Présentation des descripteurs du projet CORPTEF

Coordonnée par

Céline Guillot (Celine.Guillot@ens-lsh.fr)

Alexei Lavrentiev (Alexei.Lavrentev@ens-lsh.fr)

CNRS / Université de Lyon (ENS-LSH), UMR 5191 ICAR

Document élaboré dans le cadre du projet CoRPTeF financé par l'ANR

VERSION 4.2

Table de mise à jour du document

- 1er juillet 2008, rédaction de la 1ère version
- Septembre 2008 : 2nde version (ajouts de Anne Carlier)
- Octobre 2008 : 3ème version (ajouts de AL et CG)
- Mars 2009 (ajouts de CG)
- 23 mars 2009 : mise à jour suite à la réunion du projet

Principes généraux

1. Définition des descripteurs

Un descripteur correspond à une dimension de description d'une unité textuelle (et non, pour l'instant du moins, d'une partie d'une unité textuelle).
La définition de l'unité textuelle sur laquelle porte le descripteur repose sur la distinction de trois objets différents :

- l'œuvre (ou l'acte) ;
- l'édition ;
- le manuscrit.

L'unité textuelle correspond à une œuvre représentée par une édition (ou éventuellement par un manuscrit) et qui est numérisée afin d'être intégrée au corpus.
Chaque œuvre et chaque édition sont dotées d'un identifiant unique (sigle DEAF entier pour une édition ; sigle DEAF moins l'initiale de l'éditeur pour une œuvre).
Un descripteur d'une unité textuelle correspond donc à une propriété que cette unité hérite soit de l'œuvre (auteur, date de composition, genre, etc.), soit du manuscrit, soit de l'édition (éditeur scientifique, etc.).
L'ensemble des descripteurs rassemble les métadonnées disponibles et/ou utiles pour l'interprétation et l'exploitation des textes.
Les descripteurs sont définis *a priori* sur les textes.
A chaque descripteur correspond une liste de valeurs[1]. Ces valeurs peuvent être libres ou contrôlées par une liste d'autorité.

2. Fonctions des descripteurs

On distingue au moins trois fonctions distinctes :
▪ fonction documentaire : pouvoir retrouver un document répondant à un certain critère de description ; se faire une idée plus ou moins claire du contenu du document ; permettre les échanges de documents avec nos partenaires ;
▪ fonction méthodologique : à la fois organiser et enrichir le corpus en évaluant de quoi il est représentatif, dans un souci de diversification maximale ; pouvoir "partitionner" le corpus en sous-corpus selon plusieurs critères explicites et parfois combinés ;
▪ fonction scientifique : mener à partir du corpus des recherches synchroniques et diachroniques en langue (avec un niveau de généralité le plus grand possible) ; rendre compte de la variation des usages compte-tenu de l'ensemble des variables pertinentes ; décrire et modéliser le plus adéquatement possible les types de textes produits au Moyen Age compte-tenu des variables pertinentes.

[1] Dans la suite du document, on utilisera les majuscules pour référer aux descripteurs et les guillemets pour référer aux valeurs de ces descripteurs.

3. Fonctionnement du système des descripteurs

▪ Tous les textes sont décrits à l'aide des mêmes descripteurs, qui forment donc un système cohérent
▪ Toute la période concernée (IXe-fin XIIe siècle) doit être couverte par la même liste de descripteurs
▪ Dans un souci d'interopérabilité avec les bases de données connexes qui sont exploitées par les mêmes utilisateurs (en premier lieu la BFM), on veillera à maintenir une cohérence maximale avec les descripteurs utilisés pour les périodes suivantes ; cela permettra des recherches sur tel ou tel aspect de la langue médiévale dans son ensemble et son évolution entre le IXe et le XVe siècle (par exemple dans le cadre du projet en cours de *Grande Grammaire historique du français*).

Des trois fonctions principales assignées aux descripteurs découlent deux conséquences apparemment opposées :
▪ Pour un texte donné et pour certains descripteurs[2], on peut vouloir choisir plusieurs valeurs différentes ; par exemple, le *Roman de la Rose* peut être décrit comme un texte à la fois « littéraire » et « didactico-scientifique » (deux valeurs différentes du descripteur DOMAINE), car cela permet de retrouver ou d'inclure ce texte quand on travaille sur les textes littéraires ou didactiques (logique documentaire)
▪ En revanche, quand on utilise des outils statistiques, il est nécessaire de ne comptabiliser les occurrences d'un même texte qu'une seule fois (logique méthodologique).

