CHAPITRE V

Constitution et codage du corpus

Aperçu

La constitution des corpus est commentée : il s'agit d'une part d'établir de quoi construire la base des profils à partir de textes électroniques existants (repères méthodologiques et critères de choix), d'autre part de cerner la nature effective des documents soumis.

Ceci fait, se pose la question du format de codage, qui doit être à la fois robuste et général, mais qui doit aussi capter des structurations significatives. Un modèle (une DTD SGML, appelée *Corpus*) est construit et mis en œuvre, après avoir tiré des enseignements de l'examen approfondi des standards importants en matière de données textuelles : recommandations de la TEI, norme ISO pour la documentation électronique, modèle pour l'échange de données dans le cadre d'un projet européen de TAL (GRAALDOC). Une attention particulière est portée sur ce qui est porteur de sémantique dans les formats de structuration. Le modèle *Corpus* exprime huit formes élémentaires d'organisation du texte, qui jouent un rôle interprétatif en guidant la lecture. Ces huit formes élémentaires s'articulent pour donner une représentation de l'architecture interne du texte et des relations intertextuelles.

Des modules informatiques ont été développés pour traduire dans la DTD *Corpus* les formats d'entrée les plus concernés : du texte ASCII éventuellement mis en forme, des données SGML et en particulier des pages Web HTML tout-venant, non strictement conforme à la syntaxe SGML.

Table des matières du Chapitre V

A. LES DOCUMENTS CONSIDÉRÉS

1. Dissymétrie profil / requête

L'application DECID requiert la définition d'une base de profils, et ces profils sont construits à partir de textes. Au niveau applicatif, l'ensemble des profils forment une *base* ; au niveau de la représentation, les textes forment un *corpus*. Les propriétés attendues chez l'un doivent se retrouver chez l'autre : à l'exhaustivité, la régularité, requises pour la base, répondent la cohésion et l'équilibre du corpus. Le choix des documents pour la représentation des profils se présente comme la construction raisonnée d'un corpus, aux soins du concepteur puis de l'exploitant de DECID.

A l'inverse, les documents qui seront utilisés pour interroger DECID ne sont pas fixés, et ils sont perçus indépendamment les uns des autres. Leur réunion reflète le faisceau des pratiques des utilisateurs de DECID.

2. Les profils

a) Tirer parti des textes

L'existant suffit

La force de l'application DECID réside dans sa capacité à dériver des textes, automatiquement, une information de caractérisation des destinataires. Tout un organisme peut ainsi être décrit, pour peu que l'on dispose d'un corpus représentatif de l'ensemble.

L'automatisation de la chaîne de traitement suppose que l'on n'impose pas comme format d'entrée une structuration particulière, qui pour la plupart des corpus devrait être ajoutée manuellement ou même semi-automatiquement. En effet, le balisage d'un corpus est une charge de travail considérable. Il s'agit donc de pouvoir exploiter automatiquement des informations de structuration existantes, comme de pouvoir se passer d'informations de structuration explicite quand le texte d'entrée est démuni de tout balisage.

Pas n'importe quel existant : le discernement critique

Il faut souligner l'importance de recueillir et de se fonder sur des données de qualité, tant en ce qui concerne la propreté des fichiers que sur l'intérêt des informations qui sont représentées. Si les données pèchent par de nombreuses irrégularités ou lacunes, les traitements automatiques ne peuvent opérer que des caractérisations confuses et des rapprochements décevants. Si derrière les profils, il n'y a pas des données de valeur, nul n'a envie d'utiliser le système. Faute de se baser sur des textes de qualité, l'application est alors doublement inutile –inutilisable et inutilisée.

L'application du principe de réalité

Dans le cadre de la constitution des profils, le principe de réalité exige de savoir prendre en compte un texte tout-venant, sans lui supposer impérativement le respect d'une grammaire canonique, la limitation à un vocabulaire de référence, la présentation normalisée, ou l'organisation stricte selon un plan type. Les illustrations de corpus qui se conforment à l'une ou l'autre de ces exigences se réduisent à des usages rigides, pesants et fermés par un carcan doctrinal et formel. Se fonder sur les textes n'a de sens qu'en envisageant l'expression vivante et libre, qui est leur nature et fait leur richesse.

b) Dimensions d'un profil et rôle d'un texte

Le profil d'une personne dans DECID a deux faces : la personne peut être recherchée soit comme destinataire, soit comme source d'information. Dans le premier cas, on parle du *profil d'intérêts* de la personne, et dans le second de son *profil de compétences*.

Les documents sous-jacents, pouvant servir à la caractérisation, alimentent l'un ou l'autre de ces profils. Il n'est pas anodin de remarquer que les documents rédigés par une personne ne ressemblent qu'accessoirement aux documents dont elle se sert dans son travail ; ou que les documents cités dans une bibliographie sont toujours en léger décalage par rapport à l'ouvrage ou l'article qui les cite. L'information qui retient l'attention est celle que l'on n'aurait pas su retranscrire avant de la rencontrer. Cette forme d'hystérésis se retrouve dans les deux faces du profil. Le profil d'intérêt correspond aux documents consultés, étudiés, acquis, dans le cadre d'un projet actuel. Le profil de compétence ressort des compte-rendus, des rapports produits, des documents reçus pour relecture et avis ; il reflète souvent l'expérience acquise lors de projets relativement récents et menés à bien.

La teneur du profil se déplace donc, selon que l'on considère des documents écrits ou bien lus par la personne. Et un même document peut, avec le temps, glisser d'un profil d'intérêts à un profil de compétences. Enfin, l'organisation du travail fait que certains documents qui définissent ou reflètent l'activité d'une personne ne sont pas nécessairement écrits par elle : programme de travail établi par un supérieur hiérarchique, document de synthèse réalisé par un proche collaborateur.

> Dans plusieurs cas, le nom de la personne n'apparaît pas sur le document qu'elle présente comme caractéristique de son activité. (Merle, Fradin, Soinard 1994, p. 68)

Ou inversement

> le texte parle des activités [de l'auteur], mais on n'y perçoit pas le rôle de l'auteur dans l'activité décrite (Merle, Fradin, Soinard 1994, p. 69)

En bref, la relation de l'auteur à son texte n'est pas la seule possible pour construire un profil ; et cette relation se charge d'interprétations différentes notamment au fil du temps.

Une seconde dimension des profils peut également séparer deux familles de textes. Dans le cadre de l'application DECID, le texte est le médiateur entre l'utilisateur et le système. Il sert à la construction de représentations internes pour le calcul, qui caractérisent les destinataires et les documents à envoyer. Mais le texte est également ce qui peut être présenté à l'utilisateur comme explication d'un rapprochement. Une contrainte de confidentialité par exemple peut justifier de recourir à des textes très détaillés pour une caractérisation fine des activités des personnes, tout en ne donnant accès, comme explication, qu'à des descriptions plus évasives et moins explicites. Une autre raison, tout aussi réaliste, d'écart entre le *texte constitutif du profil* et le *texte de présentation*, se rencontre dans le cas d'une information accessible sur un réseau et en continuelle évolution. La base des profils est calculée périodiquement et se fonde sur un état des textes enregistré (donc qui peut être attesté) et synchrone. L'utilisateur peut trouver plus représentative, ou complémentaire, l'information sur la personne telle qu'elle se présente, dans son état le plus actuel.

c) Recensement des documents électroniques existants à l'échelle de la Direction

Pour avoir une description des profils sans mobiliser les agents, les textes utilisés doivent faire partie des documents courants, donner un écho de l'activité des agents, et faire l'objet d'une collecte centralisée sous forme électronique.

Dans le cadre de la Direction des Etudes et Recherches, une dizaine de familles de documents peuvent être envisagés :

- *les ARD/AID* : ce sont des textes d'environ deux pages, rédigés chaque année par les chercheurs à l'attention de leur hiérarchie, décrivant l'avancement et les objectifs annuels d'un sujet de recherche (appelé *action*). Ils servent à l'ordonnancement de l'ensemble des activités de la Direction des Etudes et Recherches, c'est-à-dire à la détermination des orientations, de la répartition, du calendrier et des moyens de la recherche. Il y a en fait deux types de documents : les Fiches sont établies en automne et explicitent le programme de recherche pour l'année suivante ; les Comptes Rendus, émis au printemps, font le point des résultats acquis et des difficultés rencontrées l'année passée.
- *La nomenclature SDG* : structuration et noms des 180 Groupes, 35 Départements et 8 Services. Les libellés des structures sont une première indication de leur domaine d'activité.

- *Les Notes Internes* : ce sont les publications internes. Elles sont extrêmement variées : compte-rendu de mission, document de travail technique, rapport de stage, document de doctrine, thèse, article ou communication scientifique, norme, rapport d'essais, etc. Un champ Type de document, renseigné lors du catalogage, doit mentionner l'une des valeurs suivantes : *livre*, *publication*, *rapport*, *congrès*, *brevet*, *thèse*, *norme*.
- *La Collection des Notes DER* : c'est une reprise d'une partie des Notes Internes, dans l'optique d'une diffusion externe. Les critères de sélection pour l'entrée dans la Collection ne sont pas clairs ; une chose est sûre, c'est que la Collection ne comprend que des Notes dont l'accessibilité est libre (aucune restriction de confidentialité).
- *Les contrats et bilans de PPRD* : ce sigle désigne les Projets Pluri-annuels de Recherche et Développement. Ceux-ci coordonnent plusieurs actions. Depuis 1997, les PPRD ont été remplacés par une organisation en *projets*.
- *Les contrats de Groupe, avenants et bilans* : ces documents présentent le programme d'un Groupe pour trois années.
- *Les CERD* : ce sont les Contrats Externes de Recherche et Développement régissant les partenariats. La partie contractuelle est très formelle et peu informative sur le contenu des travaux à mener ; ces informations sont plutôt dans l'annexe technique. On ne dispose pas de celle-ci sous forme électronique.
- *Epure* : revue trimestrielle de vulgarisation technique, illustrant de bons travaux réalisés à la DER. Un numéro regroupe trois ou quatre articles.
- *Faits Marquants* : document annuel, destiné à véhiculer l'image de marque de la DER, présentant une sélection de présentations de quelques recherches dans chaque Service. Ce document a été porté sous forme électronique pour les années 1986 à 1992.

Ce rapide inventaire donne l'impression d'une manne, dans laquelle il n'y a qu'à piocher. Mais d'autres critères interviennent.

d) Critères d'adéquation à l'application DECID

Parmi les données enregistrées sous forme électronique, DECID a besoin de :

- *Une indication de date (année)* : les profils sont repérés dans le temps, et chaque année forme une base de profils. Le profil actuel d'une personne traduit ses centres d'intérêt alors que ses profils des années passées reflètent plutôt ses compétences acquises et son expérience.
- *Le statut éventuel du document* : il s'agit de faire la part entre d'une part les documents provisoires, projets en discussion, et d'autre part ce qui reflète une activité effective.
- *Des données textuelles suffisamment riches* : un titre ou un libellé seul est insuffisant ; un résumé (résumé descriptif) est souvent trop synthétique et superficiel ; une synthèse (résumé informatif) ou / et le texte intégral sont ce qui convient le mieux. L'idéal est de disposer d'un texte structuré.
- *Une indication permettant le rattachement* du document à une personne ou à une structure (Groupe, etc.) : l'auteur n'est pas toujours identifiable. Quand c'est possible, il faut le reconnaître par son nom, ce qui rend problématique l'exploitation automatique, notamment s'il y a plusieurs personnes (il faut isoler chaque nom), s'il existe des homonymes ou au contraire si la personne change d'état-civil (passage du nom de jeune fille au nom d'épouse), si le nom se prête à des confusions entre nom et prénom, à des variantes d'orthographe ou de saisie (parce que composé, ou d'origine étrangère, par exemple). Ainsi, pour les ARD/AID, Véronique Jolly[1] a mis au point un outil qui permet l'identification du responsable à partir de son nom : la proportion d'échec est de 0,8 à 1,5 %[2].

---

[1] Véronique JOLLY est ingénieur au Département SID de la Direction des Etudes et Recherches d'EDF.

[2] Les inconnus restants proviennent pour l'essentiel :
de différences sur le nom ou le prénom entre la base et le texte : `C. KIENY` pour `KIENY JEAN CHRISTOPHE`, `KHIN YEDID C.` pour `YEDID CHHARY` ;
de noms que l'algorithme ne peut découper correctement : `C. CHAN HEW WAI` pour `CHAN HEW CAI CHAN WOHOU`, `NGUYEN NQ THUY` pour `NGUYEN NGOC QUOC THUY` ;
de noms multiples : `MM. Salomon-Nanackere` ;
de l'absence de nom [champ non renseigné] : `X`, `« chef de groupe... »`, `« remplaçant de... »`.

- *La pérennité de la source* qui fournit la version électronique : la définition des profils doit pouvoir être actualisée périodiquement. On se fonde donc sur des documents dont le circuit prévoit l'enregistrement d'une version électronique. A la DER, la base de données SPHERE centralise les données textuelles concernant l'activité : c'est de fait l'interlocuteur essentiel de DECID pour la fourniture des textes pour les profils. En revanche, des collections mises sous forme électronique dans le cadre de projets n'ont pas bénéficié d'un codage suivi : elles ne pourraient être utilisées que de façon annexe.
- *Une répartition régulière* sur toute la DER : dans l'idéal, chaque entité doit être « équitablement » décrite ; il faut éviter qu'une partie de la DER ait une représentation très détaillés, et qu'une autre n'ait que très peu d'éléments pour la construction de ses profils.
- *Une couverture* aussi *fine* que possible : toutes choses égales par ailleurs, les documents correspondant à l'activité d'une personne ou d'une petite équipe sont plus utiles que ceux faisant la synthèse de l'activité d'un Département. En effet, on peut construire une représentation de l'activité du Département connaissant l'activité de ses membres, alors que l'inverse n'est pas vrai.
- *Un volume suffisant* : cela est évidemment lié aux deux critères précédents. Le nombre de textes donne alors une indication du niveau de détail, si l'on peut considérer que les textes expriment un découpage de l'activité et que chacun aborde une spécialisation particulière. En l'occurrence, plusieurs milliers de textes d'action doivent apporter davantage d'informations que quelques dizaines de textes relatifs aux projets (PPRD).
- *Une dynamique de renouvellement* : les données très liées à un référentiel, comme les noms des équipes, ou les contrats sur lesquels elles se fondent, ont nécessairement une certaine inertie, par opposition à d'autres documents directement liés à l'activité en cours. L'image que doit donner DECID se veut suivre au plus près l'évolution des activités.
- *La fiabilité du codage*, sa fidélité par rapport à la version papier si elle existe : cela est moins crucial pour ce qui est du texte proprement dit (pour lequel une erreur devrait souvent être corrigée par le contexte, ou par d'autres occurrences correctes), que pour les autres renseignements et la structuration du fichier (ne serait-ce que la séparation d'un texte à l'autre).
- *L'accessibilité du document* : DECID protège l'accès aux documents ayant servi à la définition des profils, et prévoit d'autres modes d'explication et de guidage pour une bonne exploitation des résultats sans la visualisation de ces textes. Cependant, l'utilisation des documents les plus confidentiels est d'autant plus délicate qu'on en considère une collection complète ; pour DECID, on préférera donc se baser sur des corpus pas trop sensibles. Il est clair cependant qu'avoir une bonne représentation de l'ensemble des activités actuelles de la DER est une information stratégique, jamais anodine.

D'autres critères, moins déterminants, peuvent cependant faire préférer tel corpus à tel autre :

- *La langue* du document sur lequel est basé le profil détermine la langue des documents qui pourront en être rapprochés. L'essentiel des textes ici sont en français. Les données en anglais sont rares, peu développées (surtout traduction de titres, de résumés, mais pas de textes), et lacunaires. En l'état actuel, elles ne permettent pas la construction de profils, sinon peut-être de profils très généraux et sans doute incomplets au niveau des équipes (Groupes, Départements).
- *Le format* de codage des textes est variable selon les documents. En général, on a soit de l'ASCII (fichier texte standard), soit du SGML (fichier structuré). L'intérêt du codage SGML dépend du modèle que codent les balises et de sa pertinence pour l'application. Les autres facteurs importants sont la propreté du codage (taux d'erreurs de la version électronique) et sa richesse (conservation des accents, du découpage en paragraphes, etc.).
- Encore rarement reprises par une structuration SGML, les *régularités de structure* induites par un plan-type, ou déjà les r*égularités relatives à un genre* bien cerné, donnent accès à un traitement plus fin du document. En ce sens, l'existence de consignes de rédaction (sous la forme de documents de doctrine), qui instituent un cadre de référence, peut guider la modélisation. Bien sûr, cette dernière ne sera pas une retranscription bornée des consignes, car tout le monde n'a pas forcément connaissance du document de doctrine, et chacun se fait une idée, a son interprétation, de la forme qui est demandée. (On observe ainsi l'influence très forte du contexte : les textes

(Données fournies par Véronique JOLLY, le 27 décembre 1996).

d'action d'un même Groupe, d'un même Département, d'un même Service, ont un air de famille.) En revanche, les collections hétérogènes amalgament des documents provenant de contextes d'usage les plus divers, et rendent problématique la construction automatique d'une représentation cohérente et équilibrée.

- Le fait de disposer d'*archives* sur plusieurs années est un plus, car on peut construire immédiatement plusieurs bases de profils (avec la nuance entre profils d'intérêts -actuels- et profils de compétences), et l'on a aussi tout-de-suite des informations sur la stabilité et la nouveauté des activités.

Dans cette liste de critères ne figurent pas la présence d'attributs de *classification* (axe de recherche, thème de recherche,...), car *(i)* les grilles sont relativement peu détaillées, *(ii)* elles ne peuvent suivre l'évolution des thèmes de recherche, *(iii)* l'opération de classification, à savoir l'affectation à telle classe plutôt qu'à telle autre, est délicate et peut toujours être sujette à caution, *(iv)* l'affectation à un thème donné fige une vision univoque et *a priori* du texte, alors que DECID, en se fondant directement sur les textes, vise à s'affranchir d'un passage obligé par les cadres descriptifs connus (qu'il s'agisse de l'organigramme ou de disciplines instituées).

Cette liste de critères est générale, et doit pouvoir être utilisée à chaque fois que se pose la question du choix d'un corpus pour constituer la base de profils de DECID. La démarche suivie ici a une portée *méthodologique*. Après un recensement sérieux des documents électroniques existants, la liste de critères sert à retenir le ou les corpus les plus appropriés.

e) Choix d'un corpus : les textes d'Action

Les corpus possibles identifiés sont confrontés méthodiquement à l'ensemble des critères retenus, à l'aide d'un tableau. Il en ressort que :

- Les Actions (ARD/AID) sont les plus aptes à fonder la définition des profils, en français.
- La construction de profils en anglais souffre d'un manque de données. Des résumés (succincts) en anglais pourraient être recueillis par la Collection des Notes DER, les textes de PPRD et les articles d'*Epure*. Or ces trois corpus ne fournissent qu'une représentation partielle de la DER (les deux derniers ne donneraient guère qu'une quinzaine de résumés par an). Quant à la traduction d'un corpus français, elle requerrait un traducteur professionnel (pour être au plus proche de documents originaux en anglais) et la correction d'un expert (potentiellement l'auteur du texte français), mieux à même de connaître la terminologie précise, usitée, et les correspondances entre les deux langues.
- Le texte intégral est encore une denrée rare dans les bases de données d'entreprise. Il est vrai que les procédures de récupération de ces textes restent lourdes : recueil des versions électroniques, conversion des formats divers. Le développement de la rédaction de documents structurés (par exemple via les traitements de texte avancés, avec module de génération de SGML), du travail collectif en réseau et du partage des données, l'industrialisation de la reprise de documents papier, la constitution de mémoires d'entreprise (bases consignant l'expertise acquise par les équipes, exprimée notamment dans les documents produits), pourraient grossir le volume des documents accessibles en texte intégral.
- Les autres difficultés tiennent au mélange des genres, ou à un recueil irrégulier.
- La collection des Notes Internes de la DER pourrait fournir une base de test, non exhaustive, mais à grande échelle : il s'agirait par exemple de vérifier que chaque est Note est bien rapprochée des profils de ses auteurs pour la période correspondante. Pour le moment on dispose, sous forme électronique, des synthèses, l'idéal serait d'avoir aussi le texte des premières pages de la Note.

