Décrire un corpus d'auteurs

Un thésaurus pour les types de textes

Ioana GALLERON(1), Fatiha IDMHAND(2),
Alexei LAVRENTIEV(3), Marie-Luce DEMONET(4),
Anne REACH-NGO(5)

(1) Université Sorbonne Nouvelle, LATTICE (UMR 8094)
(2) Université de Poitiers, ITEM (UMR 8132)
(3) CNRS, IRHIM (UMR 5317)
(4) Université de Tours, CESR (UMR 7323)
(5) Université de Haute-Alsace, ILLE (UR 4363)

Il est bien connu, depuis la démonstration de Genette dans son *Introduction à l'architexte*, que la notion de « genre » ou de « type » de texte couvre des réalités extrêmement hétérogènes, en dépit d'un certain nombre d'étiquettes largement utilisées (« roman », « drame », « épique », etc.). Au fil des époques, les tentatives de classification et de description selon des critères formels se sont succédées, avec des résultats mitigés, menant jusqu'à la formulation d'un doute fondamental quant à la possibilité d'établir des catégories (Chiss, 1987). Ceci n'a pas empêché les critiques, les linguistes et les historiens des idées, tout au long du XX^{e} siècle, de produire des articles et des ouvrages d'analyse et d'histoire littéraire qui font appel à des terminologies plus ou moins consensuelles, et plus ou moins stabilisées. De nouvelles dénominations ont été introduites au fur et à mesure du développement des travaux, afin de désigner les nouvelles formes d'écriture pratiquées depuis, notamment, l'avènement d'*Internet*.

Avec l'apparition, puis la montée en puissance des publications numériques, cette situation polyphonique pose de nouveaux problèmes. L'identification des textes dans l'espace virtuel, au milieu d'une masse de publications en croissance constante, nécessite que l'on applique aux corpus des procédures de description similaires. Or, bien que des recommandations comme celles de la TEI (The TEI Consortium, 2021) ou de Dublin Core (DCMI Usage Board, 2020) permettent de répondre à cet enjeu, force est de constater que les concepteurs de publications numériques ont du mal à relever ce défi, car dès lors qu'il s'agit de fournir des informations moins objectives et

plus substantielles à propos des textes numérisés, on manque de cadres communs. Par exemple, aussi simple qu'elle puisse paraître, la présentation des rôles des différents acteurs de l'édition du texte (tels que l'auteur, le contributeur, le collaborateur, l'annotateur, etc.) ne donne pas lieu aux mêmes pratiques. On observe des difficultés encore plus importantes lorsqu'il s'agit de décrire les thèmes de l'œuvre éditée, de désigner le travail réalisé dans le cadre de l'édition, ou de caractériser le type de textes publiés. Sans croire à une possible unification en la matière, et même sans la poser comme idéal à atteindre (les réponses à ce type de questions étant différentes et elles-mêmes porteuses de sens), il est permis d'espérer un peu plus de consensus sur certains de ces points et d'œuvrer dans cette direction. La description typologique des ressources textuelles est un des aspects où la convergence semble possible au moyen d'un vocabulaire partagé. C'est ce défi qu'a voulu relever un groupe de travail du consortium CAHIER[1], en s'attaquant au besoin de définition et de clarification des genres textuels édités par les concepteurs de ressources numériques et en élaborant une typologie de ceux-ci.

Le besoin de catégorisation des textes a déjà été identifié par le passé, et nous connaissons au moins une tentative destinée à y répondre, celle de David Y. W. Lee (Lee, 2001). Cependant, cette tentative n'est pas parvenue à essaimer, d'une part parce qu'elle semble avoir été entreprise dans le cadre d'un travail individuel, sans avoir été pensée pour être transposée, adaptée ou réutilisée facilement par d'autres chercheurs, et d'autre part parce qu'elle a proposé une réduction à des genres « prototypiques » – le roman, la poésie et le théâtre – ce qui ne tient pas compte, ainsi que l'avait remarqué Guy Aston (2001), de la diversité des genres, ni de leur poids dans les différentes cultures et en particulier leur fréquence. Dans le sillage de ces réflexions et démarches, on constate que CAHIER offre deux avantages : d'un côté, il fournit un riche ensemble de publications de natures très différentes, ce qui permet de constituer une base diversifiée pour l'identification de termes à inclure dans un vocabulaire partagé ; de l'autre, il offre un espace de discussion et de test pour la solution proposée. Cet article présente et contextualise la démarche entreprise, ainsi que les résultats des travaux menés entre 2016 et 2021.

Dans un premier temps, nous revenons en détail sur les classifications déjà existantes en insistant sur leurs limites par rapport aux besoins schématiquement présentés plus haut. Ensuite, nous décrivons le thésaurus que nous avons constitué, en évoquant d'abord les modalités de sa construction, puis les critères de son organisation et ses éléments de différentiation par rapport aux outils précédents. Enfin, nous fournissons quelques exemples de notices et d'utilisations souhaitées. Nous terminons en évoquant les évolutions futures de cet outil de description, aussi bien celles qui sont déjà prévues, que celles qui sont souhaitables, mais pas encore planifiées.

[1] Pour une présentation du consortium, voir https://cahier.hypotheses.org/le-consortium. La question des typologies textuelles a été abordée dans le cadre d'un groupe de travail *ad hoc*, réunissant les membres suivants : M.-L. Demonet (U. de Tours), A. Lavrentiev (CNRS), F. Idmhand (U. de Poitiers), I. Galleron (U. Sorbonne Nouvelle), G. Williams (U. Grenoble-Alpes), A. Réach-Ngô (U. de Haute Alsace), F. Vignale (U. du Mans). L'équipe rassemble ainsi des représentants de disciplines diverses : la linguistique, les études littéraires, les études de langue et littérature étrangères, les arts du spectacle, l'histoire du livre et de l'édition et la bibliothéconomie. Cette élaboration, grâce à un dialogue interdisciplinaire, la distingue des entreprises classificatoires précédentes.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Une précision supplémentaire est nécessaire, en préambule de cet article : notre travail s'est limité aux textes écrits, qu'il s'agisse de livres conservés dans une bibliothèque, de manuscrits archivés ou même de documents « nativement numériques » inclus dans un entrepôt. Les genres oraux n'ont pas encore été pris en compte, pas plus que les enregistrements natifs, même si des proximités évidentes se dessinent entre les objets produits avec ces différents modes d'expression, par exemple dans le domaine de la narration.

