to develop into an environment suitable to reintegrate textual criticism into criticism – and, just a importantly: to ground criticism again in textual criticism. [Gabler 2010, 46]

72 De ce fait, le texte édité n'est pas innocent sur le plan herméneutique. Il instaure une nouvelle forme de médiation structurelle, critique et évaluative entre le format du texte et son contenu. La lecture, l'analyse et la compréhension des textes en sont modifiées. Si dans certains cas, elles sont balisées par une édition classique et qu'elle invite à une compréhension proche de l'horizon connu du lecteur et de l'analyste, dans d'autres cas, elles plongent le lecteur dans une boite de Pandore dont l'issu peut être autant une impasse ou un cul-de-sac qu'un nouvel horizon à explorer et découvrir.

3. 8 La lecture et l'analyse : le texte à lire analyser et interpréter (T_l)

73 Dans la variété des types des textes identifiés jusqu'à maintenant nous pouvons distinguer deux ensembles de textes selon qu'ils donnent ou ne donnent pas accès immédiat au contenu textuel comme objet de lecture et d'analyse.

74 Le premier ensemble contient les textes électroniques et les textes numériques qui, bien que porteurs de marques ou de symboles, ne sont pas comme des textes lisibles et analysables par des humains ; ils ne peuvent ancrer la compréhension. Le deuxième contient les textes-images, les textes annotés, les textes dynamiques, les textes édités. Ceux-ci sont véritablement *les textes à lire, à analyser et à interpréter* (T_l) c'est-à-dire dire ils sont des textes signifiants, objets de sémiose et ultimement de compréhension.

75 Sur le plan de la lecture, l'expérience perceptuelle de la lecture est modifiée par l'introduction de tout nouveaux facteurs physiques susceptibles d'influencer le parcours visuel. Du nombre, citons notamment la grandeur de l'écran, le lieu, la luminosité, l'angle, le format de l'écran, la polarité, le lissage des caractères, le mode d'affichage de déroulement, le fenêtrage et le mouvement des yeux. À la lumière de la quantité et de l'importance de ces paramètres d'affichage, la lecture d'un même texte sur le moniteur d'un ordinateur de bureau, une tablette ou un portable est susceptible de produire des expériences textuelles différentes.

76 Comme l'ont montré de nombreuses recherches, la lecture papier, en raison notamment de la portabilité, durabilité, maniabilité et facilité d'annotation des livres, continue d'être préférée à la lecture-écran[41]. Elle semble donner plus facilement des lectures critiques, profondes et expertes du contenu textuel. Mais, le texte numérisé gagne aussi en importance, surtout en l'absence d'équivalents papier. Tout comme le texte papier, le texte numérisé permet aussi des lectures critiques, profondes et expertes. L'hypersegmentation et l'hypertextualisation permises par les formes éditoriales variées créent une nouvelle structuration de signes par lesquels le texte est exprimé. Le texte numérisé acquiert ainsi une flexibilité sans précédent. Tout peut en effet être transformé, du titre aux commentaires, de la préface à la postface, de la légende à la note et de l'argumentation à la rhétorique. Par ailleurs, cette flexibilité peut ouvrir à un

lecture gigogne, où chaque un segment s'ouvre à d'autres segments engouffrant le lecteur dans des cybersémioticités. Une telle sorte de lecture impose au lecteur de nouvelles charges cognitives au lecteur ([DeStefano et LeFevre 2007] ; [Ackerman et Goldsmith 2011] ; [Baccino 2004]). En augmentant considérablement la quantité de matériel textuel disponible, ce nouveau monde textuel contraint nécessairement le développement et l'adoption de différentes stratégies de lecture, par exemple la fouille, l'écrémage ou le marquage.

77 Si la lecture classique séquentielle convient fort bien aux romans policiers, rien toutefois ne permet de croire que cette forme traditionnelle de lecture textuelle continuera également de prévaloir pour d'autres types de textes. Selon le contexte et les objectifs de lecture, le format textuel numérisé, annoté de liens hypertextuels renvoyant à des définitions, à des explications ainsi qu'à des critiques et commentaires de spécialistes, sera peut-être préféré au format papier traditionnel, ouvrant ainsi la lecture a un parcours textuel plus éclaté. L'impact de ce changement de mode de lecture est bien évident dans le cas d'un ouvrage comme l'*Origine des espèces* de Darwin. Par les annotations et surtout l'hypertextualisation, le lecteur peut accéder tout au long de sa lecture à un corpus paratextuel et épitextuel formé de plus de 63 éditions différentes de l'ouvrage et de plus de 1500 sources secondaires. La lecture classique est ainsi rompue au profit de parcours de textes multiples, diversifiés et participant à interconnexion textuelle véritablement révolutionnaire[42].

78 Outre la lecture, l'analyse technique est aussi profondément modifiée par la numérisation de textes. En effet, le texte à analyser, parce que dynamique, annoté, édité permet une plus grande diversité d'approches analytiques assistées par ordinateur que le permettaient celles réalisées traditionnellement « à la main ». Elles étaient souvent laborieuses. Et certaines, bien qu'imaginables, étaient cependant souvent impossibles. Un texte numérisé, dynamisé, annoté, édité, etc., peut faire l'objet de nouveaux processus de classification, de catégorisation, de comparaison ou de fouille, etc. Également, de nouvelles formes d'analyse stylistique, linguistique, discursive, thématique, conceptuelle, narrative et rhétorique apparaissent ; l'impact sur le processus interprétatif est considérable.