> « L'indexation automatique est légèrement meilleure pour le corpus 4 [page de garde, page de synthèse, sommaire, et les 3 pages suivantes] que pour le corpus 3 [texte intégral], le gain intervenant sur la précision apportée aux indexations T (+1%) [Thesaurus] et N (+3%) [Nouvelle Terminologie], le silence restant constant, ce qui accrédite l'hypothèse de n'indexer que le début des documents (premières pages) où est ciblé le « sujet » traité. » (Monteil 1993, p.17)

Les textes d'Actions permettent de construire une base de profils pour la Direction des Etudes et Recherches, en associant chaque texte au responsable de l'Action. Un moyen d'avoir aussi quelques liens vers des personnes d'autres Directions d'EDF est de considérer aussi le(s) interlocuteurs, indiqués aux côtés du responsable de l'Action, dans la partie administrative de la fiche.

Une analyse approfondie des textes d'Action est présentée en annexe.

3. Documents soumis au système pour être confrontés aux profils

a) Les observations

Les requêtes adressées à DECID sont traitées entièrement automatiquement, par une machine : cela permet d'afficher la confidentialité de la transaction, et de ne pas brider les usages de l'application. C'est important dans un contexte opérationnel, et non purement expérimental. De plus, la conservation et l'utilisation d'un tel corpus ressortirait de la déontologie. Par conséquent, les requêtes sont analysées au vol et ne sont pas mémorisées par le système. L'exploitant de DECID ne dispose pas de l'ensemble des textes de requête soumis.

Il reste l'observation de sa propre pratique personnelle, les réactions des personnes qui essayent l'application lors de démonstrations, les échos reçus dans les messages adressés à l'administrateur du serveur, le dialogue avec les plus gros utilisateurs (que leur fonction amène à constamment rechercher des interlocuteurs, ou à diffuser des documents très techniques ou confidentiels).

b) La vraisemblance

De nombreux éléments incitent l'utilisateur de l'application DECID à soumettre un texte comme requête. Il y est d'abord explicitement invité, à plusieurs reprises (page d'accueil, page de soumission de la requête, aide en ligne), et même encouragé, car c'est présenté comme le mode d'interrogation par lequel les résultats seront les meilleurs. L'ergonomie accompagne ces propos. La fenêtre d'entrée de la requête est large et s'étend sur plusieurs lignes, comme pour souligner que le système attend non pas un mot ou deux mais plutôt un paragraphe : cet affichage contraste avec celui des moteurs de recherche sur Internet, pour lesquels la zone d'entrée de la requête correspond à l'affichage de quelques dizaines de caractères. La possibilité de soumettre un (extrait de) document par un copier / coller, ou d'indiquer un fichier contenant le document-requête, figurent comme des voies rapides et efficaces d'interroger le système.

Mais les requêtes ne sont pas toutes des « textes », tant s'en faut ! Dans de nombreux cas, ce sont quelques mots ou expressions qui ont été entrés par l'utilisateur. Trois facteurs au moins expliquent cela.

La pratique habituelle d'interrogation d'un système documentaire, maintenant bien appuyée par les moteurs de recherche sur Internet, passe par l'usage de mots-clés. Autrement dit, la manière spontanée de parler à un ordinateur pour effectuer une recherche d'informations, c'est de cerner l'objet de sa recherche en quelques noms ou expressions, qui situent le domaine, éventuellement focalisent la demande autour d'une notion précise (une personne, une technique, etc.), et peut-être cadrent l'ensemble sur une période ou une région géographique. Entraînés par l'habitude, « conditionnés » par la pratique courante, les utilisateurs n'arriveraient pas à sortir d'un schéma d'interrogation par mots clés et resteraient imperturbables, insensibles aux sollicitations les engageant à changer de mode d'interrogation.

Cette analyse a un fond de vérité, mais n'explique certainement pas tout.

Quelques utilisateurs ont expliqué leur défiance envers l'interrogation par le texte, en extrapolant la situation de l'interrogation par mots clés. Le raisonnement est le suivant : un texte, c'est beaucoup de mots ; or, déjà avec très peu de mots, les résultats sont foisonnants ; combien plus alors, avec tout un texte, les résultats risquent d'être dispersés et peu précis ! Cette crainte doit être dissipée par une bonne communication (dans les présentations de DECID, sur les pages de l'application).

Mais, pour peu que l'on soit soi-même utilisateur, le motif le plus évident de ne pas utiliser de texte est tout simplement l'absence de texte à disposition. En effet, si le point de départ qui motive la recherche est un document dont on a la version électronique sous la main (dans son traitement de texte, dans sa messagerie, dans son logiciel de navigation sur le Web...), la requête la plus directe est un copier / coller ; si c'est une interrogation à partir d'une idée qu'on a dans la tête, d'une demande qui nous a été retransmise oralement, le plus naturel est d'exprimer cela par les quelques mots qui viennent à l'esprit pour formuler le problème, et il est évident que l'on n'a ni les moyens, ni la motivation, de rédiger uniquement pour le système un texte exposant la question. Et pousser dans ce

cas l'utilisateur à faire des phrases ne fait rien gagner, au contraire : la requête prend alors la forme d'une requête en langage naturel, n'ajoutant rien aux mots clés sinon de les couler dans un ensemble de formulations convenues, nécessitant une analyse élaborée (spécialement pour les négations).

La premier enseignement à tirer de la pratique et du bon sens est donc celui-ci : les requêtes soumises à DECID comportent aussi bien des textes que des amorces par quelques mots clés.

Parler ici d'amorce n'est pas anodin : l'usage auquel on peut encourager l'utilisateur qui ne part pas d'un texte, c'est d'enrichir et préciser sa recherche en la relançant avec le texte d'un premier document trouvé en réponse (pour DECID, le texte d'un profil jugé intéressant). De ce fait, le texte n'est plus un mode d'interrogation marginal, mais peut s'intégrer à toutes les sessions de recherche.

Le deuxième type de documents soumis sera donc les documents (ou extraits, ou collage d'extraits) repris de la base des profils.

c) Les suggestions d'utilisation

L'application DECID ouvre des possibilités jusqu'alors inexistantes. Son appropriation n'est donc pas immédiate, elle suppose de percevoir, au fil du travail quotidien, les moments où l'application apporterait une aide ou un plus.

Le déploiement de l'application s'accompagne donc de suggestions d'emploi. Ces suggestions orientent la forme des requêtes :

- CV (voire lettre de motivation), pour la retransmission de candidatures, ou la recherche d'une équipe d'accueil pour une mutation interne ;
- programme de séminaire ou de cours, pour la diffusion ciblée d'une annonce d'une manifestation organisée sur le site ;
- annexe technique d'un CERD (*Contrat Externe de Recherche et Développement*), pour la recherche d'experts validant l'intérêt d'un contrat ;
- synthèse d'un rapport technique, pour le renouvellement et l'ouverture de sa liste de diffusion ;
- page Web, pour le repérage des collègues travaillant sur un sujet repéré lors d'une navigation hypertexte.

d) Les retours utilisateurs

Des utilisateurs, enthousiasmés ou déçus par les résultats d'une requête ou d'une série de requêtes, font part de leurs observations à l'administrateur par le contact par messagerie prévu à cet effet (cf. annexe). Il est extrêmement rare de trouver indiqué le texte de requête, la plupart des utilisateurs semblent réticents à mentionner le sujet de leur recherche[3]. Pour autant, il arrive que la description de la situation rencontrée, ou qu'une illustration précise, permette de comprendre à quel type de requête se réfère la personne.

Beaucoup de requêtes sont évoquées en termes de mots-clés. Il s'agit même parfois *du* mot clé, de *l'*expression ou *du* terme, ce qui laisse entendre une recherche lancée à partir d'une simple désignation. Des réclamations portent sur la prise en compte des termes complexes : les notions recherchées font appel à un vocabulaire technique dans lequel les expressions composées occupent une place importante.

Certains utilisateurs tentent la soumission d'équations de recherche, formées de mots clés et d'opérateurs booléens classiques comme ET et OU (ou AND et OR). Ce mode d'expression leur manque pour traduire les relations significatives entre les notions.

Le dernier type de requête est celui des demandes « déterministes » : par exemple, trouver l'ensemble des responsables d'action citant une action d'un Groupe donné, autrement dit l'ensemble des personnes qui coopèrent officiellement avec ce Groupe. Ce besoin peut recevoir une réponse de DECID, mais cela n'est pas au centre du domaine que doit couvrir le système. En effet, cela correspondrait mieux à l'usage d'une autre application existante à la DER, le *Livre Electronique*.

[3] Cela confirme que la confidentialité, assurée par un traitement par la machine, est un point fort d'un système tout-automatique.

e) Pistes pour l'analyse de quelques familles de documents / requêtes

Formulation libre entrée au clavier

Certains utilisateurs, particulièrement ceux qui sont curieux d'observer la capacité du système à traiter du langage courant, font l'effort de tourner quelques phrases. On reste cependant dans une pratique, relativement codifiée, d'expression d'une recherche d'informations. Une exploration approfondie des formulations ferait peut-être ressortir des schémas d'expression du type, comme l'observe Denise Malrieu sur des demandes de bibliographie par correspondance.

La forme la plus courante reste l'entrée de quelques mots-clés, habitude renforcée par l'utilisation de plus en plus répandue des moteurs de recherche Web (d'ailleurs, ceux qui utilisent DECID le font via l'Intranet...).Les observations faites autour de l'usage des moteurs Web convergent : une requête consiste dans la très grande majorité des cas en 1 ou 2 mots, qui donnent le thème de la recherche (Clarke, Coormack, Tudhope 1997) (Pinkerton 1994). De ce fait, la requête est pauvre en contexte, d'où les aides à la reformulation et à l'enrichissement comme LiveTopics pour AltaVista (option *refine*).

Entrant dans une pratique bien particulière, ces requêtes seraient à considérer et à décrire comme un genre : (Malrieu 1992), (Cousins 1992) apportent déjà quelques observations et enseignements. Une piste d'étude serait d'analyser comment ces requêtes articulent les éléments du thème de recherche : énumération, progression, statuts,... ? On s'accorde souvent sur le fait que ces requêtes sont « plates » : tous les termes semblent s'équivaloir (Grefenstette 1997). Pourtant, l'ordre des termes n'est peut-être pas fortuit : n'aurait-on pas tendance à commencer par ce qui est au cœur de ses préoccupations, ce qui est le plus représentatif de la thématique, puis à compléter par des termes périphériques ou plus spécialisés ? à décrire successivement différents aspects, groupant ainsi les termes sémantiquement ?

Les Notes internes

Dans les notes internes EDF-DER, le résumé fonctionne comme un résumé indicatif, alors que la synthèse constitue plutôt un résumé informatif. En effet, le résumé signale le type d'informations que l'on peut trouver dans le document, mais sans les donner. La synthèse en revanche donne la teneur des points abordés les plus importants, les résultats. Pour avoir des descripteurs sur le fond, le contenu du document, l'utilisation de la synthèse sera donc généralement préférable.

Les Curriculum-Vitae et lettres de candidature

Un CV, en citant les expériences ou stages antérieurs, situe avec des termes techniques précis certaines qualifications et compétences acquises du candidat ; par contre, la lettre de motivation tend à faire usage de termes généraux (pour ne pas réduire le champs des domaines d'embauche éventuels, pour être compris des différents acteurs de l'entreprise...). D'autre part, le CV a un point de vue rétrospectif, la lettre un point de vue prospectif.

L'usage veut que la lettre de motivation soit manuscrite : cela rend très improbable son utilisation comme requête adressée à DECID. La soumission du CV est aussi un peu complexe car elle demande de passer du papier à l'électronique par un outil de reconnaissance optique de caractères... Les cas où cela se présente de façon beaucoup plus simple existent : une candidature transmise par courrier électronique ; le chercheur qui soumet son propre CV au système pour avoir des pistes pour une mutation interne.

Les CERD (*Contrats Externes de Recherche et Développement*)

Pour les CERD, le résumé a un vocabulaire plus sélectif que l'Annexe Technique. En effet, le résumé, très bref, reste au cœur du sujet ; l'Annexe Technique peut elle aborder des problèmes voisins liés avec l'étude présentée.

Cependant, le niveau de généricité / spécificité des termes de l'Annexe Technique peut mieux correspondre à celui des termes de l'Action (utilisée côté profils) que celui du résumé (très concis, donc usant d'hyperonymes et n'entrant pas dans les détails).

Les pages Web

Elles débordent les critères d'un genre. Pour autant, des régularités d'ensemble peuvent être décelées et utilisées.

- les « niveaux » des pages sont extrêmement variés : page d'accueil générique, exemple, illustration, article de référence, etc. (Koch 1996)
- les pages sont rarement conformes à une DTD HTML ; certaines erreurs sont courantes : omission des balises de plus haut niveau (`HTML`, `HEAD`, `BODY`), absence d'un élément attendu immédiatement un élément donné, balise fermante manquante ou ne correspondant pas à la balise ouvrante, invention de balises, mauvais placement des ancres dans l'emboîtement des éléments (la DTD prévoit que les ancres soient à l'intérieur), etc. (Woodruff, Aoki, Brewer, Gauthier, Rowe 1996) (Amitay 1997)

  Le concepteur courant est avant tout sensible à l'effet rendu à l'écran -peu importe si la syntaxe n'est pas respectée, si le logiciel de navigation s'en débrouille ; peu importe si le nom des balises ne correspond pas vraiment à l'usage qu'on en fait, si la présentation obtenue est celle que l'on cherchait.
- les informations de type catalogage (résumé, mots-clés), dans les balises prévues à cet effet, sont rares, ou parfois détournées (le *spanning* par exemple consiste en la répétition massive d'un mot-clé pour obtenir un poids dominant dans les résultats des moteurs de recherche) : les moteurs de recherche ont renoncé à en faire leur unique source d'information.
- certains documents sont conçus, ou adaptés, pour une publication hypertexte ; d'autres sont la reprise, sans réaménagement, de documents papiers existants. Ce sont deux optiques très différentes. (Oßwald 1995)

B. LES CODAGES DE TEXTES

1. Des décisions significatives et déterminantes

Le codage traduit pour la machine la vision que l'on a des textes, prépare les modes d'accès au sein du matériau textuel, et conditionne les traitements qui pourront être effectués.

Un équilibre est à trouver entre un codage pauvre, pour lequel le texte est réduit à une chaîne de caractères, et un codage très détaillé, se prêtant mal à la diversité des textes à considérer et imposant un moule particulier, une interprétation particulière, aux données.

D'autre part, le volume important de notre corpus (côté profils) et le traitement automatique à la volée des textes soumis (côté requêtes) impose que le format choisi pour la représentation des textes en entrée de DECID puisse être automatiquement produit à partir des divers formats courants originaux. Il doit donc pouvoir s'ajuster à des descriptions plus ou moins riches, depuis de simples textes ASCII à différents modèles de documents structurés (SGML).

Par delà ces remarques techniques, le codage est une interprétation du texte, puisqu'il explicite ce qui sinon serait implicite ou effacé dans la version électronique. Aussi y a-t-il une déontologie du codage des textes : l'étape d'encodage doit s'efforcer de respecter le texte, d'être en adéquation avec l'usage visé explicité, et de fournir toutes indications utiles au lecteur final éventuel. Le codage ne doit pas forcer une interprétation ni prêter au texte des indications qu'il n'a pas. Pour ce que l'on choisit de coder, on devrait donc expliciter le mécanisme interprétatif suivi. Par exemple[4] :

- Tout retour à la ligne du texte original est compris comme la fin d'un paragraphe et transcrit par une balise `<P>`.
- On se donne une liste de prénoms. Si on trouve un prénom de la liste dans le texte, et qu'il est suivi d'un mot commençant par une majuscule, alors on étiquette l'ensemble comme un nom de personne, composé d'un prénom puis d'un nom de famille.

Ceci permet de prévoir les écarts (erreurs, omissions) par rapport à l'idée intuitive que donne le nom de l'élément de son contenu.

Dans le cas où le processus de codage est difficile à formaliser, un attribut peut être ajouté pour donner une appréciation de la fiabilité des décisions de codage prises. La TEI[5] prévoit cela dans un module additionnel.

2. Cadre formel : SGML

a) Points de repères introductifs - vocabulaire

SGML, HTML, XML

Le formalisme courant pour encoder la structure d'un document électronique est SGML.

- SGML (*Standard General Markup Language*) donne un formalisme pour écrire des grammaires de documents : cela définit les éléments qui composent un document du type décrit, et leurs agencements possibles. Par exemple, un article scientifique peut être représenté comme : un titre, des auteurs, un résumé, des mots-clés, un texte composé d'une suite de paragraphes, et une bibliographie.
- HTML (*HyperText Markup Language*) est une instance particulière de SGML, utilisée pour la mise en forme des documents sur le Web.
- XML (*eXtensible Markup Language*) est un sous-ensemble de SGML, défini en sorte d'exclure les structures les plus complexes de SGML, et ainsi de permettre la manipulation de documents et d'alléger les échanges de données en se passant de la déclaration explicite d'un modèle. XML est la

[4] Ces exemples très simples ne prétendent pas être des algorithmes opératoires de traitement : on en perçoit très facilement les limites. Ils ont été choisis pour donner une illustration brève.

[5] *Text Encoding Initiative*, présentée un peu plus loin dans ce chapitre.

plus connue des nouveautés introduites pour pallier les insuffisances de HTML, rigide et uniquement axé sur la présentation (Mace, Flohr, Dobson, Graham 1998).

Il n'est pas innocent de considérer d'emblée SGML, HTML et XML. L'actualité fait porter l'attention sur HTML et maintenant XML, et relègue SGML aux oubliettes. Or SGML n'est pas une alternative, théorique et complexe, à HTML et XML : c'est le cadre général qui permet de décrire et situer tous ces modes de structuration. Une étude de SGML englobe les fondements de HTML et XML.

Des notions complémentaires, rencontrées dans les pratiques de HTML et XML, peuvent ensuite enrichir l'étude. De plus, HTML et XML synthétisent les idées et mécanismes de SGML qui se sont révélés concrètement les plus pertinents (équilibre entre simplicité et efficacité, sans nécessairement aller jusqu'au bout des constructions théoriques puissantes mais complexes et peu usitées).