1 État des lieux

Pour décrire les genres des textes, une première ressource à laquelle nous avons pensé recourir a été le langage d'indexation RAMEAU, de la Bibliothèque nationale de France[2]. RAMEAU propose des vedettes de forme désignant des genres littéraires (roman, théâtre pastoral, littérature fantastique, etc.), ainsi qu'une série de vedettes désignant des notions littéraires (par exemple allégorie, narration poétique). Il propose également des subdivisions intéressantes pour les documents ou les groupes de textes : on peut préciser l'origine géographique, la langue, le rattachement chronologique, le sujet ou la forme (entendue soit comme « forme particulière de publication des textes eux-mêmes », soit en renvoyant à des littératures et à des genres littéraires[3]). En pratique, le jeu d'étiquettes génériques et les possibilités de description fonctionnelle ou formelle des textes s'avèrent limités. L'extraction des genres et des termes spécifiques sous la vedette « Littérature[4] » offre une liste de 137 notions, parmi lesquelles on retrouve des classes fort hétérogènes (comme par exemple « Littérature – Auteurs noirs », « Littérature – Femmes écrivains », « Littérature antique », etc.), et très inégalement utilisées d'après nos consultations de la base. De fait, une recherche dans le catalogue de la BnF pour identifier des textes représentatifs d'un certain genre, à une période donnée ou pour un auteur donné, ne donne souvent aucun résultat, car une grande partie des documents n'a pas été décrite du point de vue du genre ou de la forme. On n'obtient ainsi aucun résultat lorsqu'on croise les critères « Sujet, genre ou forme : comédie » et « Date : 1720 » (ou 1735), ni lorsqu'on croise le critère « Auteur : Nivelle de La Chaussée » et « Sujet, genre ou forme : comédie ». La lecture des notices montre, en effet, que *Le Préjugé à la mode* de ce dernier n'est pas décrit comme « comédie », mais que tel est le cas pour la plus obscure *Marseilloise* de Michel Levesque de Gravelle, peut-être pour éviter les confusions générées par le titre. De même, une recherche par « sujet, genre, forme » pour « dialogue » ne donne aucune réponse pour la période 1500-1550, pourtant riche dans ce genre d'œuvres – comme le montre d'ailleurs la présence de neuf manuscrits sur Gallica.

Dans un esprit similaire, nous avons consulté l'ontologie CIDOC-CRM, devenue depuis 2006 un standard pour la documentation des héritages culturels. Malheureusement, à part deux classes (très) générales auxquelles nos textes se rattachent (E31 : Document et E33 : Linguistic object), CIDOC-CRM ne permet pas de les décrire plus finement. Plus encore, les seules propriétés des textes spécifiées par le modèle

[2] Pour plus d'informations, voir https://rameau.bnf.fr [consulté le 11 décembre 2020].

[3] Voir BnF, Service des référentiels, Centre National Rameau, *Guide d'indexation RAMEAU*, 7e édition, 2017, section 2.4, p. 93.

[4] Pour la procédure suivie et les résultats obtenus, voir https://groupes.renater.fr/wiki/cahier/rameau__-__procedure__d__extraction.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

semblent être le titre (E35) et la langue (E56). Un « objet informationnel » (surclasse à laquelle les « documents » et les « objets linguistiques » appartiennent) est réputé avoir un « type » (E55), mais les modes de création d'une taxonomie spécifique de « type » ne sont pas détaillés dans la documentation de l'ontologie, et les exemples d'usage fournis sur le site n'aident pas à y voir plus clair[5].

De son côté, la base de données Frantext[6], qui permet d'interroger un ensemble de près de 5500 textes français et de construire des sous-corpus sur le critère de « genre », s'avère d'une grande hétérogénéité en la matière. On y constate d'importants déséquilibres entre les types de textes présents dans la base, avec 1366 d'entre eux classés dans les romans, tandis que 67 des 170 termes de la liste ne sont représentés que par un texte du corpus. Outre les « romans », on a également « roman policier » (23 textes), « roman arthurien » (7 textes), « roman d'aventures » ou encore « Roman » avec majuscule, représenté par 2 textes. Si des relations hiérarchiques entre ces différents descripteurs existent, elles ne sont pas clairement établies.

En laissant de côté de tels outils, finalement inadaptés à nos travaux, le même constat d'insuffisance et d'inadaptation peut être fait au sujet de ressources plus spécifiques issues des domaines de la linguistique ou de la documentation, domaines qui, plus que le champ littéraire, s'interrogent depuis longtemps sur la classification systématique des textes. Ainsi, la proposition « Brunet-Parguez », fondée sur une « classification des libraires », est-elle utilisée depuis le XIX^{e} siècle pour organiser les stocks de livres en cinq grandes sections : « Religion », « Droit et jurisprudence », « Sciences et arts », « Belles-Lettres » et « Histoire et sciences auxiliaires de l'histoire ». Cette arborescence répond, en premier lieu, à l'objectif de rassembler les ouvrages en des groupes cohérents, en offrant suffisamment de descripteurs pour les classes « riches » (par exemple, « Poètes de langue française depuis 1628 »), mais en évitant de multiplier les catégories peu fournies, du moins dans les bibliothèques et librairies françaises[7]. Or, pour ce faire, cette classification mélange sujets, auteurs et types de textes : on y trouve donc à la fois des termes désignant des genres (comme « Grammaire », « Dictionnaire », « Introduction »…), et d'autres à la signification moins bien définie comme le descripteur « Langue hongroise, langue valaque ou daco-romaine » qui peut renvoyer à des ouvrages dans ces langues, ou à des ouvrages sur ces langues. De même, « Poèmes épiques, héroïques, mythologies, chevaleresques » peut être compris comme une seule classe de textes présentant les quatre traits, à moins qu'il ne s'agisse d'un regroupement de quatre types de poèmes, mis ensemble parce qu'aucun n'est réputé renvoyer à un trop grand nombre d'éléments le représentant.