Deux solutions sont envisageables selon le degré de « mixité » d'un texte donné :

A. Il est possible de définir une valeur principale (ou par défaut). Une ou plusieurs autres valeurs seront considérées comme secondaires et ne seront pas prises en compte dans les analyses statistiques ;

B. Il est impossible de sélectionner une valeur principale d'un descripteur sans fausser l'exactitude du classement. Dans ce cas, toutes les valeurs sont données comme principales, et leur combinaison sera considérée comme une valeur indépendante. Lors des analyses statistiques contrastives les textes de ce type seront soit exclus, soit considérés comme une catégorie à part.

▪ On distingue les descripteurs qui sont obligatoirement renseignés, de ceux qui sont facultatifs (cf. tableau en annexe) ; dans certains cas, quand il est impossible d'attribuer une valeur précise à une variable, on peut ne pas la renseigner (par exemple, on ne connaît pas la région de production du ms). Les descripteurs obligatoires sont ceux qui sont *a priori* les plus pertinents pour l'exploitation des textes et qui peuvent donner lieu à la sélection de sous-corpus ciblés.

▪ On distingue les descripteurs « universaux », applicables normalement à n'importe quelle unité textuelle et les descripteurs conditionnels, dont la pertinence dépend de la valeur d'un autre descripteur (par exemple, le THEME est pertinent iniquement pour le DOMAINE « didactico-scientifique »).

---

[2] Ou pour tous les descripteurs ? Il peut y avoir des hésitations et plusieurs hypothèses sur autre chose que les domaines et les genres, par exemple sur l'identité d'un auteur, sur la date de composition d'une œuvre, etc.

4. Liste des descripteurs (tableau en annexe)

On peut regrouper les descripteurs par grandes catégories.

A. Descripteurs liés aux dates

On regroupe dans cette catégorie des descripteurs dont les valeurs sont libres (parce que souvent approximatives) et des descripteurs dont les valeurs sont fixes (pour les tris informatiques)

Descripteurs à valeur libre :
DATE DE COMPOSITION DE L'ŒUVRE : cette date peut être très approximative[3].
DATE DE COMPOSITION DU MANUSCRIT (ms de base pour les éditions critiques) : idem ; on veillera à ce que l'écart entre la date de composition supposée de l'œuvre et la date du manuscrit utilisé soit le plus réduit possible (cf. principes de constitution du corpus).

Descripteurs à valeur fixe qui permettent de définir des intervalles de temps pour des tris informatiques :
DATE DE DEBUT DE COMPOSITION (ŒUVRE ou MANUSCRIT)
DATE DE FIN DE COMPOSITION (ŒUVRE ou MANUSCRIT)
DATE DE COMPOSITION FIXE (ŒUVRE ou MANUSCRIT) (moyenne entre la date de début et la date de fin)

B. Descripteurs liés à la dimension spatiale/dialectale

DIALECTE DE L'AUTEUR (= traits régionaux attribuables à…)
DIALECTE DU COPISTE (= traits régionaux attribuables à…)
REGION DE PROVENANCE DU MS

Le dialecte du copiste doit être distingué de la région de production du ms, la circulation des personnes ayant toujours été importante au Moyen Age (cf. Busby, 2002).
Par défaut, on ne renseigne que le dialecte du copiste à partir des informations présentes dans le DEAF.

C. Descripteur lié à l'auteur, au scribe, à la chancellerie

Ces informations seront consignées dans un descripteur nommé AUTEUR / REDACTEUR.

---

[3] En cas de datation approximative, la date est fixée au milieu de la période dans laquelle le texte parait avoir été composé. Une autre solution a été adoptée dans la base textuelle élaborée pour le *Dictionnaire de moyen français* du laboratoire ATILF (http://atilf.atilf.fr/dmf.htm) où c'est la date la plus ancienne possible de la composition du texte qui est retenue. Il semble toutefois que le choix de la date « médiane » soit partagé par la majorité des membres du Consortium pour les Corpus de Français Médiéval (CCFM, http://ccfm.ens-lsh.fr)

L'identité de l'auteur/rédacteur pourra selon les cas correspondre à une personne ou à une institution (dans le cas des chartes par exemple), et on distinguera par ailleurs au moins deux rôles possibles (auteur/scribe). Des informations relatives à l'âge et à la catégorie sociale seront données dans les rares cas où elles sont disponibles et ces informations ne concerneront que les personnes.