DTD : le Modèle de Document

Le patron donnant la structure d'une famille de documents est appelé DTD (*Document Type Definition*) de cette famille. Une DTD est un modèle, une description aussi bien de la structure organisatrice que des différents éléments qui figurent dans ce type de documents. Un document de la famille correspondant au modèle fixé par la DTD est une *instance*.

Un même document peut-être décrit par plusieurs DTD, instituant différents degrés de contraintes. Par exemple, pour la réalisation d'un document électronique, on peut avoir recours successivement à une *DTD de saisie* (qui tolère que le document soit inachevé), une *DTD d'épreuvage* (qui vérifie l'allure générale, donne une première maquette permettant la mise en forme, sans rentrer nécessairement dans le détail de tous les éléments), et une *DTD de référence* (la plus complète et la moins évolutive, elle correspond au format le plus riche en informations sur la structure, dans lequel le document est enregistré et utilisé).

La DTD est aussi liée aux corpus et aux applications pour lesquels elle est conçue : elle représente un point de vue propre à un domaine, à une approche. D'où des résumés comme ceux-ci, même pour les standards les plus répandus : HTML est une DTD pour présenter sur écran ; CALS est une DTD pour les documents techniques avec des tableaux ; la TEI est une DTD pour la littérature ; ATA est une DTD pour la documentation aéronautique ; etc.

b) Expressivité du formalisme

Points fondamentaux

Les notions et propriétés structurelles que peut coder SGML se déduisent d'une analyse de la construction de SGML, et notamment des opérateurs que la norme prévoit. Dans la perspective du codage de textes, les principaux aspects à retenir de SGML pourraient être les suivants :

- La déclaration de types d'*éléments*, identifiés par un nom ; un élément occurrence se réalise par une chaîne de caractères, il délimite une zone connexe (pas de trou, de discontinuité).
- La nature du *contenu d'un élément* : d'autres éléments, ou / et des chaînes de caractères quelconques. Le type peut également fixer que l'élément soit vide : il correspond alors à une localisation, un point entre deux caractères. Les caractères considérés sont les caractères alphanumériques, les symboles, les espaces (blancs) ; les caractères de contrôle et de présentation (comme le retour à la ligne ou les tabulations) sont ignorés. Les instructions de mise en forme qu'ils traduisent n'ont en effet pas à être mêlées aux caractères, au contenu textuel, mais sont à reprendre au niveau des éléments, en les associant à un type.
- Les *règles de composition* des éléments : elles peuvent décrire l'inclusion (en distinguant l'emboîtement direct, et l'enchâssement indirect possible à travers d'autres éléments), la succession et la juxtaposition (ordre déterminé ou non), un ensemble d'alternatives structurellement équivalentes (choix), la présence obligatoire ou facultative, la possibilité de répétition (par opposition à l'unicité). La structure induite est arborescente, elle ne permet pas le chevauchement

d'éléments[6]. En revanche, plusieurs balisages peuvent coexister dans un même document : chaque balisage ignore alors les balises qui ne font pas partie de son système.

- Des *informations associées* à un type d'éléments : le mécanisme des attributs permet de qualifier chaque occurrence de façon autonome par rapport à son contexte, sans intervenir dans la structure hiérarchique d'inclusions. Chaque attribut est unique pour le type d'éléments auquel il est associé. Sa déclaration précise s'il est facultatif, ou bien doit toujours figurer, ou encore possède une valeur par défaut (une constante, ou la reprise contextuelle de la valeur pour l'occurrence précédente). Les informations enregistrées par les attributs peuvent être davantage précisées que les chaînes de caractères de contenu : par exemple valeur prise dans une énumération de possibilités, ou caractères uniquement numériques (donc représentation d'un chiffre). La gamme de valeurs que peut prendre un attribut est donnée par son type : à part la donnée d'une énumération de valeurs au gré du concepteur de la DTD[7], il y a une petite dizaine de types prédéfinis. Ceux-ci opposent des chaînes de caractères à dominante alphabétique ou numérique, ainsi que des valeurs simples à des listes de valeurs ; on ne peut se forger un patron à l'aide d'une expression régulière par exemple. Certains types introduisent des contraintes globales régissant l'ensemble des valeurs prises sur toutes les occurrences : une même valeur pour toutes les éléments du type, ou tout au contraire valeurs toutes différentes pour toutes les occurrences et tous les éléments (on parle alors d'*identifiant*). Une autre contrainte disponible via le typage de l'attribut, utilisée pour gérer les liens d'un élément vers un autre, vérifie l'existence d'un identifiant de la valeur indiquée.
- Des *liens orientés*, d'un élément (occurrence) vers un autre : ces liens sont donc réalisés par des attributs de type particulier. La définition de liens est intéressante pour échapper à la linéarité du texte, et associer des portions de textes artificiellement isolées à cause de la connexité imposée aux éléments.

Quelques aspects complémentaires à remarquer

Les travaux autour de la conception et de l'utilisation de DTD ont dégagé des concepts qui guident la mise en œuvre.

- Un élément qui définit un groupement stable d'autres éléments, qui se décompose entièrement en constituants identifiés (éléments), est appelé *cristal* (« crystal »). Des cristaux typiques sont par exemple les noms de personnes (qui se décomposent en prénom et nom de famille), les références bibliographiques (structurées en champs), les adresses (dans lesquelles on isole le numéro, la rue, le code postal, la ville, etc.). Les cristaux apparaissent comme des unités de contexte, qui médiatisent l'accès à leurs constituants.[8]
- Le mécanisme des *inclusions* (et des *exclusions*), qui est utilisé pour indiquer la possibilité (ou l'impossibilité) de présence d'un élément dans un autre à quelque niveau de profondeur que ce soit, a pour effet que le contenu autorisé d'un élément dépend de son contexte (à savoir ici les élément qui l'incluent). Il n'est alors pas toujours possible d'isoler un fragment du document et de faire une validation locale.
- Un même document peut avoir plusieurs DTD indépendantes, incompatibles du point de vue du découpage structurel (du fait de décalages, de chevauchements), traduisant chacune une vision du document. SGML permet de considérer conjointement les DTD sur les structures (éléments) qu'elles partagent, et de traiter indépendamment les structures propres à chaque vue.

---

[6] Il y a des « astuces » pour contourner cette contrainte, par exemple en codant les parties susceptibles de se chevaucher non pas comme des éléments SGML, mais comme des zones dont on indique seulement le début et la fin, par des bornes (*milestones*) déclarées comme des éléments vides (*empty*) (on trouvera un exemple dans (Chahuneau & *al*. 1992, §4.5), dans le contexte du codage d'annotations dynamiques de textes, effectivement particulièrement susceptibles de recouvrements mutuels). Ceci se paye néanmoins au prix d'un affaiblissement général de la précision des informations qui peuvent être indiquées dans le modèle de document et exploitées dans le parcours d'une instance.

[7] La liste des valeurs d'un type énuméré doit cependant respecter quelques limites arbitraires : des attributs qui qualifient un même élément doivent avoir des ensembles de valeurs disjoints ; les valeurs d'un type énumératif sont représentées par des chaînes de longueur ne dépassant pas 8 caractères.

[8] La notion de *cristal* est présentée dans (Sperberg-McQueen, Burnard, 1995, § 2.2).

- Sans rentrer dans le détail, il faut aussi mentionner le mécanisme très polyvalent des entités. Il permet la consultation de fichiers externes, l'appel de procédures externes. Il permet aussi plusieurs formes de paramétrage. Les entités sont un moyen de représenter des classes d'éléments et d'attributs, qui partagent certaines caractéristiques de comportement. Dans le même ordre d'idées (la définition de classes d'objets), la norme HyTime, prolongement de SGML pour des données multimédia, introduit la notion de *formes architecturales* : ce sont des modèles de structure de données, pour la description d'éléments et de groupes d'attributs, qui servent de briques dans l'élaboration des DTD. Le catalogue de formes architecturales diffusé par HyTime condense une connaissance des objets et des mécanismes pertinents pour les applications multimédia.

SGML et les textes

Ceci conduit naturellement à s'interroger sur le mode de codage des textes que permet SGML, car les textes ne sont finalement qu'une partie des données couvertes par SGML. Compte-tenu de ses possibilités expressives, quelle vision SGML dessine-t-il du texte ? Comment cette vision se prête-t-elle aux propriétés des textes ? Quelle lecture a-t-on des textes à travers un codage SGML ? Est-ce que se dégage un mode d'emploi des balises particulier aux textes ?

Quelques remarques peuvent d'ores et déjà être formulées, avant de se tourner vers l'expérience reflétée par des DTD textuelles à larde diffusion.

Propriétés interprétatives

Le codage SGML est univoque, explicite et discret.

- *univoque* : pour une DTD donnée, les balises d'une instance ne peuvent traduire qu'une seule structure.
- *explicite* : l'option `OMITTAG` n'est utilisée que quand la présence de la balise est déjà déterminée et établie dans tous ses contextes d'apparition.
- *discret* : la DTD énumère l'ensemble des éléments perçus. Chaque balise indique ensuite le repérage d'un et un seul élément. Seul le mécanisme des attributs permet d'introduire une qualification dont le nombre des valeurs possibles n'est pas limité de la même manière.

Ceci va dans le sens de l'efficacité du traitement formel, informatique. Les propriétés de cette interprétation mécanique tournent le dos à celles de l'interprétation à laquelle procède le lecteur humain d'un texte.

Sections et passages

En délimitant des zones ou en marquant des points frontière, le codage matérialise des *sections* (qui, comme leur nom l'indique, se définissent en termes de *coupures* dans le texte). Cela ne correspond pas directement à la définition de *passages*. Les passages d'un texte se construisent à partir d'un point d'attention, d'un mode de lecture, alors que les sections viennent du texte. Les passages peuvent se chevaucher, jouent davantage sur la proximité et la propagation que sur la rupture et la discontinuité. Ils se définissent dynamiquement et en ce sens ne sont effectivement pas de l'ordre du codage.

Par exemple, la segmentation d'un texte en paragraphes fournit des sections ; la définition de voisinages en tout point du texte sous forme d'une fenêtre glissante (*i.e.* on considère *n* mots qui précèdent et *n* mots qui suivent le mot d'ancrage de la fenêtre) correspond davantage à l'idée de passages.

Les sections sont généralement très marquées dans les textes scientifiques et techniques. La segmentation prépare une consultation opportuniste (« je cherche le paragraphe concernant tel cas de figure »), localise (numérotage) et désigne (intitulé) l'information. Tout ceci favorise une prise de connaissance rapide (survol, table des matières).

Ces préoccupations d'efficacité informative, d'annonce et de morcellement, n'ont plus court dans le monde des œuvres littéraires, si bien que les sections y sont beaucoup moins prégnantes.

3. Quand les données sont des textes : apports de la Text Encoding Initiative

a) Présentation de la TEI, et de ses postulats fondateurs

Un groupe de travail international se penche justement sur la question du codage SGML des textes : il s'agit de la TEI, *Text Encoding Initiative*. La TEI propose des conventions de description des corpus sous forme de recommandations, de repères, dans la mise en œuvre de SGML. S'étant accordé un recul critique vis-à-vis de ce formalisme, elle l'adopte finalement et déclare y trouver matière suffisante pour décrire les textes.

> From the start, the standards perceived as most relevant to the work of the TEI were SGML and existing applications of SGML ; the TEI therefore decided that its Guidelines should, if possible, use the formalisms of SGML, with the caveat that, if the needs of research required constructs unavailable in SGML, research was to take precedence over the standard. It is a tribute to the expressive power and generally good design of SGML that no extensions to SGML were in practice found to be necessary. (Sperberg-McQueen, Burnard, 1995, § 1.2.3)

La TEI réunit des experts internationaux, qui s'appuient sur leur expérience propre. Divers usages des corpus sont ainsi représentés, pour des corpus variés. L'ambition de la TEI est de n'exclure aucun texte ; par là même elle appréhende implicitement, mais de front, la textualité.

> The Guidelines apply to texts in any natural language, of any date, in any literary genre or text type, without restriction on form or content. (TEI P3, § 1)

La TEI se veut la plus fidèle aux documents originaux : elle s'engage dans le recensement de l'ensemble des caractéristiques des textes. Les caractéristiques perçues comme significatives, toutes les distinctions faites dans l'encodage du texte original le cas échéant, doivent recevoir une description dans la TEI.

> For interchange, it must be possible to translate from any existing scheme for text encoding into the TEI scheme without loss of information. All distinctions present in the original encoding must be preserved. Any conventions used in the original encoding but not made explicit by its encoding format should be documented within the interchange format.
>
> When the TEI scheme is used as an interchange format for pre-existing encodings, only those textual features expressed explicitely in the pre-existing encoding format can be converted into their TEI equivalents. If the original encoding lacks, for example, information about paragraph breaks in the original source, so will the TEI version, even though marking this particular feature is strongly recommended to those creating new electronic texts. As the final item in the Poughkeepsie Principles makes clear, translation into the TEI scheme should not be construed as requiring the addition of any new information not present in the original encoding.
>
> (Sperberg-McQueen, Burnard 1995, § 1.3.2.)

La TEI relève deux cas de figure, où la présentation devient partie intégrante du texte et mérite d'être encodée. Le premier est celui d'un texte d'archive, dont l'exemplaire original est unique et dont le support est étudié en tant que tel par les paléographes et les philologues. Le second cas est celui où la mise en forme est perçue comme significative, mais que son interprétation peut être multiple : on évite alors d'imposer une interprétation plutôt qu'une autre, et l'on s'en tient, pour le codage, à l'indication apportée par le texte. Cela concerne souvent les procédés de mise en valeur (italiques, gras, etc.). C'est en ce sens que la TEI entend proposer un codage dit descriptif.

> In the TEI encoding scheme, descriptive markup has in general been preferred to presentational markup. TEI tags typically describe structural or other fundamental textual features, idependently of their representation on the page. In some cases, however, — e.g. for codicologists, paleographers, or analytical bibliographers — the physical appearance of the original text carier can be the primary object of interest. In others, there may be no consensus as to the meaning of all aspects of the text's physical appearance ; it must therefore be possible to represent them as explicitly as possible, without being forced to speculate as to their meaning. For use in such cases, the TEI defines elements (e.g. <hi>, for highlighted phrases of any type) which simply record some salient fact about the appearance of the source text, without requiring any overt interpretation on the part of the encoder. A great deal of work remains to be done, however, to ensure that students of a text's transmisson or physical presentation can conveniently record the relevant information in a precise, readily processable form. (Sperberg-McQueen, Burnard, 1995, § 1.3.4)

La mise au point de conventions de codage est apparue nécessaire pour rendre possible les échanges de données. En effet, le codage électronique propre d'un corpus est un travail monumental, et donc la communauté scientifique entière bénéficie d'un apport formidable si les travaux de chacun sont réutilisables par tous. L'accord porte sur plusieurs niveaux : il faut déjà s'entendre sur les caractéristiques à traduire grâce au codage, puis convenir d'une notation. C'est le premier point qui est déterminant ; et d'ailleurs la TEI introduit une certaine souplesse sur les noms donnés aux balises en les mettant en paramètres, tout en conservant en attribut le nom conventionnel permettant de rétablir l'équivalence.

> We distinguish sharply between the textual *feature* marked and the string of characters or other device used to mark it in the electronic text. That string of characters is conventionally referred to as a *tag*. [...] Note that a tag must refer to a feature, but a feature need not be tagged. The same feature can be denoted by many identifiers in different encoding schemes. Thus in the TEI tag set, the element <div1> may be used to identify occurrences of the feature *chapter*. In other tag sets the same feature might be tagged with <chapter>, <Kapitel>, <chapitre>, or <caput>. [...] The TEI provides mechanisms [...] [with which the names associated with the features] may be changed without altering either the definition of the features themselves or the syntax which governs their occurrences. (Sperberg-McQueen, Burnard, 1995, § 2.1).

Devenir un format d'échange utilisable le plus souvent possible conduit la TEI à prévoir un très large éventail d'éléments encodables. L'optique est de donner les moyens de coder tout ce qu'un utilisateur peut (raisonnablement) souhaiter enregistrer, et pas d'émettre un point de vue critique et sélectif sur ce qui vaut d'être codé[9].

La TEI choisit une approche descriptive plutôt que prescriptive ou normative. Elle se veut extensible et évolutive car il en va de sa viabilité. Enfin, les commissions de la TEI ont la saine prudence de rester ouvertes, mettant ainsi leurs détracteurs en face de leurs responsabilités et engageant les débats sur le terrain d'une critique constructive.

> We therefore offer these Guidelines to the user community for use in the same spirit of active collaboration and cooperation with which they have so far been developped. [...] we anticipate that users of the TEI Guidelines will in some instances adapt and extend them as necessary to suit particular needs ; we invite such users to engage in te further development of the Guidelines by working with us as they do so. (TEI P3, Preface)

b) La proposition de la TEI

Conformément à son optique de modularité, la TEI organise ses éléments en trois groupes :

[9] Il serait donc déplacé de faire grief à la TEI de décrire l'encodage de certains éléments –dans le cas d'un encodage que l'on jugerait discutable– si certains utilisateurs pensent avoir besoin et ont recourt à un tel encodage. En dernier ressort, le travail de la TEI ne peut être évalué que par rapport à son adéquation aux usages effectifs, non par rapport à une réflexion théorique sur ce qu'il est justifié de coder dans un texte. En revanche, elle reflète les conceptions en vigueur dans la communauté scientifique, et ce sont ces conceptions qui, à travers la TEI, peuvent être discutées.

	Pour tous les textes	Pour certains textes (8 types prédéfinis)
Eléments obligatoires (toujours disponibles, pas nécessairement utilisés)	Les balises « noyau » (*core*) : elles comprennent l'*en-tête*, qui documente l'édition électronique du texte (informations bibliographiques de catalogage, mais aussi informations sur le fichier, le contexte du codage, les règles suivies pour le codage) ; en ce qui concerne le texte lui-même, on dispose de balises pour décrire la structure générale du texte (organisation en chapitres, en parties) et de balises pour repérer certains éléments particuliers (nom de personne, mot souligné, etc.).	Les balises « de base » (*base*) : un et un seul type doit être sélectionné parmi les suivants : prose, poésie (versifiée), théâtre, retranscription de paroles orales, dictionnaire, terminologie ; et deux modes de mélange de ces genres. Par défaut, la prose n'ajoute aucune balise spécifique ; en effet, il est considéré que tout genre de texte comporte une part de prose, et donc que les éléments correspondant à la prose font partie de l'ensemble noyau.
Eléments facultatifs	Les balises « additionnelles » (*additional*) : elles correspondent essentiellement à des éléments d'appui pour un contexte d'utilisation (analyse linguistique, édition critique, regroupement en corpus, etc.)	

Le modèle que la TEI donne de la textualité ressort déjà de cette répartition des éléments. Les éléments de base rappellent (sans pleinement la traduire) l'incidence fondamentale, constitutive, des genres[10]. Les balises additionnelles manifestent la diversité des lectures suscitées par les textes. L'en-tête marque l'ancrage intertextuel et situationnel du texte, et reconnaît le caractère interprétatif de l'opération de codage. Les autres éléments noyaux refléteraient les articulations fondamentales de l'organisation interne des textes. C'est ce dernier point qui reste à explorer pour expliciter les propriétés structurelles constitutives de la textualité, telles que les a cernées la TEI.