79 Enfin, il va sans dire que la numérisation des textes affecte aussi grandement leur diffusion et leur partage. Certes, l'analyse de l'impact « distributionnel » des projets de numérisation massive, par exemple celui initié par Google pour les textes importants de l'humanité, reste à faire. Cependant, il n'en demeure pas moins que la numérisation des textes a profondément modifié les pratiques de partage du savoir, notamment au niveau du mode de fonctionnement d'organismes tels que les maisons d'édition, les librairies, les journaux, les bibliothèques et les universités. Enfin, sur le plan sémiotique, la numérisation favorise non seulement la diffusion et la distribution des textes, notamment dans de nouvelles communautés, mais également elle les intègre aux autres médias technologiques que ce soit à titre visuel ou sonore, avec l'image, le film et la musique.

80 Jusqu'alors limitée, sous sa forme classique, à l'interprétation des textes dans un horizon du sujet, de la culture et du savoir, l'herméneutique doit désormais s'ajuster au contexte numérique, tant à la nouvelle matérialité du texte qu'aux outils d'assistance et à la nouvelle pratique interprétative qui lui est liée. De nos jours, l'herméneutique classique ne peut donc se faire indépendamment d'une herméneutique matérielle. En dépit de cette nouvelle contrainte, toute démarche de ce genre aboutira néanmoins et toujours à la création d'un nouveau texte à lire (T_I) et, partant, à l'ajout d'un nouvel élément à la « galaxie » de textes liés au texte source.

4. Conclusion

81 Au fil de cette analyse, nous avons voulu mieux préciser la nature du texte non pas *numérique*, mais du texte *numérisé*. Le texte numérique n'est qu'une des formes d'encodage particulier qu'un texte peut recevoir au sein du processus de numérisation. Celle-ci , en effet, est un processus complexe qui permet de multiples transformations d'un texte. Elle produit le texte numérisé. Chaque opération de ce processus en est une de transformation d'un document textuel vers une autre forme de document textuel. Au départ, le texte source est transformé en un texte matériel : le texte encodé de manière *électronique*. Ensuite, celui-ci est transformé en divers types de textes sémiotiques : un premier, le texte « numérique » à proprement parler encode le texte par des symboles 0 et 1. Celui-ci, comme texte, est normalement illisible par des humains ; un second, dit « texte image », peut être affiché sur écran ou imprimé sur papier et lu en tant que tel, mais l'analyse y est surtout « manuelle » ; viennent ensuite le texte *dynamique*, le texte annoté et le texte *édité*.

Enfin apparaît le texte *à lire et à analyser*. Ainsi, partant d'un texte source sélectionné parmi une collection de textes, la numérisation produit non pas une copie unique dite « numérique » du texte, mais bien une véritable galaxie de textes numérisés. Interreliés et organisés hiérarchiquement, les textes numérisés formant cette galaxie ouvrent ainsi à des parcours nouveaux de lecture et d'analyse.

82 Dans une telle perspective, une critique ou une valorisation de la textualité numérisée doit être prudente. Les défauts et les qualités, les solutions et les problèmes, les avantages et les désavantages du texte numérisé ne s'appliquent pas à tous et de la même manière. Chaque format ou type de texte présente sa signature. Et il faut en saisir la forme, l'usage, la portée, la pertinence, pour en souligner les problèmes ou la valeur.

83 Enfin, la lecture et l'analyse des textes numérisés, quelle qu'en soit la richesse ou la finesse, ne peuvent jamais se faire de manière totalement automatisée, l'ordinateur ne pouvant ici jouer qu'un rôle d'assistance. Même à l'ère numérique, la lecture et l'analyse des textes demeureront une activité humaine. Elles ne peuvent être réduites à un processus intégralement algorithmique. Tout dans le monde n'est pas un modèle complètement computationnel.