Au cours des années 1970 à 2000, l'analyse du discours s'est longuement penchée sur l'étude des genres. Celle-ci a été souvent entreprise dans une perspective didactique, qui vise à améliorer l'enseignement de la reconnaissance et de la production de différents types textuels. De ce fait, les classes de textes mentionnés dans de tels articles ou ouvrages sont souvent très larges, ou très spécifiques, l'objectif n'étant pas de réaliser un inventaire raisonné de toutes les possibilités, mais d'introduire au-

[5] Voir http://www.cidoc-crm.org/useCasesPage.

[6] Voir : http://www.frantext.fr.

[7] Par exemple, « Poésie orientale – poètes arabes, hébreux, persans, arméniens et turcs ». On remarque dans cette catégorie le regroupement d'auteurs d'aires culturelles fort contrastées.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

près des élèves quelques notions usuelles. Une autre caractéristique de ces travaux est qu'ils se concentrent plutôt sur les niveaux macro-génériques (les « types idéaux » de Werlich, 1987[8], les quatre catégories binaires de Longacre, 1983[9], les huit groupes d'Adam, 1985, fondés sur les actes de langage[10], les trois niveaux descriptifs de Monica Fludernik, 2000[11], etc.), laissant à des études ultérieures le soin d'énumérer et de décrire les ramifications en types ou formes textuelles subséquentes. Proposant des réflexions sur des principes classificatoires (Petitjean, 1989) et sur les enjeux de la description linguistique de textes littéraires (Combettes, 1987), de telles approches, quoique stimulantes d'un point de vue intellectuel, manipulent des étiquettes trop peu nombreuses ou trop larges pour pouvoir être utilisées de façon efficace dans l'indexation des textes, opération au cours de laquelle l'annotateur, au plus proche de la singularité du texte, recherche un certain niveau de précision en même temps qu'il rattache son objet à une classe générale.

Enfin, un troisième type de travaux linguistiques, menés dès les années 1980, mais se développant surtout après la généralisation des ordinateurs et des corpus dans les années 2000, s'est attaché à la classification automatique des genres. Fondés sur une critique des critères externes qui définissent traditionnellement les genres, ces travaux entendent se concentrer sur les caractéristiques « internes », permettant de discriminer différents types de textes. Toutefois, si l'idéal est de parvenir à des classes fondées sur des critères objectifs, produites indépendamment de leur (re)connaissance préalable, les approches adoptées combinent usuellement des démarches inductives et déductives, qui posent comme préalable une certaine représentation (voire une certaine intuition) de l'organisation du champ discursif. Par ailleurs, l'utilisation de variables morpho-syntaxiques pour regrouper les textes s'est avérée, à la longue, plus opératoire pour traiter de manière contrastive de grands ensembles (texte juridique vs. autres types de textes), ou pour étudier certains groupes d'écrits[12], que pour proposer une typologie en remplacement des catégories incomplètes et contradictoires qui ont cours dans les études littéraires. Le constat de l'hétérogénéité constitutive des genres (Mellet, 2013) prend le pas sur l'objectif de l'établissement d'une « typologie des textes […] indépendante de la théorie des genres » (Malrieu et Rastier, 2001).

Si dans les études littéraires, il existe une longue tradition de travail avec des dictionnaires et des encyclopédies – une offre abondante de tels outils s'est développée à travers le temps, et continue à occuper une part considérable du marché du livre universitaire –, ils s'avèrent peu utilisables dans la perspective de l'annotation des textes, pour de multiples raisons. On observe ainsi que l'objectif des dictionnaires et

[8] Rappelons que Werlich oppose le « type idéal » du texte à la « forme » du texte et à l'« idiome » du texte. Les types « idéaux » sont en nombre de cinq : description, narration, exposition, argumentation et instruction. Ils sont différemment réalisés dans les écrits concrets, avec peu de textes, par exemple, qui se réclament purement du descriptif, et beaucoup du narratif.

[9] Des textes organisés selon une succession temporelle sont opposés à des textes qui en sont dépourvus, et les textes « orientés agent » à d'autres qui ne le sont pas.

[10] Les macro-catégories d'Adam sont : narratif, descriptif, explicatif, argumentatif, injonctif, prédictif, conversationnel, rhétorique.

[11] Se fondant sur une proposition plus ancienne de Virtanen et Warvik, Fludernik propose cinq macro-genres (narratif, argumentatif, instructif, conversationnel et réflectif), et un nombre indéterminé de types textuels mettant en œuvre, subséquemment, un nombre (semi ?)-déterminé de modes de discours.

[12] Dans Poudat *et al.*, 2006, l'idée d'une description textuelle qui combinerait traits lexicaux et traits morpho-syntactiques est posée, mais l'application n'est faite que sur un corpus d'écrits scientifiques.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

encyclopédies de termes littéraires n'est pas toujours de préciser le sens, ou l'empan d'application d'un concept. Cet objectif, que l'on peut qualifier de « sémasiologique », passe souvent derrière l'objectif scientifique de réalisation d'un état des lieux sur une question. Ainsi, plus que les dictionnaires de langue, les ouvrages que la communauté littéraire désigne comme « dictionnaires » permettent de comprendre les transformations des notions et des concepts utilisés dans la discipline, de découvrir différentes perspectives historiques ou théoriques sur les textes, tout en constituant par moments un lieu de discussion, qui ouvre de nouvelles perspectives sur l'objet étudié, comme le ferait un article de revue. Dès lors, ce n'est pas vers le dictionnaire que l'éditeur d'une ressource numérique aura naturellement tendance à se tourner lors du choix de mots-clés pour décrire le texte annoté. Familier de ce dernier, c'est moins un éclaircissement quant à son genre qu'une précision quant au bon usage des termes, et de leurs combinaisons, qu'il sera amené à chercher. Faut-il déclarer *Le Tartuffe* de Molière comme du « théâtre du XVII^{e} siècle », comme « comédie » (ou « comédie classique ») ou comme une « pièce en cinq actes et vers » ? Faut-il combiner ces différentes étiquettes, ou bien répartir leurs contenus sur plusieurs métadonnées ? Tout en étant susceptibles d'éclairer le sens de chacun de ces termes, les dictionnaires ne sauraient répondre à de telles questions.