> The Guidelines are built on the assumption that there is a common core of textual features shared by virtually all texts and virtually all serious work on texts. (TEI P3, § 1.2.1)

Un élément particulier du document est sa *couverture*, comprenant un certain nombre de mentions, comme le *titre*, l'*auteur*, l'*édition*. Aux frontières du texte, on trouve encore les *pièces liminaires* (au début) et les *annexes* (en fin), qui sont distinguées du *corps* du texte et l'encadrent. Leurs composants (*préface*, *sommaire*, *résumé*, *index*, *bibliographie*, *glossaire*, etc.) sont tous considérés identiquement comme des *parties* (c'est un attribut appelé *type* qui peut être utilisé pour les distinguer). La TEI prévoit le cas de recueils de textes, avec un élément *groupe de textes* qui s'intercale entre le texte englobant et les textes réunis : chaque texte peut avoir ses pièces liminaires et annexes, et le groupement des textes dispose lui aussi de ses éventuelles pièces liminaires et annexes. En revanche il n'y a qu'un seul *en-tête*, correspondant au recueil, puisqu'il est considéré comme un seul document.

Les *parties* (ou *divisions*) sont fonctionnellement définies comme ce qui regroupe les éléments du niveau des paragraphes. Elles peuvent être réparties en niveaux, les divisions de niveau *n* ne pouvant inclure que des balises de niveau *n+1* et plus. Certains éléments sont propres aux débuts ou aux fins de divisions, comme un *épigraphe* ou une *signature* : la TEI enregistre ici surtout des éléments typiques des lettres à un correspondant. Une autre forme de division du texte est indiquée

[10] Les huit ensembles de base distingués par la TEI délimitent plutôt des champs pratiques, plus larges que des genres. Ils se fondent sur la présence d'éléments d'expression particuliers et qu'il s'agit d'enregistrer, comme le vers ou la didascalie. Les genres construisent leur forme en s'appropriant ces éléments. Ainsi, un roman par lettres peut reprendre les éléments formels de la lettre sans devenir l'équivalent d'une concaténation de lettres : la lettre dans un roman n'est pas à proprement parler une lettre, et le roman en question n'est pas un arlequin. La lettre fait figure ici de genre inclus, qui est dominé par le genre englobant.

par des *bornes* : au lieu de marquer le début et la fin d'une zone, dont on exprime ainsi l'unité du contenu, on place un élément frontière qui sépare un avant d'un après. Ces bornes sont typiquement utilisées pour coder le changement de ligne ou de page. L'utilisation judicieuse de ce double mécanisme de division (parties et bornes) évite d'avoir des éléments qui se chevauchent et préserve ainsi une vue intégrée du texte. Il reste le cas de parties particulières, celles qui ne sont pas autonomes et dont le contenu est généré automatiquement en fonction du reste du texte : par exemple le *sommaire* ou l'*index*. Une balise suffit pour les coder : elle indique l'emplacement où cette partie doit s'insérer, et un attribut (type) précise ce qui doit être généré (*table des illustrations*, *index*, etc.).

Les divisions les plus fines sont les *paragraphes* (des balises de l'ensemble de base s'ajoutent éventuellement, en fonction du type de texte : par exemple, les *vers* en poésie). Ce qui les distingue des parties, c'est qu'ils ne s'imbriquent pas les uns dans les autres. D'autre part, ils apparaissent comme un niveau charnière : en deçà du paragraphe, les éléments décrivent des expressions ou des empans de texte d'une longueur de quelques mots ou quelques phrases. Au-delà, on a affaire à des modes de structuration des paragraphes : liste, regroupement (ce sont les *parties*, appelées *divisions*). La TEI propose en conséquence trois classes d'éléments : les morceaux (*chunks*), qui correspondent aux paragraphes ; les expressions (*phrases*), qui s'intéressent aux mots et aux suites de mots ; et les intermédiaires (*inter-level*), qui s'appliquent tantôt à des paragraphes, tantôt à des mots. Les divisions ont alors pour fonction de grouper des composants (*components*), ces derniers étant définis par l'union de la classes des morceaux et celle des intermédiaires.

Les *éléments intermédiaires* sont les *listes*, les *références bibliographiques*, les *annotations* (et *didascalies*), et les *citations*. Pour les listes, la TEI propose de distinguer (par un attribut) si les items sont mis sur le même plan (pas d'ordre explicite), numérotés (progression) ou étiquetés (un identificateur, comme une étiquette, introduit chaque élément de la liste).

Que prévoit de coder la TEI, en deçà du paragraphe ?

On repère les éléments contrastant avec le contexte immédiat. Ils sont d'ailleurs traditionnellement marqués par un changement typographique dans les imprimés. On a ainsi : le *soulignement* visant un *effet rhétorique ou linguistique* ; un *terme emprunté* (à une *langue étrangère*, au vocabulaire *technique*, à un auteur, à un langage formel -*programme* informatique, *formule* mathématique, et le cas particulier de SGML dont les balises citées ne doivent pas être interprétées dans le traitement) ; un mot considéré pour sa forme graphique, et dont le signifiant est mis en exergue (*mot mentionné*, *désignation* proposée comme une étiquette, *titre cité* et renvoyant à un ouvrage). La famille des emprunts s'élargit à toutes les formes de citation, y compris les *prises de parole* dans un dialogue et les *extraits* tirés d'un autre écrit.

Une autre catégorie d'éléments repérés correspondent à des informations référentielles et à des notions repères : *nom propre*, *date*, *nombre*, *abréviation*. Elles peuvent être rapportées à une forme normalisée, pour retrouver et suivre l'identité du référent à travers des variantes de formulation. Les attributs sont utilisés pour un complément sémantique, par exemple pour préciser si le nom propre désigne une *personne*, un *lieu*, une *organisation*, ou toute autre catégorie qu'on s'est donnée.

On trouve encore quelques cristaux, et leurs éléments propres, dont l'organisation interne tranche sur un continuum textuel : l'*adresse postale*, la *référence bibliographique*, les ingrédients spécifiques à un courrier (la salutation par exemple).

Une deuxième grande famille d'éléments en deçà du paragraphe concerne les interventions de l'éditeur, de l'auteur, du traducteur, ou de tel ou tel lecteur. Par ceci il faut entendre : les *annotations*, l'indication d'*éléments à consigner dans l'index*, les *corrections* par rapport à la *version d'origine*.

Ces éléments font pour la plupart appel au mécanisme de définition de liens. Ceux-ci permettent le *renvoi*, à partir d'un point ou d'une expression, à un point, une expression, un document, voire plus généralement à un élément quelconque en fonction de son emplacement relatif ou/et de la valeur d'un de ses attributs.

Pour les besoins du codage ou de l'analyse, on peut définir d'autres unités du niveau des mots et expressions : la TEI prévoit à cet usage l'élément *segment*. L'attribut *ana* indique à quoi correspond l'unité ainsi délimitée, en l'étiquetant avec une valeur prise dans une grille d'analyse qu'on donne avec l'instance. Un exemple d'utilisation est l'étiquetage par des catégories morphosyntaxiques.

Le découpage en phrases a sa propre balise, mais cela ne fait plus partie des éléments noyaux. La phrase se voit attribuer des définitions multiples : typographique (délimitation par la ponctuation),

syntaxique (autonomie grammaticale), rhétorique (balancement en périodes), etc. La TEI propose de retenir plutôt la définition typographique pour la balise `<s>`. L'élément `<s>` correspond structurellement à un découpage complet en deçà du paragraphe : ce découpage est tel que tout point du texte appartient à une unité `<s>` et à une seule.

Ceci conclut l'inventaire des éléments noyaux.

Restent les attributs généraux, prévus pour tous les éléments codés dans les textes. Ils comprennent : un *identifiant* et une *désignation* (nom ou numéro) (qui individualisent chaque occurrence d'un élément dans un texte), un lien vers un élément *précédent* et un vers un élément *suivant* (on s'inscrit clairement dans la linéarité du texte ; surtout, ceci pourrait être prévu pour rétablir la relation entre des composants discontinus d'un même élément), le rappel du *nom de la balise TEI* correspondante (utile si les balises ont été renommées). Deux attributs s'appliquent au contenu de l'élément (la zone de texte délimitée) : l'indication de la *langue*, et des informations de présentation destinées normalement à permettre de retrouver la *mise en forme originale* du texte.

c) Relevé et commentaire de choix significatifs

Une édition de qualité

La TEI se place dans une perspective éditoriale, et est sensible à la relativité de toute version qui est donnée d'un texte. Le respect de la version originale du texte transparaît dans ses choix de modélisation : le souci est constant de toujours être en mesure de rétablir la version originale, telle qu'elle peut être perçue.

La TEI est consciente de la divergence possible de choix interprétatifs (cf. l'existence de balises pour coder une mise en forme, sans préjuger de son interprétation). Elle prévoit les écarts introduits par le travail éditorial : l'indication de corrections apportées dans le texte par exemple. La multiplicité des usages des corpus, et des analyses qui peuvent y être appliquées, ont également leur place, avec l'élément « à tout faire » *segment*.

Les grilles de valeurs sont assez systématiquement reportées au niveau des attributs : cela est vrai pour le typage des parties (chapitre, section, etc.), la nature et l'identité d'un référent (personne X, objet Y), les catégories sémantiques repérées par une analyse. Cela assure la compatibilité du modèle, à vocation générale, avec l'introduction d'informations particulières.

Une description qui se place au niveau du fonctionnement des éléments : l'illustration des divisions

La TEI décrit les éléments non par leur nom mais par leur fonction, entend-on dire. En effet, pour un certain nombre d'éléments, la TEI fait abstraction de la nature de leur réalisation et n'enregistre que leur fonction. Le cas le plus flagrant est celui des parties (la famille des `<div>` et `<divn>`) : l'élément `<div>` est défini comme ce qui regroupe des paragraphes. Or il y aurait toute une gamme de natures de parties possibles : *chapitre*, *section*, *préface*, etc. Cette sémantique est renseignée par le biais de l'attribut *type*. C'est très ouvert car le système de catégories de parties est libre, il peut tout à fait varier d'un document à l'autre. Cela favorise l'adaptabilité de la DTD mais n'assure aucune comparabilité d'un document à l'autre. L'interprétation de la nature de chaque partie est médiatisée par l'énumération des valeurs qu'on s'est donné pour le codage et l'explicitation des règles suivies pour leur affectation.

Ce mécanisme de report de l'interprétation au niveau d'un attribut ménage la possibilité d'introduire dans le codage des informations plus fines adaptées à chaque point de vue, tout en s'accordant sur des formes structurelles génériques. Chaque interprétation se dote d'un référentiel, en spécifiant la gamme des valeurs d'attributs pertinente pour elle, et aucune grille ne détient l'ensemble des interprétations possibles.

Ce faisant, on abandonne une grande part de contrôle sur la structure : SGML ne pourra vérifier qu'une *section* est incluse dans un *chapitre* et non l'inverse, par exemple.

Du point de vue de la validation SGML, le nom d'un élément ne fait effectivement que déterminer ses contextes possibles. Deux éléments qui ont le même comportement structurel (et les mêmes attributs) devraient, au sens SGML, porter le même nom. A tout le moins, on peut faire

ressortir leur parenté en utilisant systématiquement une entité paramètre qui représente indifféremment l'un ou l'autre[11]. En ce qui concerne la sémantique que l'on accorde à une occurrence indépendamment de sa place dans la structure, ce sont les attributs que prévoit SGML pour l'enregistrer.

L'utilisation de la forme numérotée des divisions (`<div1>`, `<div2>`, jusqu'à `<div7>`) permet de réintroduire un contrôle sur les inclusions possibles. On note le statut particulier de la `<div0>`, autorisée seulement aux frontières du texte, dans les parties liminaires et annexes. Ces parties `<div0>` acquièrent ainsi leur autonomie par rapport au corpus du texte (par exemple, elles ne vont pas s'intégrer dans une numérotation des chapitres). Pour les divisions suivantes, la description par niveaux numérotés apporte une information de portée relative (on s'attend à ce qu'une `<div1>` corresponde à une grande articulation du texte, alors qu'une `<div7>` enfermerait le détail de quelques paragraphes). Mais la DTD est très contraignante, du fait qu'on ne peut « sauter » un niveau.

En regardant le détail des structures décrites, on relève aussi qu'une division peut être précédée par des paragraphes par exemple, mais ne peut être suivie que par d'autres divisions (du même niveau le cas échéant). Ce genre de contraintes atténue la robustesse du modèle, mais est révélateur des conventions rédactionnelles les plus généralement admises et pratiquées.

Des éléments que l'on choisit de distinguer

Alors que des éléments ont des analogies fortes sur le plan formel, la TEI ne les identifie pas sous un même élément, selon le principe vu précédemment pour les divisions, mais les présente comme des éléments de nature différente.

Premier exemple : la bibliographie d'un ouvrage n'est pas une liste. Toutes deux se présentent comme une succession d'items. Les différences se jouent sur de multiples plans. Dans la bibliographie, s'introduit la structure interne de chaque référence bibliographique. La bibliographie est dans un rapport global avec l'ensemble du texte pour lequel elle est définie. Elle se présente comme en bordure, pas incluse et interne au texte, mais ouvrant des passages vers des aspects non vraiment développés dans le texte. Elle a ses propres modes de lecture : elle peut être systématiquement consultée avant même de commencer la lecture du texte (et parfois elle conditionne cette lecture) ; elle peut être parcourue avec une attention spéciale aux auteurs ou bien aux titres, etc.

Les éléments enregistrés sous le modèle *phrase* constituent une famille considérable d'éléments ayant une analogie de comportement (qui motive la définition du modèle phrase en question) et pourtant distingués selon une véritable taxonomie (là encore dessinée par d'autres entités paramètre, traduisant d'autres modèles de contenu plus précis). Si l'on s'était arrêté à définir un élément phrase, avec différentes valeur indiquées par un attribut, on aurait simplement traduit la présence d'éléments saillants, comme semés çà et là au fil du texte. Le choix de leur identification cerne davantage l'interprétation (il faut décider de reconnaître, ou pas, un nom propre, une date), mais prépare la formation de cristaux (dans lesquels certains éléments précis sont attendus).

Un niveau charnière, et l'existence d'éléments à double niveau

Les entités paramètres définissant les *morceaux*, les *expressions* et les *intermédiaires* matérialisent une répartition des éléments en niveaux. Le niveau charnière correspond au paragraphe : au dessus se situent les morceaux, en dessous les expressions ; le paragraphe est lui-même un morceau qui ne peut contenir de morceaux. La suite de cette étude ne fera d'ailleurs que confirmer la pertinence de ce niveau charnière.

Il est donc curieux de voir la TEI en même temps affirmer cette séparation, et déclarer l'existence d'éléments intermédiaires qui s'en affranchissent, autrement dit qui se réalisent sur les deux niveaux, de part et d'autre du niveau du paragraphe. Pour mémoire, ces éléments intermédiaires sont en substance : les listes, les références bibliographiques, les annotations, les citations.

Est-il légitime de (ne pas) distinguer les listes en deçà et au-delà du paragraphe ? A l'intérieur d'un paragraphe, une structure de liste pourrait s'apparenter à une énumération (parataxe) ; alors

[11] Le mécanisme le plus juste pourrait être celui des *formes architecturales* : elles sont encore en marge de SGML et ignorées de la TEI-P3. Nous y reviendrons au sujet de la sémantique des DTD.

qu'une liste détachant chacun de ses composants par un retour à la ligne mettrait en valeur l'autonomie de chaque item et leur appréhension commune. Si l'on s'en tient à une définition typographique du paragraphe (son marquage par un retour à la ligne), rien n'interdit des cas limites, d'une énumération de noms avec retours à la ligne, ou inversement de l'indication méthodique de points développés au fil du texte (soulignée par des marqueurs graphiques de type *(i)* ..., *(ii)* ... ou des séries de connecteurs —*premièrement*, *deuxièmement*,... ; *d'abord*, *ensuite*,... ; *d'une part*, *d'autre part*— qui lexicalisent un séquencement ou une mise en parallèle de plusieurs points). La décision de reconnaître une structure de liste n'a alors rien d'évident, car elle tranche dans un continuum, du marquage lexical le plus fondu au marquage typographique le plus spécifique, de l'interprétation lissant les découpages ou focalisée sur un point isolé du reste, à l'interprétation attentive aux balancements et aux structures et se déployant à partir d'eux.

Les références bibliographiques font partie de ces cristaux de base dans une approche éditoriale. On les trouve d'ailleurs aussi bien dans l'en-tête documentant l'édition, que dans le texte lui-même. Une référence bibliographique constitue en quelque sorte un grain autonome d'information. Cette perception d'unité informationnelle explique qu'on puisse la reconnaître aussi bien enchâssée dans une phrase au fil du texte, que se détachant sous forme d'un bloc de texte. Son repérage échappe donc à la seule interprétation de la structure physique du document : la présence de guillemets, d'italiques ou de tirets ne sont pas suffisants. Interviennent aussi la morphologie des noms de personnes et la reconnaissance de maisons d'édition, la formulation typique des titres, la nature des informations attendues (date, pagination). Pourtant, cette unité d'information est particulièrement intéressante à identifier dans notre perspective, car elle est saillante et joue un rôle actif dans divers parcours de lecture professionnelle (attention portée surtout à l'auteur, ou au titre, ou à l'année d'édition, ou au nombre de pages, etc.)

Les annotations et les citations se présentent comme l'insertion d'éléments non rédigés par l'auteur. C'est en quelque sorte un marquage explicite de différentes contributions au sujet développé dans le texte. La distinction des points de vue représentés dans le texte est pertinente pour la description textuelle. Pour autant, les annotations et les citations explicites n'en donnent que les traces les plus superficielles, une grande part du travail interprétatif est ailleurs. Quelques remarques suffiront à montrer la difficulté d'exploiter ces marques et de discerner des points de vue. Une citation peut être explicitement marquée (par des guillemets, par un changement typographique) ; elle peut aussi être simplement annoncée par une formule conventionnelle (« pour reprendre les mots de Untel,... »). La citation peut être littérale, ou librement adaptée. Elle est plus ou moins consciente, car le rédacteur a en mémoire d'autres textes au moment où il écrit : à la limite, tout texte est un centon, il réécrit ce qui a été lu ailleurs, il fait écho à tout un ensemble de textes. Enfin, certaines citations et annotations viennent appuyer la thèse de l'auteur, d'autres indiquent de nouvelles perspectives qui ne sont qu'esquissées, d'autres encore sont en contestation et sont introduites pour alimenter le débat, etc. Il n'y a aucune nécessité de superposition des points de vue représentés, et de l'attribution d'un passage. L'auteur peut prendre successivement différents partis, et une citation peut s'inscrire parfaitement dans la thèse de l'auteur.

4. Autres DTD standard pour les documents textuels

a) La norme ISO 12083 (1994) : Electronic Manuscript Preparation and Markup

Présentation

Le modèle ISO 12083 est, depuis 1994, un standard international pour le codage de documents électroniques.