84

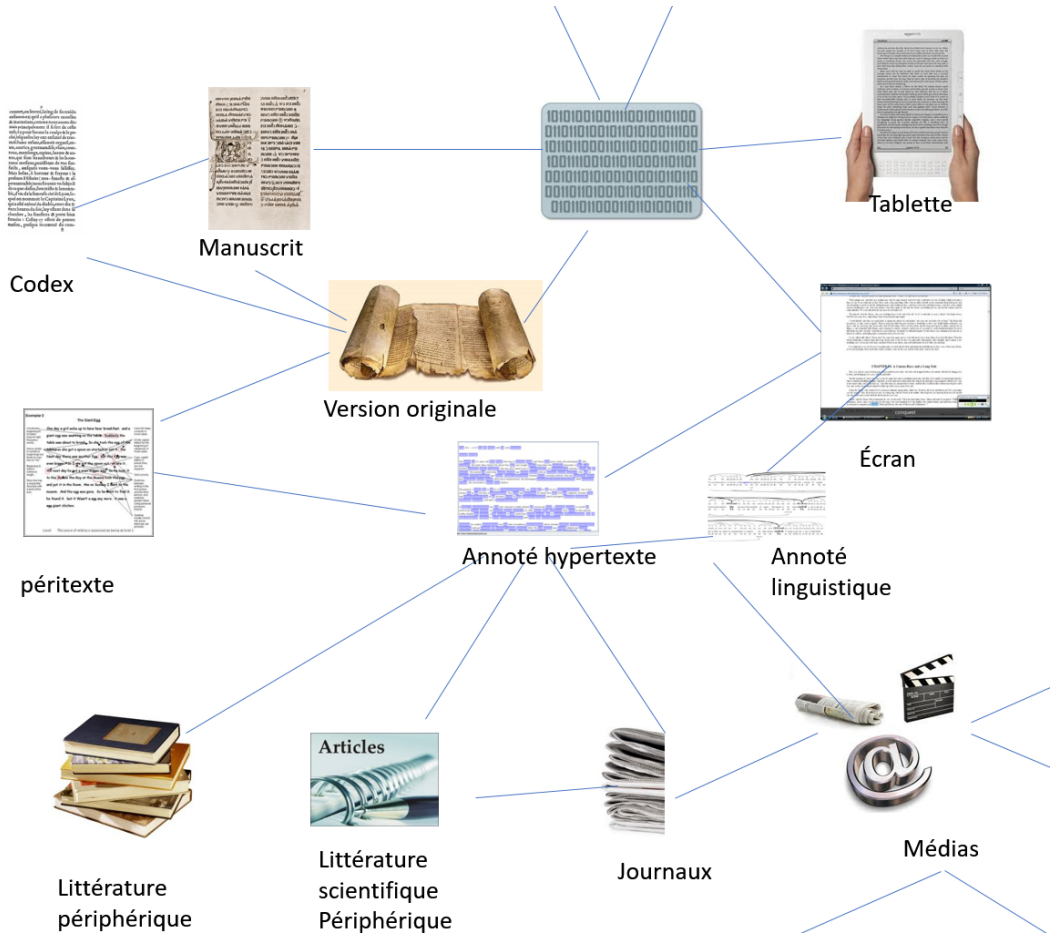

Figure 6. Galaxie numérique textuelle

85 Qu'elle soit classique ou matérielle, la pratique herméneutique est nécessairement interpelée par ces transformations, dans la mesure où elle ne porte plus sur un texte unique, mais sur une galaxie de textes. Par ailleurs, l'analyse et l'interprétation textuelles vont même jusqu'à jouer un rôle actif dans le processus de numérisation en soi, que ce soit au niveau de la mise en corpus, de la saisie électronique, de l'encodage numérique et l'annotation. À la lumière de ces transformations, l'activité interprétative se trouve du coup plongée dans un contexte dynamique radicalement différent du cadre herméneutique classique, que les thèses sémiotiques de Peirce et discursives de Foucault permettent de mieux comprendre et modéliser. L'interprétation porte toujours sur un système de signes canoniquement inscrits dans des puces, affichés ou imprimés, et elle navigue dans une galaxie de *systèmes de signes*.

Évidemment, ce nouveau paysage herméneutique n'est pas sans affecter le monde scientifique et culturel. Cette connectivité numérique, intertextuelle, hybride influenceront grandement le savoir des lecteurs et des analystes, leurs connaissances et désirs ainsi que leur langage. En ce sens, la numérisation des textes modifie profondément les fonctions sémiotiques classiques des textes ; elle les enrichit de pratiques rhétoriques originales, informées de nouvelles formes d'affirmation, de conviction, d'organisation et d'argumentation. Cela dit, malgré toutes ces transformations, Hermès veillera au grain.

Notes

[1] Les statistiques portant sur la totalité historique du matériel publié dans le monde varient. On parle d'environ 30 MK de livres, 750 MK d'articles, 25 MK de chansons, 500 MK images, 3 MK de vidéos ou programmes et 100 MK de pages internet. Dans une étude de 2011 [Thielens 2011], on comptait environ 285 exabits de données non structurées de divers types (images, son, textes, etc.). Pour un état de l'accroissement des données textuelles, voir : [Xiao 2008]Xiao, 2008.

[2] Voir les sites web de la Bibliothèque numérique mondiale (https://www.wdl.org/fr/) ainsi que du projet Gutenberg (http://www.gutenberg.org/) et de l'énoncé de sa mission par Michael Hart (http://www.gutenberg.org/wiki/Gutenberg:Project_Gutenberg_Mission_Statement_by_Michael_Hart).

[3] Ce qui n'est pas sans soulever de nombreuses questions politiques et juridiques. Sur cette question, voir [Jeanneney 2005].

[4] Les analyses philologiques du terme « texte » renvoient d'ailleurs en tout premier lieu au tissu et au tissage : « *Textus*, du participe passé de *textere*, est ce qui a été tissé, tressé, entrelacé, construit ; c'est une trame. » [Cerquiligni 1989, 59].

[5] *Wikipédia* le définit directement de cette manière : « Un livre (sens le plus courant) est un ensemble de pages reliées entre elles et contenant des signes destinés à être lus. » (https://fr.wikipedia.org/wiki/Livre)

[6] Ces propriétés matérielles du texte sont importantes. De nombreux historiens et théoriciens ont montré comment, à travers les siècles, la lente transformation du support physique du texte a été déterminante sur sa lecture et son analyse. Un texte inscrit sur la pierre ne pouvait être lu qu'oralement. Un manuscrit de copiste exigeait un lecteur spécialisé. Le codex ouvrait à un public plus large. L'insertion d'une gravure et d'une image orientait l'interprétation. La forme *livre* impose une lecture solitaire, séquentielle, un rythme spécifique. Elle contraint aussi l'analyse qui entre autres permet le soulignement, le commentaire, l'annotation (*marginalia*), etc. Chacun de ces types de supports physiques touchait de manière propre l'accès, le contenu signifiant du texte. Et aujourd'hui, le support électronique n'est pas sans offrir aussi une diversité de manipulations originales et spécifiques (déroulement, hyperliens, gigogne, archivage, transmission, affichages multiples, etc.) qui à leur tour ne sont pas sans affecter la lecture et l'analyse.

[7] Voir notamment : [Foucault 1969] ; [Rastier 2011].

[8] Voir le travail requis pour numériser les documents textuels et mis en ligne par l'Oriental Institute Open Access Publications : https://oi.uchicago.edu/research/catalog-publications.

[9] Sur ce plan, une image peut aussi être numérisée et devenir une suite d'inscriptions électroniques. Aucun humain ne peut, du moins normalement, y reconnaitre une image comme telle.

[10] Cependant, selon la qualité de la conversion, il est possible formellement et matériellement de reproduire de manière fidèle le signal analogique original. Certaines techniques de filtrage du bruit permettent de stabiliser, de nettoyer et de rendre plus précis le signal entrant (cf. l'effet Dither : effet de réverbération du bruit de la mécanique du moteur sur la numérisation).