Pour terminer cette première partie, rappelons que la prééminence du médium papier parmi les encyclopédies littéraires constitue une autre raison qui les rend inappropriées, ou peu appropriées, au travail envisagé. La référence à un concept est à la fois lourde et difficile à résoudre quand elle renvoie à un tel outil. Outre le fait que le champ des mots-clés dans une édition XML/TEI ne fournit pas naturellement une place pour les précisions bibliographiques sur les auteurs, le titre, le lieu et la date d'édition de l'ouvrage-source, de tels renvois seraient peu pratiques à la fois pour l'annotateur et pour l'utilisateur d'un corpus annoté, dans la mesure où ils impliquent l'achat ou la consultation de l'ouvrage dans une bibliothèque, ainsi qu'une manipulation prolongée de celui-ci afin d'identifier le fragment visé. Quant aux dictionnaires et encyclopédies de termes littéraires existant en ligne, ils combinent rarement les deux qualités indispensables, à savoir d'avoir à la fois été élaborés par des spécialistes, et d'avoir adopté le principe de l'*open access*. La résolution des références dépendrait ainsi de la possibilité pour le lecteur d'accéder à ces ressources, soit par ses propres moyens, soit par des comptes institutionnels. La solution du thésaurus, que nous offrons, ne présente pas ces barrières financières ou spatio-temporelles, même si, étant nativement numérique, elle présente à son tour ses propres limitations techniques, comme nous le verrons plus loin.

L'analyse des théories, classifications et outils existants montre donc qu'un thésaurus de typologies textuelles, construit spécifiquement pour les publications numériques de corpus d'auteurs, est hautement nécessaire. Un tel outil doit s'appuyer sur les ressources préexistantes, mais ne saurait en constituer une simple transposition dans le domaine du numérique, comme nous le montrerons dans les sections suivantes.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

2 L'organisation intellectuelle du thésaurus

Le thésaurus proposé par le groupe de travail de CAHIER emprunte différents traits conceptuels et procéduraux aux approches classificatoires décrites plus haut et s'en différencie en vertu même de ce panachage. Cinq impératifs ont présidé à sa création :

1. Couvrir un ensemble aussi large que possible d'étiquettes, même si l'exhaustivité en la matière ne peut pas être atteinte, ou du moins pas d'emblée.
2. Trier les étiquettes, éliminer certaines, et relier d'autres entre elles, afin de proposer un ensemble manipulable, dans lequel l'annotateur ne se perde pas parmi trop de concepts proches ou similaires.
3. Concevoir une arborescence qui permette la classification des textes selon différentes macro-catégories et aspects (ou critères) proposés à travers le temps.
4. Proposer une classification non-redondante par rapport aux autres informations sur le texte, apportées par les éléments bibliographiques attendus dans une édition scientifique, qu'elle soit traditionnelle ou numérique.
5. S'assurer que la classification proposée peut être exprimée dans la syntaxe utilisée par le Web de données.

D'un point de vue technique, le thésaurus a été construit en utilisant le gestionnaire Opentheso[13], un outil multi-langue et multi-hiérarchique, déjà utilisé par les archéologues, hébergé et mis à la disposition de la communauté scientifique par l'infrastructure Huma-Num[14]. La prise en main d'Opentheso est facile, y compris pour les profanes en matière de Web sémantique, en raison de son interface intuitive ; toutefois, Opentheso présente quelques limites, comme l'impossibilité de donner des définitions pour les micro-thésaurus[15].

Comme pour les outils élaborés dans le monde des bibliothèques, la constitution du thésaurus CAHIER est fondée sur une démarche pragmatique et collaborative, qui part d'objets concrets (ouvrages, liasses, cahiers, boîtes d'archive, etc.) ayant fait l'objet d'une numérisation et d'une publication numérique dans l'un ou l'autre des projets du consortium CAHIER. Nos points de départ sont ainsi les paramètres éditoriaux et formels de ces objets, mais aussi les auto-descriptions génériques pratiquées dans les textes mêmes. On obtient ainsi une première liste « à plat », susceptible de convenir à la description de textes.

Parallèlement, une démarche *top down* a été adoptée, fondée sur l'analyse des similitudes et des recoupements entre les typologies mentionnées dans la précédente section, mais aussi des pratiques de l'édition numérique et de l'histoire des textes. Cette confrontation a conduit à l'adoption de plusieurs « foyers classificatoires » (Petitjean, 1989), capables de rendre compte de la complexité inhérente des textes. Ceux-ci se

[13] Voir : https://opentheso.hypotheses.org.

[14] Pour plus d'informations sur l'infrastructure, voir https://www.huma-num.fr.

[15] Ceux-ci réunissent les étiquettes de base, les intitulés des classes générales dans lesquelles nous regroupons les différents descripteurs.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

partagent d'une part en paramètres de haut niveau, constituant autant de micro-thésaurus, et des « catégories de triage » qui leur sont attachées, et dans lesquelles on trouve les étiquettes à proprement parler. Nous pouvons ainsi rendre compte de la matérialité des textes (que nous couvrons à travers la classe « mode d'agencement »), détailler certains éléments de leurs conditions de production (« public cible » et « domaine »), décrire leur intention communicationnelle telle que perçue par l'éditeur/annotateur (« type de discours »), apporter des informations sur leur « forme » et enfin, caractériser leurs contenus, non pas du point de vue des thèmes ou des *topics* développés, mais du point de vue de leur « factualité », de leur « origine » et enfin de leur rattachement à une ou plusieurs catégories intuitivement reconnues, désignées en tant que « genres ».