D. Descripteurs liés à la forme du texte

FORME (valeurs « prose », « vers », « mixte », « glose »)
TYPE DE VERS (« octosyllabes », « décasyllabes », etc.)
STRUCTURE INTERNE (division en livres, parties, chapitres, etc.)

E. Descripteurs DOMAINE

Définition du DOMAINE : trait fonctionnel qui correspond à la destination principale du texte et au domaine d'activité auquel il se rattache :

- divertir → « littéraire »
- enseigner, instruire (ce qui concerne le savoir et sa transmission) → « didactico-scientifique »
- édifier (ce qui concerne le rituel et la diffusion du message chrétien) → « religieux »
- consigner/relater les événements du passé → « historique »
- réguler la vie sociale → « juridique »
- régler des questions pratiques (actes de la pratique) → « source documentaire »

Liste des valeurs du descripteur DOMAINE

- Littéraire
- Didactico-scientifique[4]
- Religieux
- Historique[5]
- Juridique
- Source documentaire (acte de la pratique)

Utilisation du descripteur DOMAINE :
On sait qu'il est difficile, parfois impossible, de proposer un domaine unique pour un texte.
On propose de distinguer deux cas de figure :

- un texte relève plutôt d'un domaine mais a aussi des affinités avec un ou plusieurs autre(s) : on distingue un domaine principal et un ou plusieurs domaines secondaire(s) ; la conséquence est que si l'on veut établir des contrastes/comparaisons

---

[4] En ce qui concerne la partition entre « scientifique » et « didactique » ou leur association à l'intérieur d'un seul DOMAINE, nous proposons dans un premier temps de réunir les deux et de voir dans une seconde phase si au vu de la masse des données il est pertinent ou non de scinder les deux valeurs.

[5] En ce qui concerne la frontière entre DOMAINES « historique » et « historico-politique », nous proposons d'adopter la solution de Serge Lusignan : on conserve la valeur « historique » et on verse dans le « didactico-scientifique » les textes plus politiques. Cela suppose d'associer au DOMAINE « didactico-scientifique » un THEME dont la valeur est « politique ».

entre textes de plusieurs domaines, ce texte sera classé à l'intérieur de son domaine principal

- un texte ne peut vraiment pas être classé dans un seul domaine principal : on admet qu'il ait deux ou plusieurs domaines principaux ; la conséquence est que lors des opérations de contrastes, ce texte sera dans une catégorie à part ; pour cette raison cette solution doit être limitée aux cas les plus extrêmes.

F. Descripteur GENRE

Trait qui correspond à certaines propriétés formelles internes au texte, mais qui sont difficiles à définir précisément ; au bout du compte, on s'accorde en général pour voir dans les genres textuels des catégories « intuitivement reconnues » et qui sont le produit d'une sorte de consensus à une époque donnée (Lee, Jauss pour le Moyen Age, etc.) ; catégorie dont les limites sont évidentes mais dont on peut en même temps difficilement se passer.
Il s'agit d'un descripteur dont la liste des valeurs est ouverte dans l'état actuel des connaissances.

Trois taxonomies au moins sont possibles :

- les GENRES tels qu'on pense que les médiévaux les voyaient ;
- les GENRES médiévaux tels qu'on les voit aujourd'hui (du point de vue de la tradition philologique) ;
- les GENRES tels qu'on les définit dans la société actuelle.

Pour l'instant on s'est surtout concentré sur la seconde.

Le descripteur GENRE est d'une certaine façon subordonné au descripteur DOMAINE, mais un même GENRE peut se rencontrer dans différents DOMAINES (par exemple le GENRE « dramatique » se rencontre dans les DOMAINES « religieux » et « littéraire »).