La TEI, très fouillée, embrassant la gamme la plus ouverte de propriétés textuelles, et se présentant comme un jeu de propositions souples, modulaires et évolutives, est souvent perçue comme destinée au milieu de la recherche, et adaptée aux corpus littéraires. Elle ne renie pas ses origines, née d'un rapprochement de trois associations partageant la préoccupation commune de l'application de

l'informatique dans le domaine des humanités : *Association for Computers and the Humanities*, *Association for Computational Linguistics*, *Association for Literary and Linguistic Computing*. Les chercheurs peuvent se baser sur la TEI pour élaborer un codage ajusté à tel besoin ou telle approche, tout en préservant des fondements communs pour l'échange et la mise en communs des corpus au sein de la communauté scientifique. La TEI a travaillé sur l'encodage de la poésie, du théâtre, des anthologies, et cela s'est concrétisé par des familles de balises dites de base. Elle a aussi étudié l'étiquetage linguistique des corpus (morphologie, syntaxe, sémantique), et l'enregistrement de ressources linguistiques (dictionnaires, terminologies). Bien que voulant couvrir « tous » les textes, elle a moins exploré la documentation technique, s'en tenant à ce qui entre en jeu dans le codage de ses propres documents de travail (par exemple pouvoir donner des exemples SGML dont les balises ne seront pas interprétées par le parseur), mais ne rentrant pas dans le détail d'autres aspects (par exemple, les tableaux : CALS[12] en fait une description beaucoup plus évoluée). Tout ceci a contribué à cette coloration « recherche littéraire » de la TEI.

Le modèle ISO 12083 se présente comme un modèle opératoire, bien arrêté, pour l'archivage de documents électroniques structurés. C'est une bonne décalque des informations de catalogage usuelles, complétée par un enregistrement des informations de structuration et de présentation communes dans les traitements de textes. Il est donc plus naturellement envisagé par les entreprises (comme EDF) comme format dans la gestion électronique de leur documentation.

ISO 12083 propose trois DTD, correspondant respectivement aux livres (y compris les rapports techniques), aux articles, et aux revues (vues comme des collections d'articles). Il ajoute une description des formules mathématiques.

Contenu et discussion

Ces DTD s'écartent de l'approche TEI sur plusieurs points :

- L'*en-tête* enregistre des informations de catalogage, mais ne prévoit pas de documenter le processus de codage. La dimension herméneutique et subjective de l'opération de codage, affirmée par la TEI, est ici ignorée : le codage est placé sous le registre de l'évidence. Le cas échéant, les spécifications de la norme doivent réduire les éventuelles variantes et hésitations interprétatives. L'objectif est d'avoir une description uniforme de tous les documents d'une base, et de lisser leurs singularités.
- Les parties de différentes natures sont codées comme des éléments différents : le *résumé*, l'*éditorial*, l'*index*, les *encarts*, etc. sont balisés comme tels. Ils précisent et enregistrent ainsi les « rubriques » attendues dans le type de document spécifié (selon la DTD choisie). Le corps du texte est toujours structuré comme une hiérarchie de parties. Il se divise en *chapitres*, eux-mêmes subdivisés en *sections* puis en *sous-sections de différents niveaux*.
- Les éléments décrivant le contenu textuel sont beaucoup moins nombreux que les éléments structurant les données de catalogage. Outre les parties évoquées ci-dessus, le modèle recense : les *titres* (seule la position en contexte peut distinguer le nom de la revue, le titre de l'article, l'intitulé de la section ou la légende de la figure) ; les *paragraphes* ; les *extraits de listing* de programme informatique (dans lesquels les balises SGML ne doivent pas être interprétées par le parseur) ; les *listes*, subdivisées en *items* ; la *mise en valeur* typographique (un attribut indique le moyen de rendement utilisé : gras, italiques, etc.) ; les *liens* et notamment les *annotations*. On ne retrouve donc pas l'éventail des éléments TEI, particulièrement au niveau des expressions en deçà du paragraphe.

[12] CALS est un acronyme pour *Computer Aided Acquisition and Logistics/Lifecycle Support*. Il s'agit d'un projet du Département de la Défense américain concernant la gestion de l'information technique relative aux systèmes d'armes.

b) GRAALDOC : Modèle des documents dans le cadre du Consortium GRAAL

Le contexte du projet GRAAL

Le projet EUREKA GRAAL (*Grammaires Réutilisables pour l'Analyse Automatique des Langues*) est un projet européen (France, Italie, Suisse, Finlande, Portugal, Grèce) récent (1992-1996), qui réunit des laboratoires de recherche, des industriels spécialistes du traitement automatique des langues, et de grandes entreprises utilisatrices d'applications faisant appel à ces technologies. Le projet GRAAL a pour objectif de construire des grammaires et des outils linguistiques génériques, destinés à être mis en œuvre et validés au sein de différentes classes d'applications. Les applications qui intéressent les partenaires sont notamment : l'indexation automatique de textes sur différents types de référentiels, l'extraction de connaissances en tant qu'aide à la constitution de terminologies ou de bases de connaissances, les interfaces en langage naturel pour la recherche documentaire ou l'interrogation de banques de données, la traduction assistée et notamment dans le cadre de langages contrôlés. Le projet s'inscrit dans la logique d'une ingénierie linguistique industrielle, au sens où il permet la réduction des coûts de réalisation d'une application, via le partage des outils et données, et des investissements, entre différentes classes d'applications ayant une composante « langage naturel » (Herviou, Monteil 1994).

La DTD GRAALDOC sert de dénominateur commun pour le codage des corpus soumis aux outils d'analyses. Elle est donc conçue pour accueillir une certaine variété de corpus, et plus particulièrement les documentations scientifiques, les notices techniques, et les documents rédigés, utilisés, ou/et circulant en entreprise. GRAALDOC correspond également par définition à format d'entrée adapté à des analyseurs grammaticaux généraux. Notre analyse porte sur la dernière version de cette DTD : la version 1.9, du 30 septembre 1994.

Analyse de la DTD

Le modèle s'articule autour de deux niveaux fondamentaux : le document (qui correspond au fichier de données) est une série d'*unités textuelles* ; et chacune de ces unités textuelles est subdivisée en *séquences*. Ce double découpage s'interprète ainsi : les unités textuelles sont les unités logiques représentées, par exemple un ensemble de documents référencés. Les séquences morcellent l'unité textuelle en unités élémentaires de traitement indépendantes, le cas type étant le découpage en phrases à soumettre à un analyseur syntaxique. Les unités textuelles sont donc les entités considérées, en fonction desquelles est pensé le contenu du fichier et le résultat de son analyse ; et les séquences correspondent à la décomposition des unités textuelles en vue d'une analyse, d'un traitement.

Cette décomposition selon deux paliers est comparable au format d'entrée de nombreux logiciels d'analyse textuelle. Dans le jargon d'ALCESTE par exemple, il s'agit d'*u.c.i.* (*unités de contexte initiales*) et d'*u.c.e.* (*unités de contexte élémentaires*), les premières traduisant le découpage logique du corpus par rapport auquel repérer les résultats, et les secondes définissant les voisinages pour le calcul distributionnel. Exprimant le référentiel des entités à caractériser en une suite d'unités adaptées au traitement, ce type de représentation sous forme de deux partitions imbriquées transcrit l'information de structuration exploitée par bon nombre d'applications de linguistique computationnelle et de lexicométrie : le modèle GRAALDOC reprend, englobe, généralise, et enrichit ce format.

Avec GRAALDOC, deux éléments s'intercalent dans ce double découpage sans reste. D'une part, une *liste de mots-clés* peut clore une unité textuelle en s'insérant à la suite des séquences. D'autre part, des *informations factuelles* peuvent apparaître n'importe où, sauf à l'intérieur d'une séquence (la séquence ne contient que du texte à analyser). Une information factuelle se présente comme un renseignement, un commentaire, qui ne fait pas partie du texte à analyser, et qui n'est pas structuré (une information factuelle ne contient aucun autre élément, donc en particulier une information factuelle ne peut se décomposer en d'autres).

Deux mécanismes sont introduits pour enrichir l'information sur les séquences. Le premier consiste à leur surimposer des formes structurelles.

- Il y a une structure de haut niveau, c'est la *section*, qui s'ouvre par un ou plusieurs *titres*. Titres mis à part (ils relèvent du niveau de base, cf. paragraphe suivant), les sections peuvent s'emboîter récursivement, ou contenir des structures de base. Les structures de haut niveau sont donc des structures facultatives, combinables entre elles, qui servent à organiser des éléments de base (ici les titres) et les structures de base.
- Les structures de base sont au nombre de deux : l'articulation *figure* / *légende*, et le *paragraphe*. Les structures de base constituent, avec les titres des sections, le niveau de base. La caractéristique du niveau de base est d'être un passage unique et obligé entre l'unité textuelle et la séquence (toute séquence relève d'un et un seul élément du niveau de base). Elles forment donc elles aussi une partition du document. L'articulation figure / légende a une sémantique bien particulière : la séquence fait partie de la légende d'une figure, dont le contenu serait généré à partir d'un autre fichier (indiqué). Le paragraphe, qui groupe une suite d'éléments (séquences ou / et listes), voit sa signification naturelle (la constitution d'unités cohérentes) s'affaiblir voire être neutralisée, puisque la présence obligatoire d'une structure de base conduit à son usage comme structure « par défaut ».
- La *liste* se présente comme une série d'*items* ; les items ne peuvent se composer que de séquences ou d'autres listes imbriquées. La liste a un statut particulier, elle intervient en deçà (à l'intérieur) des structures de base. Elle est considérée comme une « super-séquence », un groupe (une hiérarchie) de séquences, qui se comporte comme une séquence simple : elle apparaît aux mêmes endroits dans la structure.

Le second mécanisme pour enrichir l'information sur les séquences est l'élément *attribut de séquence*. C'est un élément vide, dont on peut insérer le nombre d'exemplaires que l'on veut en début de séquence, avant le texte contenu dans la séquence. Chaque attribut de séquence spécifie une paire *nom* - *valeur*, qui accompagne la séquence pendant le traitement. L'astuce de ce mécanisme réside dans le fait que l'on peut qualifier souplement chaque séquence, sans avoir à fixer dans la DTD (et par conséquent pour toutes les instances) un ensemble d'attributs et leurs domaines de valeur. En revanche, SGML ne contrôle ni la présence, ni l'absence, ni la validité des informations ainsi ajoutées.

Le modèle GRAALDOC utilise cinq attributs (outre l'attribut qui permet de référencer le fichier associé à une figure) :

- des *identifiants* facultatifs : pour le document, les unités textuelles, et les sections. Les identifiants servent de désignation pour la destination d'un lien : les éléments précédents apparaissent donc comme les unités que l'on peut avoir à désigner, à mentionner.
- une forme de *numérotation*, pas nécessairement numérique (on peut avoir des codes, des lettres, des étiquettes) : pour les séquences, puis pour les unités textuelles et les titres. Deux usages peuvent être pressentis : un repérage systématique et ordonné suivant le déroulement du document (ce qui convient bien aux séquences et aux unités textuelles) ; l'enregistrement de la numérotation des parties d'un texte (pour les titres). Ce deuxième usage rattache au titre une information qui appartient en fait à la partie qui correspond à la portée du titre. C'est certainement la difficulté d'avoir toujours cette information de portée qui conduit à introduire ce biais, avec celui d'avoir plusieurs titres en ouverture d'une section.
- la *langue* : pour le titre. L'utilisation privilégiée de corpus extraits de fonds documentaires d'entreprise ressort très nettement : les seules variations de langue prévues sont celles du titre original, en langue étrangère.
- le *type* : pour l'unité textuelle. Ceci renverrait à une typologie, complètement indéterminée dans la DTD.
- une *marque introductive* : pour les informations factuelles, les paragraphes, les items. C'est une mention qui introduit l'élément, les usages prévus étant un symbole (un tiret, un numéro, etc.) ou un mot-clé (« Note : », « Attention : », etc.).

Discussion

Après l'approche éditoriale de la TEI, et l'approche documentaire de la norme ISO 12083, GRAALDOC apparaît nettement marquée par sa conception pour le traitement automatique du langage naturel, et particulièrement l'analyse syntaxique. La DTD GRAALDOC est (ouvertement)

« orientée traitement ». La structure qu'elle décrit gravite entièrement autour des deux pôles que constituent l'unité textuelle et la séquence. Et la séquence se mesure à l'aune des grammaires : son domaine se situe à l'intérieur d'un titre ou d'un paragraphe.

Le découpage en séquences influence tout le modèle. Il est symptomatique par exemple qu'une information sur le texte, formalisée par une paire attribut-valeur, n'a de rattachement prévu qu'au niveau des séquences ; le seul autre moyen d'introduire une information est de recourir à l'élément information factuelle, qui pour être très souple n'a aucune structuration interne. La séquence est le niveau qui départage sans reste le domaine des attributs de séquence et celui des informations factuelles. On note également le lien déterminant entre listes et séquences : chaque item d'une liste est directement une série séquences ou de listes. Il n'y a jamais de paragraphes au sein d'une liste.

La DTD GRAALDOC fait preuve de pragmatisme : elle est prévue pour s'adapter à des structurations grossières comme à des structurations plus raffinées. Par prudence, elle prévoit la déclaration de plusieurs titres pour une même section : le traitement n'est pas toujours capable de repérer une zone de texte associée à chaque niveau de titre. La technique des attributs, pour enregistrer les informations complémentaires et les grilles d'analyse particulières, est parfaitement mise en œuvre. Une place est ainsi préparée pour l'indication du genre textuel... la dangereuse question de la typologie des genres étant sagement laissée à la discrétion de l'utilisateur.

5. La sémantique des DTD

a) L'étiquetage sémantique

Identification et homologation de référents

La DTD peut prévoir, en définissant les balises correspondantes, de reconnaître certaines mentions dans le texte : les dates, les personnes, les lieux... Elle peut accentuer son optique dénotationnelle en explicitant le lien entre les différentes mentions d'une même réalité, d'un même référent, par exemple en rapprochant :

Jean-Luc Sanson... M. Sanson... Le chef du Groupe TTI... il...Son projet pour le groupe...

Comme on le constate vite, cela peut aller des variantes de formulation aux anaphores les plus discrètes. Nul ne conteste que procéder à ce genre de repérage automatiquement serait une prouesse : le calcul propose, et l'homme valide, on est donc tout au plus dans du semi-automatique. Mais même avec la plus grande expertise humaine, la faisabilité d'un tel codage est discutable, pour une pertinence linguistique douteuse. La décision de repérer et d'identifier une certaine entité est une opération interprétative, et ne relève ni de l'évidence, ni de la permanence, ni de l'universalité, ni d'une délimitation franche ou d'une extraction. L'étiquetage induit une sémantique atomisante et figée. Or la dynamique interprétative prend par exemple tout son sens de discernement actif dans l'ambivalence Dr Jekyll / Mr Hyde[13] : toute identification statique serait réductrice... et bien des romans policiers y perdraient leur charme. Ou encore, est-il sensé de systématiquement projeter les indications temporelles de la fiction sur le registre des dates historiques ? Leur valeur sémantique peut être bien ailleurs : romantisme de l'automne, solennité du moment... A propos du codage des dates, le commentaire ingénu donné dans la TEI-Lite[14] est exemplaire : on perçoit la double difficulté de voir comme une date toutes sortes d'indications temporelles, puis de vouloir rapporter ces occurrences à un calendrier rigoureux.

> L'attribut *value* indique une forme normalisée pour la date ou l'heure, au moyen d'un format reconnu tel que celui qui est prescrit par la norme ISO 8601. Les dates ou les heures partielles (par exemple « 1990 », « septembre 1990 », « autour de midi ») peuvent habituellement être exprimées en omettant simplement une partie de la valeur donnée ; ou bien, les dates ou les heures imprécises (par exemple « début août », « entre dix et douze heures ») peuvent être exprimées comme une plage de dates ou d'heures. Si l'une ou l'autre extrémité de la plage d'heure ou de date est connue avec

[13] On aura reconnu le(s) héros de Stevenson.
[14] TEI-Lite : version allégée de la TEI (*Text Encoding Initiative*).

certitude (par exemple, « avant 1230 », « quelques jours après Hallowe'en »), l'attribut *exact* peut être employé pour le préciser.

Exemples :

```
<date value='1980-02-21'>21 Feb 1980</date>
<date value='1990'>1990</date>
<date value='1990-09'>September 1990</date>
Given on the <date value='1977-06-12'>Twelfth Day of June in the Year
of Our Lord One Thousand Nine Hundred and Seventy-seven of the Republic the
Two Hundredth and first and of the University the Eighty-Sixth.</date>
[...]
<p>C'était une belle matinée de la <date value='1323-11'>fin
novembre</date> ...
```

(Burnard, Sperberg-McQueen 1996, §11.2)

La plupart des types d'éléments repérés correspondent à des entités désignables par un nom propre (lieux, personnes, organismes, etc.) ou du moins dotées d'une dénomination « canonique » (date exprimée dans un calendrier de référence). En poursuivant la même logique, et en l'étendant au lexique d'un dictionnaire de langue, on identifierait des occurrences en relation de synonymie. Les difficultés mentionnées précédemment n'en seraient qu'amplifiées : non localisation du sens au palier du mot, non équivalence des occurrences (manifestée par la non transitivité des relations),...

Application d'un référentiel : annotation partielle

Si l'on dispose d'un dictionnaire de référence, on peut vouloir systématiquement assigner une signification à chaque unité découpée dans le texte. La signification peut se limiter à un ensemble de traits généraux et fondamentaux, apportant l'information suffisante pour des traitements ultérieurs, sans prétendre épuiser le sens des mots. L'utilisation de ces informations prend la forme de patrons, de restrictions de sélection, etc.

Les annotations introduites correspondent à une grille particulière d'analyse. Il serait illusoire de croire cerner ainsi la sémantique *véhiculée* par les unités du texte, de viser à *extraire* le *contenu* de ce *gisement* ou de cette *mine* d'informations que constitue une base de documents. Pour prétendre à une certaine pertinence, le codage avec annotations sémantiques doit être pensé et situé dans un cadre applicatif.

Là encore, l'étiquetage par des balises est brutal (un item est ou n'est pas étiqueté, reçoit ou ne reçoit pas telle identification, commence et finit à tel et tel point de la ligne textuelle). C'est se placer dans une interprétation sous le régime de l'évidence, voire dans certains cas de l'univocité (si certains choix de codage sont exclusifs).

Bilan

D'une manière générale, ces codages sont lourds, ils multiplient les balises. Ils sont délicats à réaliser et à formaliser, car ils réifient un travail interprétatif. Le résultat d'un traitement tout automatique, approximatif, partiel et relatif, peut trouver sa place comme étape dans une chaîne de traitements, mais se justifie moins comme version à part entière, édition électronique, du texte original.

Pour un format général d'entrée, on peut préférer s'en tenir à un codage minimal, qui ne transcrit que les marquages déjà explicites (mise en forme) et potentiellement significatifs pour l'application (retenir les fins de paragraphes, pas les fins de lignes par exemple). Et on ne cherchera pas toujours à résoudre immédiatement leur équivocité.

b) Les modèles orientés contenu

Les « DTD sémantiques »

Le codage le plus riche est conçu en termes de contenu, et non de structure apparente.

L'analyse d'un ensemble de documents homogène, correspondant à un type, permet de se doter d'une DTD « sémantique » : chaque rubrique attendue est prévue et identifiée. Par exemple, la zone de texte correspondant à une bibliographie ou à une citation est reconnue comme telle. La DTD est dite sémantique par opposition à une DTD qui ne donne que des indications de forme : « ceci est

un intertitre » est une indication de forme, mais « ceci est la partie qui décrit le but de l'action » est repérage sémantique, le contenu de la partie est en quelque sorte interprété.