La figure suivante présente une vue schématique de l'arborescence du thésaurus, en se concentrant sur les trois premiers (et principaux) niveaux de profondeur : celui des paramètres, celui des catégories de triage et celui des étiquettes descriptives à proprement parler. Il importe de souligner qu'il s'agit ici d'une présentation simplifiée, à la fois en ce qui concerne les niveaux de profondeur dans notre thésaurus, et en ce qui concerne les relations susceptibles d'exister entre les différents niveaux et concepts, dont le positionnement n'est pas strictement hiérarchique, contrairement à ce que suggère l'image.

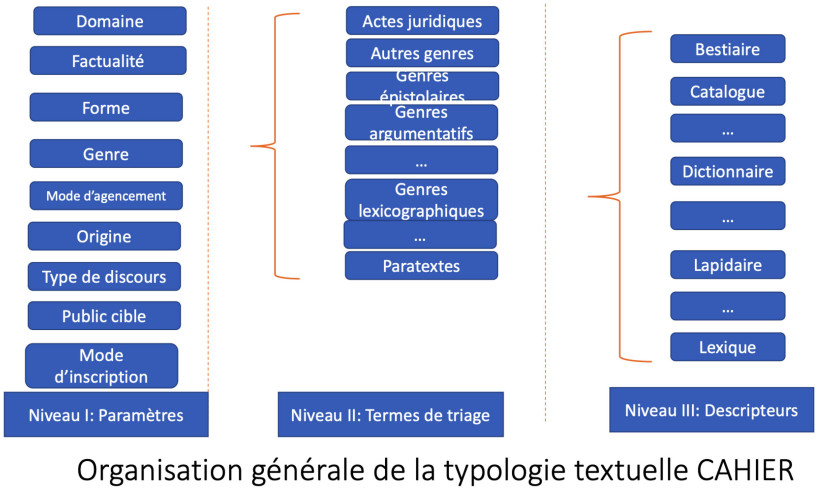


Fig. 1 : Organisation du thésaurus « Typologies textuelles »[16]

[16] Voir : https://opentheso.huma-num.fr/opentheso/?idt=43.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Selon cette arborescence, on peut donc décrire un texte selon neuf aspects. Nous donnons dans ce qui suit une description de chacun de ces paramètres.

1° Le « Domaine »

On trouvera sous cette appellation des étiquettes renvoyant à un découpage intellectuel du savoir, ou plus largement des productions de l'esprit humain, le texte se rattachant à un des « territoires » ainsi délimités. La déclaration du « domaine » implique de se poser la question de la sphère où le texte est destiné à circuler : « intime », « juridique », « régional », « religieux », telles sont quelques-unes des étiquettes que nous proposons dans ce micro-thésaurus.

L'intégration du domaine dans la classification des textes est argumentée de façon convaincante par Lee dans sa propre proposition typologique (2001). Toutefois, il n'en fait pas une classe de descripteurs combinables avec d'autres, même si, dans les faits, il l'intègre en tant qu'élément définitoire des genres oraux et écrits qu'il identifie dans le BNC (British National Corpus) : par exemple W_ac_nat_science, W-newsp_brdsht_nat_misc, S_sportslive, etc. Outre les critiques que l'on peut formuler au sujet des domaines retenus, cette utilisation de l'aspect le rend du coup moins opératoire, plus rigide, et en tout cas peu transposable à d'autres situations que celle du corpus de départ.

2° La « Factualité »

On classe ici les textes en trois grandes catégories : fictifs, non-fictifs ou mixtes. Cet aspect s'inspire, sans la reprendre exactement, de la catégorisation ternaire des textes par Werner Wolf (1993), selon qu'ils sont « illusionnistes », « anti-illusionnistes » ou « non-illusionnistes » (voir aussi Genette, 1979 et 1986).

3° La « Forme »

Un texte peut être en prose, en vers, dialogué, ou enfin graphique (tabulaire, calligramme, nuage…). Nous nous écartons donc, ici, de la définition de la « forme » dans la classification de Lee, qui suit Van Dijk (1985).

4° Le « Genre de texte »

Il s'agit, incontestablement, du micro-thésaurus le plus riche, avec plus de 220 membres[17] donnant lieu, pour certains, à plusieurs subdivisions.

[17] 222 à ce jour.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Les concepts proposés dans cette classe proviennent d'abord des auto-descriptions génériques des textes numérisés, mentionnées plus haut. Celles-ci ont été normalisées[18], puis enrichies avec des dénominations puisées dans des ouvrages de référence[19], dans différentes entreprises numériques[20] et, non en dernier lieu, dans la connaissance encyclopédique des différents spécialistes du texte ayant collaboré à ce groupe de travail.

Nous ne nous lançons pas ici dans une description détaillée de cette classe, qui serait fastidieuse, mais les notices individuelles que nous présentons plus loin comme exemples permettront de mieux en saisir les contenus. En raison de la ramification potentielle en plusieurs sous-genres et sous-sous-genres, cette section a demandé le plus de travail d'organisation et de définition.

5° Le « Mode d'agencement »

Cette nouvelle catégorie a été élaborée par le groupe de travail pour pallier un manque. Dans les systèmes actuels, la description formelle des supports va jusqu'à la précision des dimensions des feuillets ou des encres, mais laisse de côté les descriptions plus globales, pourtant extrêmement usuelles et fort éclairantes. Cette catégorie permet de désigner des modes d'organisation des textes tels que des « recueils », des « agendas », des « cahiers », des « liasses », des « mélanges » ou des « posters » dont la forme peut conditionner et orienter les modalités d'appréhension et d'interprétation de leurs contenus par les lecteurs. Le « mode d'agencement » possède ainsi une véritable capacité de description typologique.