[11] La manutention de certains documents anciens peut exiger des technologies plus fines, qui éviteront de les fragiliser davantage : par exemple, éviter de forcer la reliure, la délicatesse du changement de page, etc. L'ampleur et la finalité de la numérisation exigent des scanneurs appropriés : manuels ou automatiques. Dans les grands projets, la saisie est confiée à des entreprises spécialisées via des sous-traitances (*outsourcing*). Les plus simples sont à couverts plats ou en V, les plus sophistiqués sont à angles, ou même robotiques. Le choix d'un type particulier de scanneur dépendra des fins poursuivies. Il ne sera pas le même pour copies dactylographiées à archiver et à envoyer, à mettre sur Internet, ou pour ceux à inclure dans une collection de bibliothèques patrimoniales ou à conserver pour la postérité.

[12] Par exemple, la qualité des lentilles et la stabilité physique de l'appareil. Il existe divers types de scanneurs, les uns plus sophistiqués que les autres, et leur coût est évidemment lié à leur performance potentielle.

[13] Le noir et blanc doivent être limités à des copies de travail. Il y a trop de perte d'informations. Mieux vaut la couleur et avec la plus haute résolution et finesse possible pour des fins de conservation. Il est toujours possible alors de revenir à des copies de travail plus économiques en espace mémoire.

[14] On ne peut négliger les multiples sous-opérations impliquées dans une numérisation. De nombreuses fonctionnalités logicielles peuvent assister la numérisation et la rendre ergonomiquement plus facile. Par exemple, le choix des types de pages, des sections de pages, des copies multiples, de l'automatisation des fonctions, des outils d'édition, de l'ajustement des couleurs, etc.

[15] La sémantique de ce code : il réfère aux deux grandes classes de variations électriques de l'inscription électronique du signal, soit le positif ou le négatif.

[16] Les autres formats classiques de sortie de l'image sont le GIF (*Graphics Interchange Format*) et le PDF (*Portable Document Format*). Pour la numérisation académique, on utilisera surtout les formats TIFF, JPEG et PDF. En fait, une combinaison des trois sera souvent utile selon la multiplicité des usages faite des documents numérisés. Le format TIFF est le format le plus compatible avec les multiples plateformes logicielles et celui qui conserve le plus d'informations. Évidemment, il coûte cher en espace mémoire, mais il assure une stabilité dans la conservation et l'interchangeabilité logicielle. Le format JPEG est un format de compression. Il est le plus populaire pour des documents à mettre sur Internet et pour échanger. L'œil ne saisira pas la différence, mais pour l'archivage, ou des agrandissements et des manipulations plus fines, ce format sera problématique.

[17] Von Neumann « lisait » directement le code binaire : http://w3.salemstate.edu/~tevans/VonNeuma.htm.

[18] Extrait de la collection Dibner : http://webapp1.dlib.indiana.edu/newton/project/about.do.

[19] Plus la résolution est fine, meilleure est la précision de la copie électronique. La résolution doit être haute si la finalité est la constitution d'archives professionnelles ; elle peut être moins haute si la numérisation n'est qu'une étape vers la production d'un texte vivant (soumis au ROC). Le choix de la résolution sera préférablement haut (autour de 1000 dpi) pour de l'archivage patrimonial, mais pour une reconnaissance optique de caractères, quelque 400 dpi suffisent.

[20] Il est intéressant de noter que l'expression « numérique » associée à « clavier » renvoie habituellement au pavé numérique, c'est-à-dire le clavier avec des chiffres. Le clavier ordinaire d'un ordinateur est une technologie mécanique qui transforme une pression effectuée sur une touche en un signal électrique qui, à son tour, est transformé en un code numérique. Lorsque les textes à copier sont complexes, la saisie passe souvent par l'intermédiaire de plusieurs personnes. Elles encoderont manuellement, en parallèle et de manière comparée, le texte-image (ou le texte source lui-même) pour assurer la plus grande fidélité du texte vivant avec l'original.

[21] Il faut distinguer le jeu de caractères codés avec leur représentation en bits. Par exemple, le code ASCII est lui-même encodé en 8 bits dans la norme ISO 8859.

[22] Et dans ce cas, il faudra des stratégies complexes de vérification et de correction : par exemple, faire des copies parallèles qui sont comparées et co-corriger.

[23] Les logiciels de ROC ne travaillent jamais sur le document papier d'origine ; ils en font toujours, mais de manière transparente, une copie image, et c'est cette image via la représentation numérique des configurations lumineuses qui est soumise au logiciel de reconnaissance.

[24] Le format TXT élimine tous les marqueurs et ne garde que les espaces entre les mots alors que le RTF en conserve quelques marqueurs importants (comme les paragraphes et les italiques).

[25] Il est intéressant de noter la métonymie sous-jacente à cette nomination de « logiciel de traitement de textes ». *Word*™, par exemple, ne traite pas du texte sémiotique, mais des signes linguistiques encodés de manière standard. Pour ce logiciel, il n'y a pas de différence informatique entre « la klr ok kf prp oi klr » et « Il lit ce livre au lit ». Les deux sont des suites de signes linguistiques même si la première suite n'a aucun sens.

[26] Le texte est quelques fois appelé « vivant » ou en anglais « living » par les entreprises informatiques spécialisées en logiciels ROC. Mais le terme dynamique semble utilisé le plus souvent.

[27] Un nettoyage et un filtrage sont souvent par la suite nécessaires. Mais ces tâches peuvent être assistées par des outils informatiques. Par exemple, un extracteur de lexique peut fournir la liste des chaînes de caractères formant les « mots » mal identifiés.

[28] Voir [Goldfarb 1981] ; [Reid 1980] ; [Renear, Mylonas et Durand 1996] ; [DeRose 1999]DeRose, 1999.

[29] Voir [Marshall 1998] ; [Bird et Liberman 2001].

[30] Voir [Ma, Audibert et Nazarenko 2009].

[31] Une modification de XML a donné lieu (en 2007) à XHTML, mais ce dernier a été ensuite abandonné.