L'appellation *DTD sémantique* est une manière commode de parler, ce n'est pas un terme consigné dans la norme SGML. Il serait d'ailleurs difficile de lui donner une définition formelle. Pour un dialogue au sein de la communauté SGML, on préférera une formulation analytique plus explicite, par exemple en opposant les *DTD orientées contenu* et les *DTD orientées structure logique documentaire*. En revanche, dans un petit groupe de travail comme celui d'EDF, *DTD sémantique* est un raccourci de langage, qui a en outre l'avantage de souligner l'enjeu de cette approche : un gain de signification apporté par le codage.

Concevoir les DTD en ce sens est en fait le B.A. BA du sain usage d'un codage structuré. La règle d'or : éviter de s'en tenir au typographique (*titre 1*, *titre 2*, *italiques*) quand on peut reconnaître l'interprétation qui serait conventionnellement attribuée -évidente pour le lecteur, mais invisible pour la machine si on ne lui dit pas !- : *résumé*, *titre de revue*, etc. Les formateurs à SGML, et les praticiens chevronnés[15], ne manquent pas d'insister sur ce point : car ils savent qu'il faut tourner le dos à ce qui était devenu naturel avec l'usage des styles dans les traitements de texte.

Une DTD sémantique est plus puissante qu'une DTD calquée sur la structure logique, elle-même partiellement dérivée des marques de présentation données par la structure physique et la typographie[16], au sens où un même élément typographique peut être la représentation de différents éléments sémantiques. C'est un travail d'interprétation qui choisit de reconnaître, en faisant appel à de multiples indices contextuels, une valeur sémantique différente à différents éléments de même apparence. Un document dont on a reconnu et étiqueté la sémantique des constituants, se prête ensuite à de multiples projections ou vues : affichage de certains éléments, mise en valeur tel et tel élément de telle ou telle façon. La qualité et la portée des traitements automatiques sont incomparablement augmentés (Futtersack 1995).

Discussion : codage informatif, codage réducteur, un point de vue

Si donc le codage sémantique vise à reconnaître, identifier, déterminer, la nature de chaque élément du texte, les éléments que le modèle prévoit sont comme autant de rubriques, que le texte vient remplir.

A la limite, tout se passe comme si on transposait le texte en données factuelles, puisque articulé en informations élémentaires. Vu à travers les mailles du codage sémantique, il reste bien *du* texte, mais plus directement *le* texte, car il est mis en lambeaux pour être réparti entre les différentes rubriques. Sur l'axe de l'opposition *données textuelles* vs *données factuelles*, on s'éloigne des premières pour se rapprocher insensiblement des secondes.

Une inquiétude peut poindre : en cernant l'identité de chaque particule du texte, n'est-on pas en train de bafouer la liberté et l'équivocité naturelle des textes ? C'est anti-herméneutique, dans la mesure où l'on force une ou quelques interprétations : « c'est comme cela qu'il faut voir ce texte ».

Relativisons un peu : en ce qui concerne la libre variation des textes, les textes prennent place dans des pratiques de production et de lecture, et leur abord est médiatisé par les régularités d'un genre. Par exemple, considérer tel texte comme une recette de cuisine permet d'y reconnaître des éléments « ingrédients », ou tel autre comme une pièce de théâtre autorise à en recenser les personnages. Le codage sémantique peut être utilisé pour transcrire l'éclairage d'*un* genre. Le tout est de le savoir, et de l'utiliser à bon escient. Plusieurs codages d'un même document peuvent aussi coexister pacifiquement, sans que jamais soit close la possibilité de poser la trace d'une nouvelle lecture, d'un nouveau codage.

---

[15] Je remercie vivement ceux qui m'ont aidé à faire les premiers pas dans le monde SGML, et que j'ai ici bien présents à l'esprit. Du côté des formateurs, Valérie REINER (AIS / Berger-Levrault) ; et en tête de file des praticiens chevronnés, Jean-Luc SANSON, chef du Groupe *Technologies du Traitement de l'Information*, pilier du projet *Bibliothèque Electronique*, inlassable avocat de la documentation structurée à la DER d'EDF.

[16] On peut concevoir des DTD traduisant la structure physique des documents (pages, blocs de texte, tableaux, illustrations). Une telle DTD peut encoder les résultats d'une reconnaissance optique de document, avant que d'autres traitements n'en déduisent une structure logique (enchaînement des blocs, reconnaissance des intertitres, etc.) (Lefèvre, Reynaud 1993).

c) Fragments, îles : la notion d'unité d'information mise en valeur par XML

XML trouve pour le moment l'essentiel de ses applications dans les échanges de transactions : par exemple, commandes et factures commerciales. Les données correspondent à des champs de bases de données : elles sont extraites d'une base et transmises pour compléter une autre base. Plus exactement, les données sont des compositions de champs répondant à une requête. Un tel échange d'informations s'analyse en unités d'informations, elles-mêmes construites à partir d'éléments (extraits d'une base).

XML est alors basé sur la notion de *fragments* ou d'*îles*[17]. Un fragment ou une île est une unité d'information, sémantiquement autonome, et syntaxiquement bien formée et complète. C'est un morceau isolable de l'échange de données. La décomposition d'un échange en fragments ou îles est essentielle au fonctionnement de XML : elle assure la possibilité d'effectuer sur chaque transaction des analyses et validations syntaxiques, indépendamment de tout autre contexte ; une transaction est à elle-même son propre contexte, elle crée son contexte implicite.

Le terme de *fragments* est surtout utilisé en France, à l'initiative de François Chahuneau, Directeur d'une entreprise de premier plan dans la réalisation d'outils logiciels autour de XML. Dans les milieux anglophones, le concept analogue apparaît sous le nom de *data islands*, sur la page Web du premier éditeur de logiciels au rang mondial. De fait, il n'y a pas de terme officiel, car on ne peut donner une définition claire et formalisée valide au plan général : on peut en effet avoir des petites structures syntaxiquement complètes, qui n'ont en fait pas d'autonomie sémantique, et sont toujours liées à une structure supérieure qui les intègre. La définition des fragments s'avère relative : ils se définissent pour une instance XML donnée, ou dans un contexte applicatif déterminé.

Le caractère imagé des deux dénominations relevées souligne des aspects différents. *Fragment* est ce que l'on peut détacher d'un tout, c'est une division, une unité de description, autonome mais relative à un tout. *Ile* insiste particulièrement sur la délimitation, qui cerne une unité complète au sein d'un échange comportant d'autres données et informations.

D'un point de vue général, par delà les extractions de bases de données, le fragment peut être compris comme l'unité potentiellement adressable (*i.e.* que l'on peut vouloir désigner) qui se fait jour dans une pratique[18]. Pour un système de consultation de documentation technique, cela pourrait correspondre exactement à tout ce qui est muni d'un titre, d'une légende ou d'une numérotation, qui justement sert à désigner : parties de divers niveaux, figures et tableaux, définitions, équations. La tendance est à l'homogénéisation des longueurs des sections : les titres réguliers et fréquents favorisent la lecture rapide, en servant de points d'appui. Un fragment se comporte sémantiquement comme une unité de contexte, et comme une unité du parcours de lecture. La structuration en fragments relève d'un découpage sémantique et interprétatif.

d) La structure fait sens : les formes architecturales

Les formes architecturales[19] ont été brièvement mentionnées lors de la présentation générale de SGML. Ce sont des modèles de structure de données, pour la description d'éléments et de groupes d'attributs, qui servent de briques dans l'élaboration des DTD. En somme, une forme architecturale déclare un type d'unité, un *objet* au sens informatique du terme. Une forme architecturale définit : *(i)* les éléments qui entrent dans la composition de sa structure interne ; *(ii)* l'organisation de ces éléments ; *(iii)* les motifs d'attributs qui la qualifient globalement, ainsi que ceux qui qualifient ses

[17] C'est une réflexion avec Véronique JOLLY et Jean-Louis VULDY, du Département SID d'EDF-DER, qui a mis en évidence la portée sémantique du concept naissant de *fragments*, dans la perspective ici adoptée.

[18] Cette proposition de définition heuristique, et l'illustration qui suit, sont redevables à Jean-Louis VULDY (EDF-DER, Département SID).

[19] C'est Jean-Luc SANSON, puis Véronique JOLLY, (tous deux au Département SID d'EDF-DER), qui ont signalé et éclairé la valeur des formes architecturales, dans le cadre de cette réflexion sur la sémantique portée par la structuration.

composants. Dans le cadre d'une application, chaque forme architecturale peut se voir assigner *(iv)* des traitements qui lui sont propres et reflètent le sens qu'on lui accorde[20].

Par exemple, les liens hypertextes peuvent être décrits par plusieurs formes architecturales. On peut ainsi distinguer des liens qui constituent un simple renvoi, des liens qui se groupent en un faisceau de renvois, des liens qui s'enchaînent pour former un parcours, etc. Chaque cas peut avoir ses propres attributs : par exemple, une indication d'ordre structure certains ensembles de liens. Et un navigateur (*browser*) peut représenter et exploiter différemment les différentes natures de liens. On peut imaginer : l'affichage contextuel, à la demande, d'une liste d'alternatives ; ou le déroulement d'un parcours guidé prédéfini.

La notion de forme architecturale a été définie et formalisée dans le cadre de la norme HyTime. HyTime utilise la syntaxe SGML dans le cadre du codage de données multimédias. Ceci nécessite la prise en compte des contraintes de synchronisation, de la diversité des modes de manipulation, etc. Les formes architecturales constituent une technique pour proposer des modules de codage de référence, significatifs et validés, qui s'intègrent de façon souple pour construire les DTD adaptées à chaque contexte.

HyTime met en œuvre le concept de forme architecturale qu'il a introduit, en en proposant un catalogue. Ces formes architecturales condensent une connaissance des objets et des mécanismes pertinents pour les applications *multimédia*, mais finalement très peu décrivent des aspects spécifiques des *textes*. On en regrette d'autant plus le peu d'échos reçus par le concept de forme architecturale hors de HyTime. SGML aurait pourtant réintégré le concept, ainsi généralisé, dans le cadre d'une annexe.

Ainsi, la dernière version de la TEI (TEI-P3) ignore les formes architecturales. Cela peut se comprendre, dans la mesure où la norme HyTime a été publiée en 1992, alors que les travaux de la TEI ont été initiés plusieurs années auparavant (fin 1987). La TEI s'en tient aux entités paramètres pour factoriser des comportements communs et dessiner ainsi des familles d'objets : cependant, l'entité paramètre n'a qu'un statut trop vague de notation, et ne peut constituer elle-même une unité, elle n'a pas la consistance d'une forme architecturale. Mais l'usage régulier des entités paramètres, et l'orientation modulaire générale affirmée, préparent la TEI à une description plus riche en termes de formes architecturales.

Les exemples de formes architecturales propres aux textes sont rares. Dans HyTime, on peut relever : les modèles de liens ; les répertoires de formes lexicographiques (un attribut indique la règle qui préside à l'ordre des entrées). Si l'on y réfléchit, d'autres structures textuelles pourraient se prêter à une interprétation en termes de formes architecturales : les différents types de listes (liste non ordonnée, liste numérotée, liste correspondant à un ensemble d'alternatives spécifiées, etc.) ; un index ; l'association d'un titre et de la partie à laquelle il se rapporte.

Ainsi, définir une forme architecturale pour le chapitre ou la section serait un moyen de reconnaître l'incidence de cette structuration, par exemple dans un séquencement de la lecture, tout en gardant une approche orientée contenu. Pour les textes d'ordonnancement EDF par exemple, la DTD expliciterait la présence et l'enchaînement d'un élément *Contexte* du projet, d'un élément *But* du projet, tout en indiquant que le *Contexte* prend la forme architecturale d'une *forme-architecturale-de-section*, le *But* également, etc. La structure sémantique n'est plus pensée de façon antagoniste à la structure logique documentaire, mais ces deux structures sont mises en correspondance. Chaque structure contribue dans son registre à l'organisation du document. Les formes architecturales permettent par exemple d'enregistrer de la façon la plus efficace l'information de présentation qui relève de la structure logique.

Notre recherche sur les structures impliquées dans la textualité va dans le sens d'une recherche de formes architecturales sous-jacentes à la description de textes : quels motifs d'éléments font sens, quelles incidences différentes ils ont pour un « traitement » de lecture et d'interprétation.

[20] On reconnaît là les *méthodes* d'une *conception orientée objet*.

C. FORMAT DES TEXTES POUR L'APPLICATION DECID

1. Conception du modèle

a) Orientations fondatrices

Proposition d'une structuration minimale, point d'équilibre entre robustesse et informativité

La recherche d'une DTD pour décrire différents documents s'apparente à la recherche du plus grand dénominateur commun.

Le système DECID doit analyser tout texte qui lui est soumis pour se construire une représentation interne utilisée dans le calcul des rapprochements texte-texte. Le format de réception ou d'entrée des textes délimite l'information qui peut être utilisée dans les traitements.

Quand se mêlent les genres de documents (descriptif d'activité, pages Web, courrier d'annonce de séminaire, etc.), même en s'en tenant aux documents de travail à dominante scientifique et technique, seule une DTD « minimale » convient, pour assurer une description régulière et homogène. Elle s'en tient au niveau de l'expression.

Le pari est ici de savoir tirer parti au mieux de tout texte soumis à DECID, sans imposer de convention particulière de mise en forme, mais en interprétant souplement tout ce qui peut « faire signe ».[21]

Point d'appui : incidences de la présentation sur la lecture

La DTD se base sur ce que l'on peut percevoir dans la mise en forme du document, qui guide la lecture et en ce sens intervient dans la représentation que l'on se fait du texte.

Une illustration « historique »

L'importance de la présentation au plan du sens que le lecteur donne au texte a déjà été reconnue. En 1981, Daniel Hérault appelle de ses vœux un *module optique*, en amont d'une analyse automatique des textes. Il s'agit de travailler sur

> [une représentation du texte] strictement *optiquement* équivalente à la présentation originale. Toutes variantes typographiques, ou même celles que permettent maintenant les machines à écrire les plus modernes, sont soigneusement conservées, que ce soit dans le choix des différents corps pour la hiérarchie des titres, sous-titres et intertitres, que ce soit encore dans le repérage des « mises en relief » par le biais de lettres italiques, grosses, scriptes ou espacées, que ce soit enfin pour tout ce qui concerne la mise en page comme les passages à la ligne, la création de paragraphes ou l'organisation des notes et des indications de renvoi leur correspondant. Toute cette énumération est loin d'être complète [...]. (Hérault 1981, p. 92)

Les trois temps qui rythment l'énumération détachent selon nous trois formes d'articulation du texte : les oppositions titre *vs* texte courant, le soulignement ponctuel d'éléments au fil du texte, le découpage en paragraphes (sans y mêler les notes et renvois).

Poursuivons notre lecture : quelles structures retiennent plus particulièrement l'attention du concepteur d'un « module hypersémantique » ?

---

[21] (Tazi 1988) cherche également à reconnaître des *Fonctions de Structuration Textuelle*, mais il le fait à l'aide d'un parseur, pour un document se conformant à une *grammaire* préétablie, et dactylographié suivant certaines *conventions* de présentation données (par exemple, un titre n'est reconnu que s'il est souligné ou/et numéroté). Son approche est plutôt normative, et la nôtre se veut descriptive.
(Pascual 1991) reprend ces *Fonctions de Structuration Textuelle* (*diviser*, *dénombrer*, *lister*, *titrer*, *détacher*, *énoncer*) et les analyse de façon systématique par des 'métaphrases', dans une perspective encore différente de la nôtre, la génération de textes. Son travail est néanmoins pertinent pour nous, pour la formalisation des FST à laquelle il aboutit.

> Il est bien connu que la plupart des textes non-littéraires ont une profonde vocation didactique et que cette vocation impose une forme de présentation qui permette d'insister sur des points précis, qu'il s'agisse de définitions, de propriétés ou de remarques. Ces « mises en relief » ne peuvent être produites que par les artifices typographiques classiques : utilisation de l'italique, du gras, du changement de corps, des guillemets, des tirets, de la hiérarchie des titres, des index alphabétiques, etc. ... [...] Nous nous sommes limités à la manipulation de l'ensemble des titres, sous-titres, inter-titres, qui regroupe aussi les titres marginaux, ceux des tableaux, des figures, etc. (Hérault 1981, p. 110)
>
> Les mises en relief typographiques les plus intéressantes sont évidemment celles qui permettent d'*isoler* un mot ou un syntagme plus ou moins complexe. Cette notion d'isolement est assez difficile à définir nettement et, pour le moment, nous avons choisi simplement d'effectuer le partage à partir du nombre de mots. Au-delà d'une dizaine, il ne semble pas qu'il s'agisse d'un réel isolement. (Hérault 1981, p. 111)[22]

Si ces structures sont enregistrées, c'est qu'elles sont significatives et entrent en ligne de compte dans le traitement : Daniel Hérault les utilise en fait comme filtre, pour focaliser son traitement sur ces zones, où trouver le plus sûrement les « objets » du texte.

S'appuyer sur la présentation à l'ère des documents structurés

La lecture des lignes qui précèdent pourraient raviver quelques inquiétudes, quant au bien-fondé d'explorer la présentation du texte, quand les codages permettent aujourd'hui d'expliciter toute cette information, et même davantage (la présentation est *générée* à partir des balises du texte). Résumons la manière dont nous concevons une lecture basée sur la présentation, et les raisons et le sens d'une description de l'apparence des textes.

Premier point, la présentation est le dénominateur commun à tous les documents que DECID doit traiter, issus de diverses DTD (HTML, *Livre Electronique*), voire sans DTD (texte ASCII entré au clavier).

Second point : pour un document structuré, c'est surtout la manière d'afficher ou de présenter le contenu de tel élément qui guide son interprétation, plutôt que son nom (généralement masqué). Pas plus que le formalisme SGML ne peut contrôler le contenu textuel des éléments (et vérifier leur adéquation à l'élément utilisé pour le codage), le nom de la balise n'a aucun moyen d'imposer une interprétation. En pratique, on observera que le « détournement interprétatif » d'éléments HTML est courant sur les pages Web : il faut aller consulter les *sources* du document pour s'apercevoir qu'un petit copyright en bas de la page est codé comme un intitulé (*heading*), qu'une manière commode d'obtenir la large marge a été de déclarer tous les paragraphes comme les éléments d'une liste après plusieurs niveaux d'imbrication, etc. Les usages non conformes des balises sont plus ou moins conscients : une page Web est mise au point en fonction de son affichage, et quasiment jamais à l'aide d'un parseur ; c'est par sa présentation et l'effet rendu qu'elle est validée, et non par le respect d'une sémantique –ni même d'une syntaxe– des balises (Amitay 1997, §2).

La lecture d'un texte n'est pas une pure lecture de mots. La mise en forme matérielle du texte entre de plain-pied dans une sémantique linguistique, différentielle, unifiée, et interprétative. D'abord, la lecture interprète des effets de présentation, plus ou moins ouvertement conventionnels, en *interaction* avec le sens issu des unités linguistiques : présentation et unités linguistiques se rejoignent dans la matérialité des signes. Puis, la présentation fait sens par ses différences : c'est le *principe de contraste* indiqué par (Virbel 1987, §2.5, p. 90), car les identités ou les différences systématiques sont comprises les unes comme des instructions d'équivalence de valeur, les autres comme l'affirmation d'une distinction significative. Ensuite, l'apparence du texte guide l'interprétation des unités de tous niveaux, à savoir de l'ordre du mot, du paragraphe, du texte, voire d'un intertexte au sein duquel le texte se situe. Enfin, la présentation fonctionne comme une invitation à telle ou telle lecture, elle donne des points d'appui sans tout expliciter, elle module et oriente le parcours du texte et la construction du sens que l'on en retient.