6° L'« Origine » d'un texte

Cet aspect permet d'indiquer s'il s'agit d'une œuvre originale ou bien d'une adaptation, d'un commentaire, d'une continuation, d'une histoire dérivée, d'un « prequel », d'une recension, d'un résumé, etc.

7° Le « Public cible »

Ce renseignement est destiné à permettre l'identification des textes « genrés » ou des textes destinés à une catégorie d'âge particulière. Dans les

[18] Pour ne donner qu'un exemple, le théâtre français du XVIII^e siècle est particulièrement inventif lorsqu'il s'agit de caractériser ses productions : on relève des sous-titres comme « pièce héroï-tragi-comique », « fait historique », « opuscule dramatique », etc. Dans certains cas, ces déclarations visent à cerner les contours d'un genre, dans d'autres, il s'agit de jouer sur les termes, principalement dans une perspective parodique. Il convient donc d'opérer un tri entre les différentes appellations et de se contenter, dans certains cas, d'une description plus « neutre » de l'ouvrage (par exemple en tant que « parodie »).

[19] *Routledge Encyclopaedia of Narrative, Dictionnaire de termes littéraires, Dictionnaire de la caducité des genres littéraires, Dictionnaire des genres et notions littéraires.* Voir les références dans la bibliographie générale.

[20] Outre une extraction des éléments typologiques pratiqués par Frantext, nous avons utilisé la liste des « genres littéraires » publiée par Wikipedia (https://fr.wikipedia.org/wiki/Genre_litt\%C3\%A9raire.), et regardé les descripteurs textuels du BNC (https://www.english-corpora.org/bnc consulté le 15 mars 2022).

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

éditions, l'étiquetage selon cet aspect doit s'appuyer sur un élément explicite du paratexte qui renvoie à un public donné, plutôt que sur l'intuition de l'éditeur. Sans précision de la part de leur auteur, un conte ou un roman ne peuvent pas être déclarés comme étant destinés à « la jeunesse », même si les contenus ressemblent à ceux que ce type de public lit habituellement. À l'inverse, toutes les indications auctoriales ne sont pas à prendre pour argent comptant : pour n'en donner qu'un exemple, *Les Petits Contes nègres pour les enfants des blancs*, de Blaise Cendrars, ne s'adresse certainement pas qu'aux petits Européens et Nord-Américains.

8° Le « Mode d'inscription »

Un texte peut être déclaré comme étant « imprimé », « manuscrit », « tapuscrit » ou « numérique », avec, dans ce dernier cas, deux sous-catégories, celle des textes « nativement numériques » et celle des textes « numérisés » (ou obtenus par rétronumérisation). La nécessité de prendre en compte le médium dans la description typologique est généralement absente des théorisations ; elle figure en revanche dans l'approche pragmatique de Lee, qui propose, sous l'aspect *medium*, des catégories similaires à celles que nous avons adoptées.

Une certaine superposition existe entre cet aspect « support » et d'autres métadonnées recommandées pour les éditions numériques (par exemple, en TEI, dans le pavé <sourceDesc> du *header* d'un document). Il ne s'agit toutefois pas d'une véritable redondance, puisque l'information concernant l'appartenance à l'une ou l'autre de ces catégories reste le plus souvent implicite dans les informations bibliographiques à propos de la source. Par ailleurs, l'éventuel redoublement de l'information a l'avantage d'augmenter la « trouvabilité » du texte, un aspect sur lequel les éditions numériques sont actuellement, et assez paradoxalement, plutôt en déficit lorsqu'on les compare avec les ouvrages imprimés accueillis dans les bibliothèques.

9° Le « Type de discours »

Cet aspect permet de caractériser les textes dans une perspective illocutionnaire (Searle). Nous reprenons ici les grandes étiquettes proposées respectivement par E. Werlich et J.-M. Adam (descriptif, narratif, expositif, argumentatif, instructif…), mais nous leur en avons ajouté d'autres, qui les complètent ou les nuancent (prédictif, satirique, normatif, etc.). Cette classe correspond également, plus ou moins, à celle que Lee appelle *language*, et sous laquelle il place des caractéristiques liées au « registre ».

Selon la présentation schématique donnée dans la figure 1, les descripteurs à proprement parler devraient se trouver au niveau III de notre thésaurus. En pratique, on trouvera déjà certaines de ces étiquettes au niveau II : ainsi, tandis qu'il faut ouvrir « fictif » pour accéder aux étiquettes que le terme recouvre (*fantasy*, réaliste,

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

science-fiction, utopie/ dystopie ou autre), « non-fictif » peut être utilisé directement en tant que terme-clé. À l'inverse, dans quelques cas, les concepts peuvent se trouver à un niveau IV de profondeur, c'est-à-dire contenus dans des sous-groupes, constituant eux-mêmes une sous-division d'un sous-groupe de niveau supérieur, comme c'est par exemple le cas de « minute notariale » :