Le DOMAINE « source documentaire » comporte une série de valeurs de GENRE qui lui sont propres

Sources documentaires	Charte-lettre
	Registre[6]
	Compte
	Censier-cadastre

Il existe une certaine affinité entre le descripteur GENRE et l'opposition entre textes narratifs ; argumentatifs, expositifs et autres (lyrique, etc.). On proposera une mise en correspondance entre les genres et cet aspect de la typologie textuelle.

G. Descripteur THEME

Le descripteur THEME est a priori utilisé uniquement pour les textes qui relèvent du DOMAINE « didactico-scientifique » et il n'est pas obligatoire pour ces textes[7].

---

[6] Registre de délibérations, de sentences, de plaidoiries etc.

Liste des valeurs du descripteur THEME

- Nature : bestiaire, lapidaire, plantaire
- Géographie
- Trivium : grammaire, logique, rhétorique[8]
- Quadrivium : science des nombres et comput
- Médecine
- Droit
- Théologie
- Pastorale
- Encyclopédie
- Politique
- Vie quotidienne

H. Descripteur RELATION

Ce descripteur permet d'indiquer si le texte entretient une relation avec un autre texte.
Les valeurs possibles sont :

Descripteur RELATION	Valeurs
	« traduction »
	« adaptation »
	« remaniement »
	« commentaire »
	« inapplicable »[9]
	« indéterminé »[10]
	« mise en prose » / « mise en vers »

On accepte plusieurs valeurs pour un même texte. Par exemple, les *Quatre livres de rois* sont à la fois une traduciton du latin et une mise en prose d'une version plus ancienne en vers.

I. Descripteur SOURCE

Ce descripteur identifie la ou les œuvres ou autorités à la source du texte.
Les valeurs possibles sont :

Descripteur SOURCE	Valeurs
	« Bible »
	« auctoritas »[11]

---

[7] Nous proposons finalement de conserver le descripteur THEME et non DISCIPLINE SCIENTIFIQUE, car cela permet de caractériser par exemple une œuvre comme le *Roman de la rose* avec un THEME « encyclopédie ». Cette information trouverait mal sa place dans une catégorie DISCIPLINE SCIENTIFIQUE et elle correspond bien à une réalité médiévale.

[8] On indiquera par exemple pour un ouvrage grammatical : « trivium:grammaire ».

[9] Si le texte est original.

[10] Si la relation est incertaine

[11] Cette valeur est utilisée pour les traductions de textes non bibliques.

	« autre »
	« original »

Un champ supplémentaire est utilisé pour indiquer chaque fois que cela est possible le titre du ou des textes source.

Annexe 1
Tableau qui donne la liste des descripteurs avec valeur fixe / contrôlée, obligatoire / facultatif et exemples de valeurs pour chaque genre (listes d'autorité en plus pour les valeurs contrôlées).

Tableau 1. Liste des descripteurs

Descripteur	**Valeur contrôlée (C) ou libre (L)**	**Valeurs**	**Obligatoire (O) ou facultatif (F)**
(A1) Date de composition	C et L	[sous-siècle] (« début 13e », « milieu 13e », etc.)	O
(A2) Date du manuscrit	C	[sous-siècle], [siècle]	O
(B1) Dialecte de l'auteur	C	[nom du dialecte] (si identifiable : « picard », « normand » ou « bourguignon »)	O
		« traits de [nom du dialecte] »	
		« non déterminé »	
(B2) Dialecte	C	idem	F
(B3) Région du manuscrit	C	[nom de la région]	F
		« inconnu »	
(C1) Auteur	L	[nom] [âge] [catégorie]	O et F
(C2) Scribe	L	[nom]	F
(D) Forme	C	« vers », « prose » ou « mixte »	O
(E) Domaine	C	« littéraire », « religieux », « didactico-scientifique », « historique » « juridique » « source documentaire »	O
(F) Genre	C (liste ouverte)	« roman », « épique », etc.	O
(G) Thème	C	« nature », « géographie », « trivium », etc.	O (pour le domaine didactico-scientifique) F (pour les autres domaines)
(H) Relation	C	« traduction », « remaniement », « commentaire », « inapplicable », « indéterminé »	O
(I) Origine	L	« Bible », « auctoritas », « autre », « original ».	O (qd il y a une relation explicite avec un autre texte ou un auteur)
		[Nom du texte source]	F