[22] Les guillemets et les tirets « jouent en réalité plusieurs rôles : ils peuvent, par exemple, ne concerner qu'un seul mot ou encadrer une partie plus ou moins longue du texte. Il est donc indispensable de disposer d'un module-guillemet et d'un module-tiret, qui soient capables d'analyser correctement toutes les situations qui peuvent se produire (et qui sont beaucoup plus nombreuses que celles qui viennent d'être citées). » (Hérault 1981, p. 110)

L'influence générale et déterminante de la présentation étant établie, c'est elle qui fonde le choix des structures à décrire, en fonction de leur rôle dans la construction d'une lecture. Pour autant, on ne s'interdit pas (bien au contraire !) d'utiliser la forme structurée des documents quand on y a accès. Ce qui importe alors, c'est d'interpréter les balises non simplement en fonction de leur nom, mais aussi en fonction des usages auxquels elles se prêtent effectivement, notamment dans les modes de restitution qui leur sont associés pour les applications d'affichage (consultation) et d'impression (lecture).

b) Structures textuelles retenues

Zones

Les zones sont importantes à enregistrer car elles interviennent dans le contrôle de la propagation des sèmes, par les frontières qu'elles posent.

Le fait qu'on se situe dans le discours des documents d'entreprise a une incidence directe sur les types de zones définissables et pertinentes pour notre application. En effet, s'il s'agissait de caractériser les textes d'un corpus littéraire, certains romans, tendus dans un seul souffle, défieraient les tentatives de parcellisation. C'est encore un rappel que la structure choisie, bien que minimale et générale, n'a rien d'universel.

Le découpage de base

Ce qui est directement enregistrable, ce sont les *alinéas*, à savoir les portions de texte délimitées par un retour chariot.

Pour autant, les frontières sémantiques recherchées sont celles des *paragraphes*. Le lecteur peut par exemple saisir comme une unité une suite d'alinéas très courts, qui ne sont pas perçus comme ayant chacun la même autonomie qu'un paragraphe, et participent alors d'une même zone d'influence. Les alinéas sont destinés à servir de briques de base dans la construction des paragraphes. La mise au format des données se contente d'enregistrer les traces que constituent les alinéas ; l'interprétation des alinéas en paragraphes, et la construction de ces derniers, se fait au cours de l'analyse même du texte[23].

Le découpage de base a ceci de caractéristique, qu'il morcelle le texte sans recouvrements et sans reste (les mathématiques appellent cela une *partition*). Or la lecture, le déroulement des lignes, impriment une continuité. Les frontières qui rythment le cours du texte ne sont pas imperméables : il y a des jeux de diffusion et de proximités. Le codage ne fait qu'enregistrer les indices qu'utilisera le traitement pour rendre compte de ces effets, dynamiques, de perception de *passages*.

Les parties à fonctionnement méta-textuel

Les exemples les plus évidents de ce que nous appelons parties à fonctionnement méta-textuel sont les titres et intertitres, et les rapports d'inclusion et de portée (un résumé vis-à-vis du texte qu'il représente, par exemple).

L'information enregistrée est double. D'une part, la *délimitation* d'une zone, dans laquelle se trouvent les éléments qui entrent en rapport. D'autre part, la dénivellation de cette zone sur deux niveaux : le niveau courant, de référence, et le niveau *méta*, qui se présente comme une image condensée du premier.

---

[23] Dès l'examen de cette première structure textuelle, il apparaît que même les structures les plus « objectives » – le cas ici des paragraphes, par contraste notamment avec les surlignages présentés plus loin– ont leur part d'interprétation subjective. Une plongée rétrospective, dans l'histoire du document écrit –l'encre rouge des rubriques, les mots implicites de la *scriptio continua*–, ne ferait que confirmer cette subjectivité généralisée des structures dont les traces semblent aujourd'hui si nettes.
Selon leur lien à la mise en forme matérielle du document 'originale', les structures s'échelonnent entre le pôle de l'objectivité et le pôle de la subjectivité. Dans leur versant objectif, les structures sont fixées, la lecture doit les reconnaître, les découvrir ; du côté du subjectif, les structures sont conçues pour proliférer, comme autant de traces d'opérations interprétatives, formes créées par la lectures.

Le niveau *méta* est le lieu privilégié des formulations synthétiques. Il souligne une isotopie et suggère la lecture selon cette isotopie. Il peut ainsi s'incorporer aux attentes du lecteur. Cela peut se traduire dans le traitement en leur conférant une persistance, une présence implicite et sous-jacente, sur toute l'étendue de la zone concernée.

Les parties à fonctionnement infra-textuel

C'est le registre des commentaires, des notes de bas de page, des annotations marginales, des développements annexes, dans leur rapport au texte de rattachement. C'est également celui de bon nombre de liens hypertextes, dont la fonction est d'ouvrir sur un autre document, une autre unité autonome. Il y a là bien sûr une extrême diversité de modes de relation au texte central : cependant, rien n'assure que l'on puisse en faire une typologie satisfaisante. Leur point commun, qui est d'apporter un éclairage spécifique complémentaire, donne déjà une finesse de description intéressante pour un traitement automatique.

Comment décrire et utiliser les parties à fonctionnement infra-textuel ? Ces parties s'affichent comme des digressions ou des compléments. Leur rôle dans la mise en relation de deux documents est ambivalent. Elles ne sont pas au cœur du texte, et en cela ne pas les retrouver dans les deux documents ne doit pas pénaliser le rapprochement. Toutefois, des rapprochements intéressants sont aussi susceptibles d'être générés à partir du texte dans le contexte et l'éclairage particulier d'un élément *infra*.

Les éléments *infra* se comporteront donc dans notre traitement comme des « plus ». L'introduction du niveau *infra* ajoute un effet de prisme, en diffractant des colorations particulières du texte. Il impulse aussi une dynamique centrifuge : les éléments *infra* font la passerelle entre les éléments centraux du texte et d'autres aspects en relation.

Dans un premier temps, on pourra considérer que du point de vue du codage, la description des parties infra-textuelles est similaire à celle des parties supra-textuelles. Elle comporte la délimitation d'une zone d'influence, et dans cette zone le marquage des éléments en rapport infra-textuel avec le reste. Le *point* d'ancrage éventuel d'une partie infra-textuelle (par exemple un numéro de note, une ancre HTML) est une indication pour la détermination d'une *zone* de rattachement, qui étend son influence avec une portée plus ou moins floue. Cette zone est en définitive délimitée par un choix, jugé acceptable et approximatif.

Les structures quasi-parallèles

Les listes et les tableaux disposent conjointement une série d'éléments à l'œil du lecteur. Ces structures sont des opérateurs à la fois conjonctifs et disjonctifs[24].

En effet, les items énumérés par une liste, ou les cellules aux croisements des lignes et des colonnes, sont perçus comme disposés dans un même plan, appartenant à un même système, composant un tout. La structure quasi-parallèle (ré)unit, elle met en conjonction.

Mais cette structure est aussi une analyse qui distingue et différencie les différents items.

La typographie transcrit clairement ces deux mouvements, conjonctif et disjonctif. Pour la conjonction : le choix d'une police uniforme et parfois différente du texte alentour ; l'encadrement ; une distance à la marge particulière (indentation) ; l'adoption d'une série uniforme ou cohérente de marques introductives (numérotation par des minuscules alphabétiques, ou bien des lettres grecques, ou bien un autre système de chiffres ou de symboles ; et si la suite des lettres fait un saut (*a*, *b*, *d*,...), cela peut être le symptôme d'une partie manquant au tout). Pour la disjonction : le quadrillage ; les tirets, puces, pois, flèches, et autres variantes choisies pour leur contraste visuel.

Le modèle Corpus (pour DECID) s'en tient, au moins dans un premier temps, à retranscrire comme liste ce que la mise en page présente comme tel, sachant qu'on ne capte pas ainsi tout ce qui

[24] Ces structures quasi-parallèles correspondent, dans les travaux réalisés à l'IRIT (Tazi, Pascual, Virbel), aux *Fonctions de Structuration Textuelles diviser*, *dénombrer* et *lister*. Leur description fine en métaphrases, chez (Pascual 1991, §II.2.2, p. 64 sq.), souligne également ce double mouvement, de réunion et de distinction de *n* parties à part entière.
Les FST restantes (*titrer*, *détacher*, *énoncer*) relèvent plutôt des parties à fonctionnement méta ou infra-textuel.

pourrait correspondre à une organisation parallèle ou séquentielle de type liste, notamment les énumérations au fil du texte.

Saillances

Si l'ordinateur peut ne voir le texte que comme une plate chaîne de caractères, le lecteur n'embrasse pas le texte uniformément. Pour lui se détachent des éléments saillants, qui ancrent son attention. Interviennent en ce sens des mises en valeur typographiques, l'insistance du rédacteur sur un point, l'information qu'a en tête le lecteur et qu'il recherche, le paragraphe qui lui apporte un élément nouveau et important pour sa réflexion, etc.

Les formes de saillance que nous retenons s'inspirent d'une situation de lecture de travail classique. Intéressé par un mot, un passage, le lecteur se saisit d'un crayon et fait ressortir l'élément du reste de la page. Il modèle le document : quand il y revient, il *s'y retrouve* ; il a posé des « signes de piste » pour retrouver son *parcours* interprétatif.[25]

Les mises en relief que nous nous proposons de coder sont le moyen de modeler le texte au gré de différentes attentes et dynamiques de lecture, de poser des points de focalisation, sans lacérer ni le dénaturer le texte. Tout le contexte est préservé, mais le lecteur peut ainsi orienter le système vers une facette particulière qui l'intéresse.

Le surlignage horizontal

Quelquefois le lecteur arrête son attention sur un terme bien choisi, une notion-clé, le nom d'une personne, d'une technique précise. Il veut retenir *l'expression* en question.

Il la marque par exemple en la soulignant, en l'encadrant, ou par une flèche dans la marge : nous en faisons un surlignage *horizontal*.

Le surlignage vertical

Tantôt c'est un passage que le lecteur trouve plus particulièrement intéressant. C'est l'idée sous-jacente qui lui plaît : elle aurait pu être formulée autrement. Et s'il se remémore ce point-là du texte, il ne l'exprimera pas nécessairement de la même façon et dans les mêmes termes que l'auteur.

L'annotation que le lecteur laisse sur la page est alors typiquement un trait (droit, courbe, ondulé) dans la marge sur toute la hauteur du passage retenu. Le passage est désigné dans son déploiement *vertical*.

---

[25] Le surlignage est une forme d'annotation. Une étude plus approfondie des pratiques d'annotation fait apparaître que les zones ainsi repérées sont parfois catégorisées, étiquetées (Virbel 1994, §II, p. 94).
Ceci conduit les concepteurs d'un poste de lecture assistée par ordinateur à prévoir des « unités logiques de document » :
« A *Logical Document Unit* delimits a document fragment between a start and an end-point, and classifies this fragment as belonging to a certain *type*, identified by a *type name*. [...] *Zones* are simplified case of Logical Document Units since they are identified solely by a fixed background color : this is the highlighter metaphor. It is useful in cases when one wants for instance to mark a passage as 'important' without wishing to be more specific for the moment. » (Chahuneau & *al*. 1992, §2.2.1)
Nos surlignages correspondent à cette notion de *zone* plutôt qu'à celle d'*unité logique de document*. Pour DECID, tous les surlignages renvoient implicitement à une même catégorie : l'indication de leur importance, au plan du calcul des rapprochements.
Une autre forme de structure est envisagée, pour la description de parties typées, qui jouent un rôle dans la lecture : il s'agirait de *rubriques*, typées par une étiquette, associées à des zones de texte, et relevant d'un *plan*. Le *plan* regroupe des *rubriques* qui prennent ainsi sens les unes par rapport aux autres. Le *plan* est aux *rubriques* ce que le *rangement* est aux *boîtes* (cf. structures décrivant les rapports intertextuels), pour des structures qui se réalisent à l'intérieur des textes. Ces structures seraient une manière de coder des plans-types. (Pascual 1991, §II.2.2.B, p. 67) relève également l'existence de telles *rubriques*, participant à l'*architecture textuelle*, mais sans les organiser par le concept de *plan*.

Rapports intertextuels

Les textes tissent entre eux quantité de liens qui, bien que souvent implicites, jouent un grand rôle dans la perception que l'on a d'un document particulier. Le document s'inscrit sur ce fond, et il serait bien difficile –et artificiel– de l'en arracher.[26]

Par exemple, je relève mon courrier et j'y trouve :

- les tracts de deux syndicats en prévision d'une manifestation prochaine (rangement : *syndicats*, boîtes : *F.O.*, *C.G.T., C.F.D.T., C.F.T.C.*),
- le dernier numéro de l'hebdomadaire informatique auquel je me suis abonné (rangement : *périodiques*, boîte : *01 informatique* ; j'en extrais au passage un article, dont je me fais une copie pour la ranger avec les descriptions de logiciels d'analyse textuelle),

[26] En proposant une *Sémantique Interprétative Intertextuelle*, (Thlivitis 1998) a également mis au point une structure de description de relations intertextuelles. Il se base sur le concept de classe sémantique, en définissant de façon unifiée des « classes sémiques de niveau textuel (avec les lexies [*i.e.* mots] comme éléments textuels), de niveau intertextuel (avec les textes en tant qu'éléments) et de niveau inter-intertextuel (avec les intertextes en tant qu'éléments) » (*ibid*., §2.1.2, p. 38). Les classes sont organisées par des traits génériques (s'appliquant à tous les éléments de la classe) et spécifiques (opposant des éléments deux à deux). Des exemples de traits organisant les textes sont donnés dans les deux extraits ci-après :
« les traits sémantiques d'un texte au sein d'une anagnose peuvent être de deux types :
- concernant le texte en tant qu'unité et par rapport aux autres textes de l'anagnose, e.g. /parodie/, /écrit avant/, etc.
- concernant le texte en tant que contenant des entités textuelles de niveau inférieur, notamment des lexies. En effet une *isotopie* entre lexies d'un texte peut devenir trait sémantique [...] au sein de l'anagnose. »
(*ibid*., §2.1.4, p. 43)
« [Une classe de textes peut avoir une validité globale (elle est valable pour toutes les interprétations), ou locale (elle est propre à une interprétation)].
[Dans] le cas le plus global, [...] la classe [...] contient des traits sémantiques stabilisés et généralement reconnus. C'est le cas, par exemple, de la caractérisation d'un texte selon un genre (*e.g.* /parodie/, /poésie romantique/) ou même selon des propriétés plus pragmatiques que l'on aurait intérêt à introduire de manière linguistique comme traits distinctifs entre textes entiers : *e.g.* l'*auteur* ou l'*époque d'écriture*, quand, bien sûr, ces informations peuvent être vérifiées et généralement acceptées. [...] ces traits sont par défaut applicables à toute nouvelle analyse.
[Dans] le cas local, [...] la classe de textes [...] exprime les objectifs particuliers d'un lecteur relatifs à son analyse actuelle [...]. Par exemple, les relations de *transformation intertextuelle*, d'*influence*, de *même genre*, etc., concernent les textes en tant qu'unités selon la spécificité de la compétence interprétative du lecteur. De même pour les relations issues d'une hypothèse sur l'auteur ou sur l'époque d'écriture d'un texte dont les origines ne sont pas claires. »
(*ibid*., §2.2.6.2, p. 66)
Par ailleurs, on note que Théodore THLIVITIS choisit en fait de décrire les classes de textes non comme des ensembles, mais comme des listes :
« L'intertexte a une *structure*. Les éléments-textes sont structurés. La structure primaire que nous avons envisagée est simplement linéaire, dépendant de la disposition des textes dans le « texte » de l'intertexte.
Cette structure est interprétable. Typiquement modélisée par la *tactique de l'expression* [cf. (Rastier 1989)] elle peut, par exemple, servir à un lecteur pour expliciter l'ordre d'apparition d'un ensemble de textes. De même, elle peut exprimer les relations thématiques entre différents textes d'un auteur, à l'aide de rassemblements tactiques ; etc. » (*ibid*., §2.1.3, p. 40)
Notre modèle des rapports textuels est compatible avec certaines des observations précédentes. Quand il s'agit de traduire un effet de succession ou d'organisation linéaire entre différents textes, nous proposons le concept de *pile*. Nous pensons qu'il y a aussi des cas où ce qui est interprété n'est pas d'ordre linéaire, mais de groupements qui traduisent des oppositions, des ressemblances, et l'existence d'entités globales (un regroupement forme un tout) : les concepts de *rangement* et de *boîtes* sont conçus pour exprimer cela.
Nos choix descriptifs contrastent deux sur deux points avec ceux de Théodore THLIVITIS. Premièrement, la notion d'ordre entre les textes n'est pas toujours pertinente. Deuxièmement, la description de la structure ne se centre pas sur des traits avec leur valeur sémantique propre, mais sur des effets collectifs (que pourraient traduire de tels traits). Autrement dit, nous n'exploiterions pas le fait qu'un texte soit écrit par tel auteur (qu'en dire, pour nos calculs ?), mais le fait que tout un ensemble de textes soient l'œuvre d'une même personne, à la différence du reste du corpus (les calculs vont s'appuyer sur ce contraste).

- la Note de mon collègue Untel que je lui avais réclamée avec insistance depuis deux semaines (rangement : les échos des projets avec lesquels je suis en relation, boîte : le projet de mon collègue Untel).

Ces rapports intertextuels se retrouvent aussi dans un corpus littéraire : je peux le vouloir le considérer par genre, ou par auteur, ou par période. Dans ce dernier cas, je peux vouloir ne pas trancher sur le statut d'une œuvre, permettre le recouvrement de périodes, etc.

En pratique, cela fournit des regroupements, des oppositions et des séries intertextuelles, exploitées comme des zones de localité d'ordre supérieur.

Ces rangements sont à géométrie variable, et relatifs à un contexte, à une pratique, à une visée interprétative. Ils n'échappent pas même à une certaine contingence :

> Certains [interlocuteurs] disent avoir changé de système [de rangement] en changeant de poste ou de bureau. Ceci nous rappelle que la cellule de travail (homme + bureau) est bien la bonne cellule d'observation. La structure matérielle du bureau, avec ses placards, ses tables, etc. conditionne le nombre et la forme des piles possibles. [...] un bureau avec plus de surfaces planes accessibles, par exemple, incite à produire plus de piles, et donc de nature différente. (Fischler & Lahlou 1995, §5.2.5, p. 39)

Rangements dans des boîtes

On peut considérer que l'on a à sa disposition un ensemble de boîtes archive, dans lesquelles on peut ranger les documents. Plusieurs rangements sont possibles (par exemple : par année, ou par auteur). Certains rangements peuvent me conduire à dupliquer un document (plusieurs auteurs,…), ou ne pas prendre en compte certains documents (document anonyme).