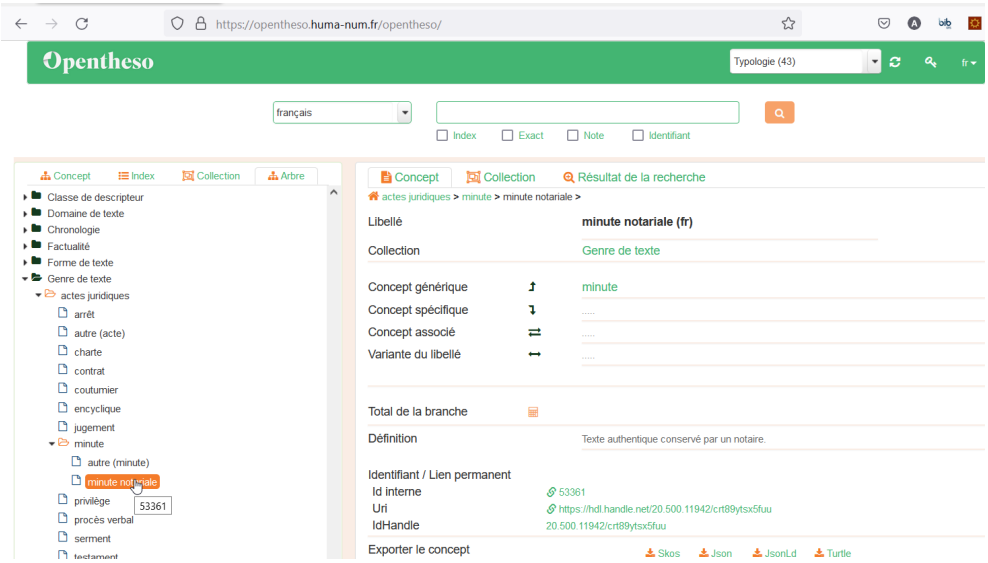


Fig. 2 : Place de « Minute notariale » dans le thésaurus « Typologies textuelles »

Toutefois, notre volonté a été de limiter au maximum les emboîtements, afin d'assurer la lisibilité de l'ensemble et de faciliter l'identification des descripteurs appropriés. L'arborescence est donc le résultat d'une recherche d'équilibre entre richesse et profondeur, d'un arbitrage entre une vision « à plat », plus spécifique des dictionnaires et plus appropriée aux fongibilités génériques, et une vision verticale, organisatrice, qui entend aller du général au particulier et du tout à la partie. En outre, comme dans toute ontologie, nous avons tenté de limiter le nombre de concepts proposés, en indiquant d'une part des termes à privilégier (« sirventès » plutôt que « sirvente », par exemple, « devinette » plutôt que « sirandane[21] »), et en misant d'autre part sur la combinatoire pour atteindre la précision descriptive.

Parmi les notices consacrées aux étiquettes, la plupart offre, pour l'instant, un socle minimal de renseignements, comme la classe à laquelle les concepts se rattachent, le concept générique dont il constitue une déclinaison, et, dans tous les cas, une brève définition.

[21] À la fois « sirvente » et « sirandane » figurent cependant dans notre thésaurus en tant que synonymes.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

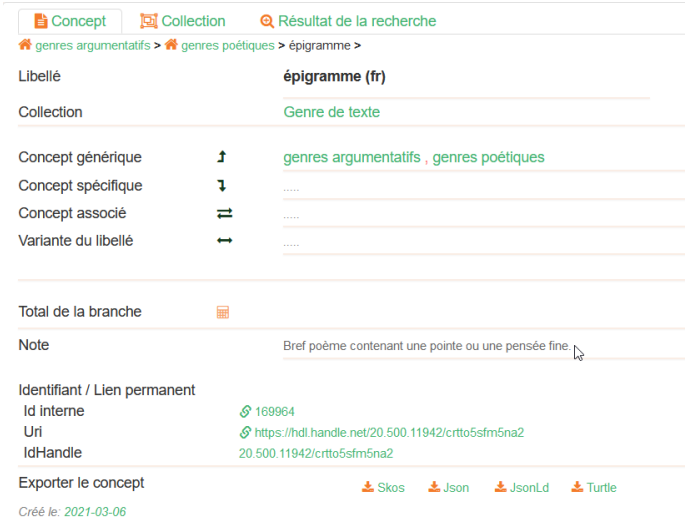


Fig. 3 : La notice « épigramme »

D'autres étiquettes sont plus élaborées et proposent des concepts associés, des rattachements chronologiques, des synonymes, des précisions éditoriales, voire des traductions. Nous reproduisons ci-dessous la notice consacrée à « chant », qui permet de montrer comment, par l'intermédiaire d'un concept associé et d'une note, le domaine d'application de l'étiquette se précise, notamment par rapport au terme concurrent de « chanson ».

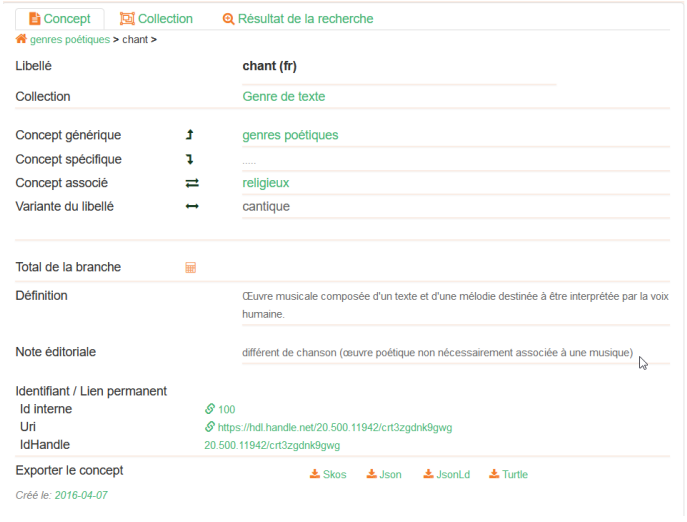


Fig. 4 : La notice « chant »

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

Enfin, une autre possibilité que nous avons utilisée pour préciser l'incidence des étiquettes est l'association avec d'autres concepts, définis sous l'une des classes de descripteurs mentionnés plus haut, comme on peut le voir dans la notice « chronique », reliée à « historique » (domaine de texte) et à « non-fictif » (factualité).



Fig. 5 : La notice « chronique »

3 Décrire les types de textes : quelques exemples d'utilisation

Avec 366 concepts documentés, le thésaurus « Typologies textuelles » est prêt à l'emploi, quoique nous ayons déjà prévu plusieurs évolutions et améliorations. Si une présentation détaillée, en vue de son utilisation, a également été proposée dans un guide pratique qui a été mis à la disposition de la communauté scientifique sur HAL[22], il nous paraît indispensable de rappeler ici les principales recommandations que nous entendons faire à la communauté des éditeurs numériques.