[32] Une ontologie est une spécification explicite et formelle d'une conceptualisation partagée d'un domaine d'intérêt. Les concepts y sont traditionnellement organisés en un graphe dont les relations peuvent être soit des relations sémantico-logiques, soit des relations de composition et d'héritage (conformément au paradigme objet). L'interprétation des ontologies est souvent de nature épistémique, dans la mesure où elles représentent des connaissances.

[33] Voir [Buitelaar, Cimiano et Magnini 2005] ; [Ma, Audibert et Nazarenko 2009].

[34] Il y a toujours une théorie latente qui opère dans la préparation du balisage. Comme le dit M. Sperberg-McQueen, « *Markup reflects a theory of text* » [Sperberg-McQueen 1991, 34].

[35] Les multiples projets d'annotation linguistique constituent à ce titre de bons exemples.

[36] Voir [Mueller 2008].

[37] Sur Wikipédia, l'accès aux textes de Kierkegaard (https://en.wikipedia.org/wiki/Søren_Kierkegaard) se fait par plusieurs sites interreliés où textes, paratextes et péritextes sont mis en interrelation par des commentateurs annotateurs et fort probablement revisés par la fondation Kierkegaard.

[38] Hébergé par *The Bertrand Russell Society* (https://users.drew.edu/~jlenz/brtexts.html).

[39] Hébergé par *The British Wittgenstein Society* (http://www.britishwittgensteinsociety.org) et *The Cambridge Wittgenstein Archive* (http://www.wittgen-cam.ac.uk/).

[40] Voir le site web Claude Bernard (http://www.claude-bernard.co.uk).

[41] Les premières recherches menées par Dillon ont démontré d'importantes différences dans les deux types d'expérience textuelle, notamment en ce qui a trait à la rapidité, à la précision, à la fatigue visuelle ainsi qu'à la compréhension [Dillon 1992].

[42] Pour [Kelly 2006], les liens hypertextes et les marqueurs représentent deux des plus importantes inventions des cinquante dernières années.

Works Cited

Ackerman et Goldsmith 2011 Ackerman, Rafaket, Goldsmith, Morris (2011). "Metacognitive regulation of text learning: On screen versus on paper". *Journal of Experimental Psychology: Applied*, 17-1 (2011): 18-32.

Adam 1999 Adam, Jean-Michel. *Linguistique textuelle: des genres de discours aux textes*. Paris, Nathan (1999).

Baccino 2004 Baccino, Thierry. *La lecture électronique: De la vision à la compréhension*. Grenoble, Presses Universitaires de Grenoble (2004).

Bird et Liberman 2001 Bird, Steven, Liberman, Mark. "A Formal Framework for Linguistic Annotation (revised version)". *Speech Communication*, 33, 1-2 (2001): 23-60.

Buitelaar, Cimiano et Magnini 2005 Buitelaar, Paul, Cimiano, Philipp, Magnini, Bernardo, " Ontology Learning from text: An Overview ". In Buitelaar, Paul, Cimiano, Philipp, Magnini, Bernardo Magnini (dir.). *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, IOS Press (coll. "Frontiers in Artificial Intelligence and Applications"): 3-12 (2005).

Cerquiligni 1989 Cerquiligni, Bernard. *Éloge de la variance : histoire critique de la philologie*. Paris, Éditions du Seuil (1989).

DeRose 1999 DeRose, Steven J., van Dam, Andries, "Document structure and markup in the FRESS Hypertext System." *Markup Languages* 1(1), Winter 1999: 7-32.

DeStefano et LeFevre 2007 De Stefano Diana, Lefevre Jo-Anne. "Cognitive load in hypertext reading: A review". *Computers in Human Behavior*. 23-3 : 1616-1641 (2007).

Desclés 1996 Desclés, Jean-Pierre. “Cognition, compilation, langage”. In Chazal, Gérard, Terrasse, Marie-Noëlle (dir.). *Philosophie du langage et informatique*. Paris, Hermès: 103-145 (1996).

Dillon 1992 Dillon, Andrew. “Reading from Paper Versus Screens: a Critical Review of the Empirical Literature”. *Ergonomics*, 35-10 (1992): 1297-1326.

Eberle Sinatra et Forest 2016 Eberle Sinatra, Michael, Forest, Dominic. “Lire à l’ère du numérique: *Le nénuphar et l’araignée* de Claire Legendre”. “Sens public”. 22 décembre 2016 : http://www.sens-public.org/article1230.html.

Eberle Sinatra et Vitali-Rosati 2014 Eberle Sinatra, Michael, Vitali-Rosai, Marcello (dir.). *Pratiques de l’édition numérique*. Montréal, Presses de l’Université de Montréal, coll. “ Parcours numérique ” (2014).

Foucault 1969 Foucault, Michel. *Archéologie du savoir*. Paris, Gallimard (1969).

Gabler 2010 Gabler, Hans W. “Theorizing the Digital Scholarly Edition”. *Literature Compass*. 7-2 (2010): 43-56.

Genette 1979 Genette, Gérard. *Introduction à l’architexte*. Paris, Seuil (2001).

Genette 1987 Genette, Gérard. *Seuils*. Paris, Le Seuil (1987).

Goldfarb 1981 Goldfarb, Charles F. “A Generalized Approach to Document Markup”. In *Proceedings of the ACM SIGPLAN-SIGOA Symposium on Text Manipulation*. New York, ACM (1981).

Habert et al. 1997 Habert, Benoît, Nazarenko, Adeline, Salem, André, *et al. Les linguistiques de corpus*. Paris, Armand Colin (1997).

Halliday et Hasan 1976 Halliday Michael, Hasan, Ruqaiya. *Cohesion in English.* Londres, Longman (1976).

Jeanneney 2005 Jeanneney, Jean-Noël. “Quand Google défie l’Europe”. *Le Monde*. 22 janvier 2005.

Jeanneney 2010 Jeanneney, Jean-Noël. “Quand Google défie l’Europe”. *Plaidoyer pour un sursaut*. Fayard, Mille et une nuits, Paris (2010).