L'idée est d'inciter cette communauté à inscrire dans les métadonnées de l'édition numérique au moins un mot-clé pour quatre des paramètres mentionnés plus haut : domaine, factualité, forme et genre. Les cinq autres classes sont optionnelles, quoique fortement recommandées. Par ailleurs, il peut y avoir plusieurs mots clés de domaine, de forme et de genre, afin de décrire les textes complexes ou composites. Enfin, quoique les catégories de triage puissent, d'un point de vue technique, être utilisées comme descripteurs, il est préférable de chercher des concepts spécifiques, situés à des niveaux inférieurs de profondeur. Si aucun terme spécifique ne convient pour caractériser un texte qui relève pourtant d'une catégorie de triage, le concept « autre » est disponible pour chacune de ces catégories génériques (par exemple, « autre (genres argumentatifs) »).

[22] Galleron, Ioana, Idmhand, Fatiha, Lavrentiev, Alexei, Demonet, Marie-Luce, Réach-Ngô, Anne. (2021). *Décrire les textes dans le cadre d'une édition numérique : le thésaurus "Typologie textuelle" du Consortium CAHIER*, halshs-03402679 , https://halshs.archives-ouvertes.fr/halshs-03402679.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

En suivant ces principes, la pièce de théâtre *Le Préjugé à la mode* de Nivelle de La Chaussée, déjà citée plus haut, pourrait être décrite ainsi dans un format XML/TEI :

```
<teiHeader>
        <fileDesc>
                <titleStmt>
                        <title>Le Préjugé à la mode</title>
                        <author>La Chaussée, Pierre Claude Nivelle de (1692-1736)</author>
                </titleStmt>
                […]
        </fileDesc>
        […]
        <profileDesc>
                <textClass>
                        <keywords scheme='https://opentheso.huma-
                num.fr/opentheso/api/theso/Typologie'>
                                <term type='domaine'>littérature</term>
                                <term type='factualite'>fiction</term>
                                <term type='forme'>vers</term>
                                <term type='genre'>drame bourgeois</term>
                                <term type='agencement'>livre</term>
                                <term type='origine'>oeuvre originale</term>
                                <term type='discours'>dramatique</term>
                                <term type='public'>non défini (public cible)</term>
                                <term type='mode_inscription'>imprimé</term>
                        </keywords>
                </textClass>
        </profileDesc>
</teiHeader>
```

À noter qu'il s'agit là d'une présentation simplifiée, pour améliorer la lisibilité. Dans les faits, l'éditeur du texte numérique est invité à ajouter une référence à l'URI de chaque concept utilisé, de la façon suivante :

```
<term type='factualite'ref='https://hdl.handle.net/20.500.11942/crt22aacdhvi7'>
fiction
</term>
```

Bien entendu, l'utilisation de cette série de termes n'est pas exclusive du recours à d'autres, plus adaptés à la description des thématiques du texte, ou d'autres aspects esthétiques ou stylistiques que l'éditeur souhaite mettre en avant. Comme nous l'avons indiqué dans le préambule, cette proposition vise à améliorer l'identification de textes similaires à travers des corpus numériques divers, un objectif que les mots-clés actuellement utilisés dans les études littéraires ne permettent pas d'atteindre, car ils sont fondés sur des conceptualisations et des pratiques trop disparates.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352

4 Conclusions

La typologie proposée par CAHIER n'est ni définitive, ni exhaustive, et elle ne prétend pas répondre sur le fond à la question de la classification des textes. En prenant acte de l'existence d'exercices de classification des textes et de leur portée opératoire, elle fournit la « traduction » numérique d'un vocabulaire relativement stabilisé, défini de façon aussi consensuelle et brève que possible. Son utilisation peut reposer sur une analyse préalable lexicale et/ou morpho-syntactique des textes, manuelle ou assistée par ordinateur, mais ne l'exige pas, et les traits définitoires fournis pour chaque notion ne prétendent pas refléter (absolument) de tels critères formels. De ce fait, la référence à la typologie CAHIER pour la constitution des métadonnées d'une publication numérique continue à engager la responsabilité de l'annotateur en ce qui concerne la pertinence des étiquettes utilisées. L'organisation des termes dans le cadre d'un thésaurus hiérarchisé, la proposition de catégories de nature bien plus diverses que dans les exercices précédents de classification, et la souplesse de leur combinaison se veulent ainsi une aide au choix de descripteurs les plus appropriés possible.

Afin de vérifier la robustesse de ce thésaurus, des tests ont été réalisés au sein de la communauté CAHIER. Ils ont permis d'identifier des termes manquants, de corriger des définitions mal formulées et, surtout, de compléter et finaliser le guide d'utilisation (Galleron *et al.*, 2021) à partir des hésitations et des incompréhensions manifestées par ces premiers annotateurs. Se pose maintenant la question de la création d'interfaces permettant de requêter de grands volumes de textes numérisés et équipés de la série de mots-clés que nous recommandons. C'est une piste que le groupe de travail veut explorer dans les prochaines années.

Enfin, force est de constater que la construction d'un thésaurus de termes littéraires n'est pas une activité que les spécialistes de ce domaine envisagent naturellement comme constituant de la recherche. Cependant, ne pas s'impliquer, en tant que communauté, dans la mise au point de tels répertoires, c'est passer à côté d'un des enjeux les plus actuels des humanités numériques, et de la construction du Web sémantique en général. L'équipement de nos éditions électroniques avec des mots-clés issus de la riche littérature sur les types et les genres mais bénéficiant de la visibilité et du potentiel de mise en lien des objets numériques, est de nature à stimuler de nouvelles recherches et de nouvelles perspectives sur le patrimoine écrit. Grâce à notre proposition de thésaurus, nous espérons contribuer à ce renouvellement des études littéraires, philologiques, linguistiques et historiques.

Published in : IDMHAND, Fatiha ; GALLERON, Ioana (dir.), *Dix ans de corpus d'auteurs*, Editions des archives contemporaines, ISBN : 9782813004352, doi : 10.17184/eac.9782813004352