Jeanneret 2014 Jeanneret, Yves. *Critique de la trivialité. Les médiations de la communication, enjeu de pouvoir*. Paris, Éd. Non Standard (2014).

Kelly 2006 Kelly, Kevin. “Scan this book!”. *New York Times*. 14 mai 2006: http://www.nytimes.com/2006/05/14/magazine/14publishing.html.

Kulkarni et Rokade 2014 Kulkarni, Kiran C., Rokade, Shashikant. “Review on Automatic Annotation Search From Web Database International Journal of Emerging Technology and Advanced Engineering Website”: www.ijetae.com, 4-1 (2014).

Ma, Audibert et Nazarenko 2009 Ma, Yue, Audibert, Laurent, Nazarenko, Adeline. “Ontologies étendues pour l’annotation sémantique”. In Gandon, Fabien L. *IC2009: Actes des 20e Journées francophones d’ingénierie des connaissances*. Hammamet, Tunisie, Mai 25-29. Grenoble, Presses universitaires de Grenoble (2009).

Mangen, Walgermo et Bronnick 2013 Mangen, Anne, Walgermo, Bente R., Bronnick, Kolbjørn. (2013) “Reading linear texts on paper versus computer screen: Effects on reading comprehension”. *International Journal of Educational Research*, 58 (2013): 61-68.

Marshall 1998 Marshall, Catherine. “The Future of Annotation in a Digital (Paper) World.” presented at the 35th Annual SGLIS Clinic: Successes and Failures of Digital Libraries, University of Illinois at Urbana-Champaign (1998).

Mayaffre 2002 Mayaffre, Damon. *Les corpus réflexifs: entre architextualité et hypertextualité*. *Corpus*, 1 (2002) : https://corpus.revues.org/11.

Meyers 2005 Meyers, Adam. “Introduction to Frontiers in Corpus Annotation II Pie”. In *The Sky Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*: 1-4 (2005).

Moretti 2013 Moretti, Franco. *Distant Reading.* Londres et New York, Verso (2013).

Morrison, Popham et Wikander 2013 Morrison, Alan, Popham, Michael, Wikander, Karen. “Creating and Documenting Electronic Texts”. *AHDS Guides to Good Practice. Oxford Text Archive* (2013): http://ota.ox.ac.uk/documents/creating/cdet/.

Mueller 2008 Mueller, Martin. “Digital Shakespeare, or Toward a Literary Informatics”. *Shakespeare*, 4-3: 284-301 (2008).

Newton v.1665 Newton, Isaac. *Trinity College Notebook*. Cambridge University Digital Library (1661-1665): http://cudl.lib.cam.ac.uk/view/MS-ADD-03996/1.

Noyes et Garland 2008 Noyes, Jan, Garland, Kate. “Computer- vs. Paper-based Tasks: Are They Equivalent?”. *Ergonomics*. 51-9 (2008): 1352-1375.

Pincemin 2007 Pincemin, Bénédicte. “Introduction”. *Corpus.* “Interprétation, contextes, codage”. 6 (2007): 5-15.

Rastier 2001 Rastier, François. *Arts et sciences du texte*. Paris, Presses universitaires de France (2001).

Rastier 2011 Rastier, François. *La mesure et le grain: sémantique de corpus*. Paris, Honoré Champion (2011).

Reid 1980 Reid, Brian. “A High-Level Approach to Computer Document Formatting”. In *Proceedings of the 7th Annual ACM Symposium on Programming Languages*. New York, ACM (1980).

Renear, Mylonas et Durand 1996 Renear, Allen H., Mylonas, Elli, Durand, David. “Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies”. In Ide, Nancy, Hockey, Susan (dir.). *Research in Humanities Computing*. Londres, Oxford University Press (1996).

Smith 1987 Smith, Joan M. “The Standard Generalized Markup Language (SGML) for Humanities Publishing”. *Literary and Linguistic Computing*. 2-3: 171-75 (1987).

Souchier et Jeanneret 1999 Souchier, Emmanuel, Jeanneret, Yves. “Pour une pratique de “l’écrit d’écran””. *Xoana*. 6: 98-99 (1999).

Sperberg-McQueen 1991 Sperberg-McQueen, Michael. “Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts”. *Literary and Linguistic Computing*. 6-1: 34-46 (1991).

Thielens 2011 Thielens, John. “Big Data Wizardry: Pay Attention To What's Behind The Curtain” (2011): https://www.forbes. com/sites/ciocentral/2012/02/23/big-data-wizardry-pay-attention-to-whats-behind-the-curtain/#384aca06752d.

Tolzmann, Hessel et Peiss 2001 Tolzmann, Don Heinrich, Hessel, Alfred, Peiss, Reuben. *The Memo of Manki*. New Castle, Oak Knoll Press (2001).

Vandendorpe 2009 Vandendorpe, Christian. *From Papyrus to Hypertext.* Urbana-Champaign, Illinois University Press (2009).

Veronis 2000 Veronis Jean. “Annotation automatique de corpus: état de la technique”. *Ingénierie des langues*. *Hermes*, 111-118 (2000): 1-52.

Virbel 1993 Virbel, Jacques. “Reading and Managing Texts on the Bibliothèque de France Station”. In Delany, Paul, Landow, George P. (éd.). *The Digital Word: Text Based Computing in the Humanities.* Cambridge, MIT Press (1993).

Weinnreich 1972 Weinnreich, Uriel. *Explorations in Semantic Theory.* Berlin, De Gruyter Mouton (1972).

Wästlund, Reinikka, Norlander et Acher 2005 Wästlund, Erik, Reinikka, Henrik, Norlander, Torsten, Acher, Trevor. “Effects of VDT and Paper Presentation on Consumption and Production of Information: Psychological and Physiological Factors”. *Computers in Human Behavior*. 21 (2005): 377 *sq*.

Xiao 2008 Xiao, Richard Z. “Well-known and influential corpora”. In Lüdeling, Anke, Merja, Kyto (dir.). *Corpus Linguistics: An International Handbook*, vol. 1. Berlin, De Gruyter Mouton (2008): 383-457.

DHQ: Digital Humanities Quarterly

2018
Volume 12 Number 1

Digital text: hermeneutic issues [fr]

Jean Guy Meunier <meunier_dot_jean-guy_at_uqam_dot_ca>, Université du Québec à Montréal

Abstract

The digitization of texts is omnipresent in the digital humanities. It seems to present itself only as a modification of the material medium: from text on paper to digital text. But it does more than that. Digitization also affects the text as a semiotic object. The multiple operations of this technology implement interpretative decisions that are not without their effects on the semiotic text; that is to say, the text that offers itself for reading and analysis. In this sense, the digitization of texts is not neutral. It is an important moment of material hermeneutics.

Note on Translation

For articles in languages other than English, DHQ provides an English-language abstract to support searching and discovery, and to enable those not fluent in the article's original language to get a basic understanding of its contents. In many cases, machine translation may be helpful for those seeking more detailed access. While DHQ does not typically have the resources to translate articles in full, we welcome contributions of effort from readers. If you are interested in translating any article into another language, please contact us at editors@digitalhumanities.org and we will be happy to work with you.

Works Cited

Ackerman et Goldsmith 2011 Ackerman, Rafaket, Goldsmith, Morris (2011). “Metacognitive regulation of text learning: On screen versus on paper”. *Journal of Experimental Psychology: Applied*, 17-1 (2011): 18-32.

Adam 1999 Adam, Jean-Michel. *Linguistique textuelle: des genres de discours aux textes*. Paris, Nathan (1999).

Baccino 2004 Baccino, Thierry. *La lecture électronique: De la vision à la compréhension*. Grenoble, Presses Universitaires de Grenoble (2004).

Bird et Liberman 2001 Bird, Steven, Liberman, Mark. “A Formal Framework for Linguistic Annotation (revised version)”. *Speech Communication*, 33, 1-2 (2001): 23-60.

Buitelaar, Cimiano et Magnini 2005 Buitelaar, Paul, Cimiano, Philipp, Magnini, Bernardo, “ Ontology Learning from text: An Overview ”. In Buitelaar, Paul, Cimiano, Philipp, Magnini, Bernardo Magnini (dir.). *Ontology Learning from Text: Methods, Evaluation and Applications*. Amsterdam, IOS Press (coll. “Frontiers in Artificial Intelligence and Applications”): 3-12 (2005).

Cerquiligni 1989 Cerquiligni, Bernard. *Éloge de la variance : histoire critique de la philologie*. Paris, Éditions du Seuil (1989).

DeRose 1999 DeRose, Steven J., van Dam, Andries, “Document structure and markup in the FRESS Hypertext System.” *Markup Languages* 1(1), Winter 1999: 7-32.

DeStefano et LeFevre 2007 De Stefano Diana, Lefevre Jo-Anne. “Cognitive load in hypertext reading: A review”. *Computers in Human Behavior*. 23-3 : 1616-1641 (2007).

Desclés 1996 Desclés, Jean-Pierre. “Cognition, compilation, langage”. In Chazal, Gérard, Terrasse, Marie-Noëlle (dir.). *Philosophie du langage et informatique*. Paris, Hermès: 103-145 (1996).

Dillon 1992 Dillon, Andrew. “Reading from Paper Versus Screens: a Critical Review of the Empirical Literature”.

Ergonomics, 35-10 (1992): 1297-1326.

Eberle Sinatra et Forest 2016 Eberle Sinatra, Michael, Forest, Dominic. “Lire à l’ère du numérique: *Le nénuphar et l’araignée* de Claire Legendre”. “Sens public”. 22 décembre 2016 : http://www.sens-public.org/article1230.html.

Eberle Sinatra et Vitali-Rosati 2014 Eberle Sinatra, Michael, Vitali-Rosai, Marcello (dir.). *Pratiques de l’édition numérique*. Montréal, Presses de l’Université de Montréal, coll. “ Parcours numérique ” (2014).

Foucault 1969 Foucault, Michel. *Archéologie du savoir*. Paris, Gallimard (1969).

Gabler 2010 Gabler, Hans W. “Theorizing the Digital Scholarly Edition”. *Literature Compass*. 7-2 (2010): 43-56.

Genette 1979 Genette, Gérard. *Introduction à l’architexte*. Paris, Seuil (2001).

Genette 1987 Genette, Gérard. *Seuils*. Paris, Le Seuil (1987).

Goldfarb 1981 Goldfarb, Charles F. “A Generalized Approach to Document Markup”. In *Proceedings of the ACM SIGPLAN-SIGOA Symposium on Text Manipulation*. New York, ACM (1981).

Habert et al. 1997 Habert, Benoît, Nazarenko, Adeline, Salem, André, *et al. Les linguistiques de corpus*. Paris, Armand Colin (1997).

Halliday et Hasan 1976 Halliday Michael, Hasan, Ruqaiya. *Cohesion in English.* Londres, Longman (1976).

Jeanneney 2005 Jeanneney, Jean-Noël. “Quand Google défie l’Europe”. *Le Monde*. 22 janvier 2005.

Jeanneney 2010 Jeanneney, Jean-Noël. “Quand Google défie l’Europe”. *Plaidoyer pour un sursaut*. Fayard, Mille et une nuits, Paris (2010).

Jeanneret 2014 Jeanneret, Yves. *Critique de la trivialité. Les médiations de la communication, enjeu de pouvoir*. Paris, Éd. Non Standard (2014).

Kelly 2006 Kelly, Kevin. “Scan this book!”. *New York Times*. 14 mai 2006: http://www.nytimes.com/2006/05/14/magazine/14publishing.html.

Kulkarni et Rokade 2014 Kulkarni, Kiran C., Rokade, Shashikant. “Review on Automatic Annotation Search From Web Database International Journal of Emerging Technology and Advanced Engineering Website”: www.ijetae.com, 4-1 (2014).

Ma, Audibert et Nazarenko 2009 Ma, Yue, Audibert, Laurent, Nazarenko, Adeline. “Ontologies étendues pour l’annotation sémantique”. In Gandon, Fabien L. *IC2009: Actes des 20e Journées francophones d’ingénierie des connaissances*. Hammamet, Tunisie, Mai 25-29. Grenoble, Presses universitaires de Grenoble (2009).

Mangen, Walgermo et Bronnick 2013 Mangen, Anne, Walgermo, Bente R., Bronnick, Kolbjørn. (2013) “Reading linear texts on paper versus computer screen: Effects on reading comprehension”. *International Journal of Educational Research*, 58 (2013): 61-68.

Marshall 1998 Marshall, Catherine. “The Future of Annotation in a Digital (Paper) World.” presented at the 35th Annual SGLIS Clinic: Successes and Failures of Digital Libraries, University of Illinois at Urbana-Champaign (1998).

Mayaffre 2002 Mayaffre, Damon. *Les corpus réflexifs: entre architextualité et hypertextualité*. *Corpus*, 1 (2002) : https://corpus.revues.org/11.

Meyers 2005 Meyers, Adam. “Introduction to Frontiers in Corpus Annotation II Pie”. In *The Sky Proceedings of the Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*: 1-4 (2005).

Moretti 2013 Moretti, Franco. *Distant Reading.* Londres et New York, Verso (2013).

Morrison, Popham et Wikander 2013 Morrison, Alan, Popham, Michael, Wikander, Karen. “Creating and Documenting Electronic Texts”. *AHDS Guides to Good Practice. Oxford Text Archive* (2013): http://ota.ox.ac.uk/documents/creating/cdet/.

Mueller 2008 Mueller, Martin. “Digital Shakespeare, or Toward a Literary Informatics”. *Shakespeare*, 4-3: 284-301 (2008).

Newton v.1665 Newton, Isaac. *Trinity College Notebook*. Cambridge University Digital Library (1661-1665): http://cudl.lib.cam.ac.uk/view/MS-ADD-03996/1.

Noyes et Garland 2008 Noyes, Jan, Garland, Kate. “Computer- vs. Paper-based Tasks: Are They Equivalent?”.

Ergonomics. 51-9 (2008): 1352-1375.

Pincemin 2007 Pincemin, Bénédicte. “Introduction”. *Corpus.* “Interprétation, contextes, codage”. 6 (2007): 5-15.

Rastier 2001 Rastier, François. *Arts et sciences du texte*. Paris, Presses universitaires de France (2001).

Rastier 2011 Rastier, François. *La mesure et le grain: sémantique de corpus*. Paris, Honoré Champion (2011).

Reid 1980 Reid, Brian. “A High-Level Approach to Computer Document Formatting”. In *Proceedings of the 7th Annual ACM Symposium on Programming Languages*. New York, ACM (1980).

Renear, Mylonas et Durand 1996 Renear, Allen H., Mylonas, Elli, Durand, David. “Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies”. In Ide, Nancy, Hockey, Susan (dir.). *Research in Humanities Computing*. Londres, Oxford University Press (1996).

Smith 1987 Smith, Joan M. “The Standard Generalized Markup Language (SGML) for Humanities Publishing”. *Literary and Linguistic Computing*. 2-3: 171-75 (1987).

Souchier et Jeanneret 1999 Souchier, Emmanuel, Jeanneret, Yves. “Pour une pratique de “l’écrit d’écran””. *Xoana*. 6: 98-99 (1999).

Sperberg-McQueen 1991 Sperberg-McQueen, Michael. “Text in the Electronic Age: Textual Study and Text Encoding, with Examples from Medieval Texts”. *Literary and Linguistic Computing*. 6-1: 34-46 (1991).

Thielens 2011 Thielens, John. “Big Data Wizardry: Pay Attention To What's Behind The Curtain” (2011): https://www. forbes. com/sites/ciocentral/2012/02/23/big-data-wizardry-pay-attention-to-whats-behind-the-curtain/#384aca06752d.

Tolzmann, Hessel et Peiss 2001 Tolzmann, Don Heinrich, Hessel, Alfred, Peiss, Reuben. *The Memo of Manki*. New Castle, Oak Knoll Press (2001).

Vandendorpe 2009 Vandendorpe, Christian. *From Papyrus to Hypertext.* Urbana-Champaign, Illinois University Press (2009).

Veronis 2000 Veronis Jean. “Annotation automatique de corpus: état de la technique”. *Ingénierie des langues*. *Hermes*, 111-118 (2000): 1-52.

Virbel 1993 Virbel, Jacques. “Reading and Managing Texts on the Bibliothèque de France Station”. In Delany, Paul, Landow, George P. (éd.). *The Digital Word: Text Based Computing in the Humanities.* Cambridge, MIT Press (1993).

Weinnreich 1972 Weinnreich, Uriel. *Explorations in Semantic Theory.* Berlin, De Gruyter Mouton (1972).

Wästlund, Reinikka, Norlander et Acher 2005 Wästlund, Erik, Reinikka, Henrik, Norlander, Torsten, Acher, Trevor. “Effects of VDT and Paper Presentation on Consumption and Production of Information: Psychological and Physiological Factors”. *Computers in Human Behavior*. 21 (2005): 377 *sq*.

Xiao 2008 Xiao, Richard Z. “Well-known and influential corpora”. In Lüdeling, Anke, Merja, Kyto (dir.). *Corpus Linguistics: An International Handbook*, vol. 1. Berlin, De Gruyter Mouton (2008): 383-457.