En résumé, c'est donc une sorte de classement, mais avec les propriétés suivantes : coexistence de plusieurs classifications, possibilité de ne pas trancher entre plusieurs classes (multiclassement), acceptation d'avoir dans certain cas des inclassables (classement non exhaustif). Chaque document peut donc librement être affecté à aucune, une ou plusieurs boîtes, dans un ou plusieurs rangements.

Chaque rangement est un plan de comparaison des documents : les boîtes structurent alors les ressemblances / différences d'un document à l'autre.

Les boîtes sont étiquetées, non pour déterminer et contrôler ce qu'elles peuvent contenir, mais pour décharger la mémoire (c'est un indicateur qui suggère le contenu de la boîte, sans avoir besoin de le réexaminer à chaque fois) et pour guider l'interprétation (en tant que quoi tel document a été classé dans telle boîte)[27].

Piles stratifiées

Les piles sont une réalité inévitable des bureaux... Une pile correspond à une catégorie de documents, et ce que nous voulons surtout retenir ici dans cette métaphore c'est l'ordre (le plus souvent chronologique) qui s'instaure entre les documents. Il y a « le dessus », « le milieu », « le bas » de la pile, avec des degrés d'accessibilité décroissant. La stratification implicite peut bien sûr être plus ou moins détaillée (*le dessus* / *le reste*, vs *les 5 premiers centimètres* / *documents plutôt au-dessus* / *les 5 cm du milieu* / *le bas*), et plus ou moins égalitaire (*la moitié du dessus* / *la moitié du dessous* vs *le premier document* / *le reste*).

Les niveaux d'empilement sont interprétés : ne dit-on pas d'un document qu'il est « enterré quelque part dans (sous) cette pile » ? ou que les dernières versions arrivées (au-dessus) périment les versions précédentes (quelque part en dessous) ?

Cette notion de pile stratifiée est un moyen d'organiser graduellement un ensemble de documents.

c) Descriptif précis des éléments et de leur articulation

Les lignes qui suivent sont une glose de la DTD (donnée en annexe).

[27] (Fischler & Lahlou 1995, §5.1.1, p. 28) observent, dans l'organisation des bureaux et l'utilisation des boîtes archive, « deux fonctions de l'étiquette : effet démultiplicateur de la mémoire, transférabilité de la mémoire externe » (pour le successeur ou pour un collègue par exemple).

Un corpus (CORPUS) est une série de documents (DOC). Une place est prévue pour l'enregistrement d'une référence documentaire le cas échéant (REF).

Chaque document peut être situé dans un nombre quelconque de boîtes (BOX) rapportées chacune à un rangement (ARR). Notamment, il peut ne recevoir aucun rangement, ou être classé dans plusieurs boîtes d'un même rangement (le nom du rangement sera répété à chaque fois). Cette notation, très souple, sert à enregistrer des relations entre les documents : même auteur, même année, etc.

Chaque document a un contenu textuel (TEXT), commençant par un titre (TIT). Le texte se laisse découper en une suite d'alinéas (NLS pour *New Line Section*) ; cependant certains alinéas ont une forme d'intertitre (HEAD remplace alors NLS).

Deux types de structuration peuvent être appliqués (récursivement) aux segments du texte : signalement de zones se comportant comme un résumé, une formulation synthétique (META), par rapport au texte avoisinant délimité par les balises SCOP ; regroupement de différents passages (ITEM) situés dans une même partie du texte (BLOC) se présentant comme un ensemble de mentions complémentaires, de même « niveau ».[28]

A l'intérieur du titre, des intertitres ou des alinéas, les balises IDEA et EXPR servent à marquer des parties mises en valeur respectivement pour l'idée sous-jacente et pour les termes employés. Il est possible de souligner une formulation à l'intérieur d'une zone importante sur le plan du contenu, c'est-à-dire d'inclure un élément EXPR dans un élément IDEA ; l'inverse ne semble pas pertinent et n'est donc pas prévu.

La DTD actuelle ne rend pas encore compte des parties à fonctionnement infra-textuel et des piles stratifiées, qu'on n'a pas prévu d'utiliser dans les premières applications. Leur introduction dans la DTD serait immédiate ; c'est leur exploitation qui exigerait une évolution des modules de traitement.

d) Mise en perspective par rapport aux DTD connues

Un accord rassurant et un écart bénéfique

On ne s'étonnera pas que la DTD Corpus, qui se veut minimale et adaptée à la plus grande diversité de textes, se fonde sur des éléments qui apparaissent quasiment tous, d'une manière ou d'une autre, dans les trois DTD ou familles de DTD prises comme exemples de structuration de textes[29]. On retrouve : l'articulation corpus / documents et la présence d'un niveau de base (également niveau charnière), le rôle particulier du titre du texte, les divisions ou sections (SCOP), la distinction des titres

[28] On a choisi, dans la DTD Corpus actuelle, de déclarer ITEM comme un composant de BLOC, et META comme un composant de SCOP. Ce sont dans les deux cas des composants parmi d'autres possibles, ce qui permet de décrire des structures « entrecoupées » (une liste avec quelques paragraphes de remarques insérés, par exemple). En revanche, pour pouvoir rendre compte d'une organisation de base en une séquence de chapitres, ensuite rythmée par un regroupement en parties, il faudrait user du mécanisme d'inclusion (un élément BLOC autoriserait l'inclusion d'ITEM, un élément SCOP l'inclusion de META) : les chapitres seraient des ITEM qui relèvent du premier ascendant BLOC, *sans être masqués* par les SCOP, s'intercalant entre eux et le BLOC, pour traduire les parties. Nous considérons dans un premier temps que ces structurations particulières sont rares, et nous évitons le mécanisme d'inclusion, qui augmenterait sensiblement la complexité des traitements.
Ces deux cas de structures complexes font partie des cas examinés par (Pascual 1991, §II.2.4, p. 83 sq.) ; les 23 autres cas ne signalent pas d'autres difficultés particulières pour une représentation par la DTD Corpus.

[29] Les structures textuelles élémentaires que nous proposons (zones, saillances, rapports intertextuels) peuvent également être confrontés à des études sur les relations structurelles de textes menées dans une autre perspective que celle du codage. Ainsi, (Virbel 1992) procède par une méthode harrissienne pour recenser les formes de structuration, telles qu'elles sont méta-linguistiquement nommées, et il en propose une classification. Celle-ci entre dans le détail des modes de réalisation des rapports (ex. : indexation, démonstration, annotation, références, etc.). L'ensemble des rapports structuraux évoqués nous semblent couverts par notre modèle, essentiellement par les rapports intertextuels (*représentations* et *substituts*, lorsqu'ils induisent une relation d'analogie entre textes), les parties à fonctionnement méta-textuel (*attachement* par *adjonction* d'un titre), et les parties à fonctionnement infra-textuel (*attachement* par *adjonction* de commentaires, d'extensions).

des sections, les structures de listes. Ces structures apparaissent comme des formes architecturales des corpus et des textes.

Des traductions entre le format de la DTD Corpus et un format TEI, ISO 12083 ou GRAALDOC sont tout à fait envisageables. Elles préservent les structurations fondamentales mentionnées ci-dessus ; en revanche, les autres informations de structuration seraient perdues.

Il y a donc une bonne compatibilité entre la DTD Corpus et les DTD existantes. Il faut s'en réjouir : cet accord sur des structures fondamentales du texte est rassurant. Pourquoi une nouvelle DTD ? La DTD Corpus est une réponse adaptée à une application (DECID) : elle a l'avantage d'*expliciter clairement et avec exactitude* les éléments et structures pertinentes dans ce contexte, et de *ne pas encombrer* et brouiller la description par des éléments ici inutilisés hérités d'un modèle trop général. La DTD Corpus est incontestablement plus simple et plus légère que ses consœurs.

Points de différence

La DTD Corpus est originale dans sa manière de rendre compte de l'intertextualité, en déclarant des couples attribut-valeur libres au niveau des documents. Cette technique a la même souplesse et la même articulation que les attributs de séquence de GRAALDOC.

L'en-tête, justifiée dans une perspective d'archivage et de réutilisation, est absente de la DTD Corpus, qui n'est qu'un format de travail. Incidemment, cela a l'avantage de porter toute l'attention sur le texte, et non sur les informations relatives au texte. Les DTD documentaires, comme ISO 12083, auraient plutôt tendance à faire une description privilégiée du document secondaire, et ne détaillent pas autant la structuration des documents primaires.

Chaque DTD prévoit une balise pour délimiter les zones où l'on mentionne du code SGML, qui ne doit pas être interprété lors du traitement : c'est un artefact lié au formalisme, qui n'a pas une signification particulière au plan de la textualité. Il serait toujours temps de l'introduire, si nécessaire, dans la DTD Corpus.

La notion de lien n'est pas retranscrite dans la DTD Corpus, sinon par la relation de portée sous-jacente aux fonctionnements méta- ou infra-textuels. Les renvois restent difficiles à identifier et à modéliser dans une analyse basée sur la présentation. La réflexion reste ouverte.

Un certain nombre de balises utilisées par les autres DTD pour enregistrer des mentions particulières (par exemple : date, nom de personne), ne sont pas du ressort de la structure apparente du texte, mais de son analyse et de son interprétation. Leur repérage n'est clairement pas une « donnée » pour la majorité des textes des corpus de DECID. Il est aussi hasardeux de recenser *a priori* les natures des informations pertinentes.

De même, des cristaux particuliers, comme l'adresse ou des références bibliographiques, ne sont pas du même niveau d'abstraction et de généralité que la structuration en paragraphes et les autres relations décrites. Ils peuvent faire l'objet d'un éventuel repérage interne au traitement. Leur pertinence pour intégration à la DTD Corpus resterait à démontrer.

Les attributs codant l'identification d'un élément, et son enchaînement à un élément précédent et à un élément suivant, ne sont pas repris dans la DTD Corpus. En effet, dans l'utilisation prévue, ces attributs seraient une indication redondante avec l'information de localisation que l'on tirerait du texte structuré.

2. Mise en œuvre du modèle : une herméneutique pour des formats électroniques ?

a) Interprétation de fichiers ASCII : mise en évidence de deux types de lignes

Un fichier texte ASCII se présente comme une chaîne de caractères, dans laquelle les caractères n'ont pas tous le même statut.

D'une part, les *caractères graphiques* correspondent à la transcription de signes qui s'alignent pour composer les mots et les phrases du texte. Ce sont les lettres, chiffres, ponctuations, symboles graphiques. Le *blanc* (espacement) en fait aussi partie, dans la mesure où il s'insère dans la suite des

signes affichés. Ceci calque l'écriture alphabétique et permet un bon enregistrement du matériau linguistique, au sens où les caractères et le blanc séparateur s'apparentent respectivement aux unités de première et de seconde articulation du langage.

D'autre part, les *caractères de contrôle* codent un jeu prédéfini et limité d'instructions pour le traitement du fichier par l'ordinateur et ses périphériques (l'imprimante). Un bon nombre d'entre eux relèvent des protocoles de transmission et de réception des fichiers et d'échange de données : ceci est lié à la machinerie électronique et n'intéresse pas les textes eux-mêmes. Concernant plus spécifiquement les textes, on relève quelques instructions de mise en forme de ce qui peut être imprimé : *blanc insécable*, *tabulations*, *changement de ligne*, *changement de page*.

Quand il s'agit de repérer le découpage d'un texte en paragraphes (plus précisément en alinéas, au sens de notre modèle), le traitement s'appuie sur les caractères de contrôle. Cependant, on s'aperçoit que l'usage des caractères de contrôle n'est pas univoque. Deux modes d'usage apparaissent.

Le premier, c'est le cas du fichier texte saisi dans un éditeur de texte rudimentaire, puis dont la mise en page a été entièrement ajustée à la main. Typiquement, la taille de chaque ligne est fixée en introduisant un retour-chariot tous les 80 caractères environ, l'ouverture d'un nouveau paragraphe est marquée par un léger retrait du début, ou une ligne d'espacement.

Le deuxième forme est celle d'un fichier saisi au kilomètre, en indiquant simplement par un « enter » que l'on passe à un nouveau paragraphe. Au logiciel revient d'afficher cela lisiblement, en faisant une césure automatique des lignes.

Se profilent donc deux acceptions pour le mot *ligne* : soit la ligne correspond à ce que visuellement on appelle comme tel, sur une page de papier ; soit c'est en fait la marque d'une fin de paragraphe. Cette dernière acception est l'usage commun en informatique : le fichier est lu (instructions `Get_Line`, etc.) et construit (`Put_Line`, `writeln`, etc.) comme une suite de lignes. Les lignes enregistrent donc les segments de base du texte, et l'information de segmentation que l'on choisit usuellement de retenir est le découpage en paragraphes, intrinsèque au texte car invariant d'une édition à l'autre.

Les programmes de lecture et de conversion des fichiers textes simples (ASCII) au format de la DTD Corpus sont donc au nombre de deux, pour tenir compte de l'interprétation différente des caractères de contrôle et notamment du caractère fin de ligne (retour-chariot). Il faut utiliser l'un ou l'autre, selon la manière de lire le fichier d'entrée.

Le module `asc2dcd` (lire *ASCII to DECID*, i.e. *ASCII vers DECID*) est le plus simple : chaque ligne informatique est une section de base du texte, un *alinéa*.

Le module `csr2dcd` (*césure to DECID*) doit recoller entre elles les lignes informatiques pour reconstituer les paragraphes. Il repère le changement de paragraphe par les retraits en début de ligne (espacement horizontal) et / ou un saut de ligne(s) (espacement vertical).

Dans les deux cas, la DTD corpus reste exploitée de façon relativement pauvre : on s'en tient à des indications sur le découpage en sections ; le repérage de structures de listes ou de rapports méta-textuels serait lourd et hasardeux. [30]

[30] Pour les formats que nous avons traités, l'interprétation de mise en forme a pu se faire par un parcours séquentiel du fichier, en une seule passe, donc sans point de vue global. (Pascual & Virbel 1992), s'intéressant à des formats non balisés et avec une typographie plus élaborée que l'ASCII, posent à juste titre la question de l'exploitation d'une vue globale du document, pour la reconnaissance de son architecture. Pour le cas de ces formats, ils montrent également l'utilisation d'indices (méta)lexicaux (« chapitre », « introduction ») et de patrons d'enchaînements (ex. : « chapitre » + n° + titre + « introduction ») pour cerner les possibilités de portée et d'emboîtement des parties reconnues.

b) Interprétation de fichiers SGML : savoir prendre en compte les instances non conformes

Obstacles à une approche normative

Sachant que chaque document structuré est construit selon la DTD dont il se réclame, nous avions bravement entrepris un programme Balise[31] pour projeter les éléments HTML (décrits dans la DTD HTML 2.0 -la plus usitée- ou 3.2 -celle qui a cours-) dans des éléments de notre DTD Corpus.

Première constatation : la quasi-totalité des pages sur Internet ne sont pas formellement conformes à la DTD publique. Même le site distribuant cette DTD se permet des écarts[32] ! Si notre module informatique veut donc avoir une quelconque utilité, il doit *être capable d'interpréter une instance inspirée de la DTD HTML, mais non moulée sur elle*.

Deuxième constatation : SGML contrôle la syntaxe de structuration des éléments, pas le contenu textuel des éléments. Certes, le nom d'un élément incite à en faire un certain emploi (`TITLE` sera un titre, `P` un paragraphe, etc.), mais rien n'interdit d'inventer de nouveaux usages quand au final cela rend l'effet voulu (faire une jolie présentation centrée, en prenant une liste à un niveau d'emboîtement suffisant pour avoir l'indentation assez grande recherchée). L'enseignement à tirer de cette observation est que *l'interprétation ne doit pas se fonder uniquement sur le nom des éléments, mais d'abord sur l'apparence donnée à chaque élément par les logiciels de navigation*. Cette apparence n'est pas normalisée, on en retiendra donc les traits caractéristiques qu'on retrouve dans les principaux logiciels (*Netscape*, *Internet Explorer*).

La non-conformité des pages HTML à la DTD de référence pouvait conduire à deux solutions tactiques différentes. La première : assouplir la DTD, jusqu'à la rendre compatible avec une proportion satisfaisante des documents à traiter (voie par exemple proposée par (Amitay 1997, §2, pp. 17-18)). C'est un réaménagement délicat (portée du choix de relâcher telle ou telle contrainte, préservation de la cohérence d'ensemble), obtenus par tâtonnements, et qui obéit à une logique de nivellement par le bas. Nous lui préférons la solution d'un module qui interprète non seulement les passages conformes mais aussi les irrégularités.

Régularités sur lesquelles s'appuyer

L'observation des sources de pages Web nous a permis d'identifier deux facteurs textuels explicatifs de la plupart des irrégularités syntaxiques HTML. D'une part, *l'implicite*, qui se manifeste de plusieurs manières : *(i)* juxtaposition et non imbrication : « je commence un paragraphe –c'est donc évident que l'intitulé que je lui ai donné s'arrête là » ; *(ii)* restitution par défaut du niveau de base : « je commence une nouvelle page et je tape quelques lignes, je m'attends à les voir affichées comme un paragraphe ». D'autre part, une focalisation locale, qui fait que le rédacteur oublie l'organisation générale du document et se concentre sur le *déroulement* de son texte, l'enchaînement de deux éléments. Par exemple, il commence une structure de haut niveau (mettons `BODY`) et, pris dans son développement, il oublie d'en signaler la fin.

La modélisation se conçoit en termes de niveaux. Chaque balise a un niveau associé. Les balises s'emboîtent par niveau croissant. Selon la balise, le niveau de la balise suivante peut ou non rester le même (croissance stricte ou non). La présentation du programme de lecture du fichier est détaillée dans la section suivante.

---

[31] *Balise* est un des logiciels les plus puissant pour exploiter des documents structurés : vérification de la conformité à une DTD (est-ce que ce fichier que je reçois est déjà au format Corpus), transformation d'une instance (par exemple, interversion de deux chapitres d'un livre), passage d'une DTD à une autre (ex. : HTML vers Corpus), génération de données non nécessairement SGML (ex. : extraction et présentation tabulée de l'information *référence (code) de l'Action - titre de l'Action*).

[32] En parsant la page d'accueil du site Internet du *World Wide Web Consortium* (W3C) avec la DTD HTML 3.2 qu'il fournit, nous avons obtenu une série de messages d'erreur, dénonçant des attributs inexistants et des ancres mal placées.

Une lecture orientée

Pour avoir un traitement robuste, l'algorithme reconnaît toute balise SGML, et fait donc bien la part entre le balisage et le contenu textuel. Si le nom de la balise lui est inconnu, il l'ignore et passe à la suite. Cela revient à considérer que l'on n'a pas nécessairement toute l'information sur le codage du texte, mais que l'on a décrit un sous-ensemble pertinent et suffisant dans le cadre de l'application. D'une certaine manière, on matérialise déjà ainsi le fait que l'on opère une lecture du texte parmi d'autres possibles, lecture correspondant à certaines attentes, ici la lecture de HTML dans la perspective de la DTD Corpus.

Le module réalisé effectue la lecture suivante :

Balise HTML 3.2	*Interprétation*	*Contribution dans la DTD Corpus*	*Remarques*
HTML	Html	DOC	Chaque page HTML est un document ; un corpus pourrait être formé par la mise bout à bout de plusieurs pages dans un seul fichier.
HEAD	Header_Side		Délimite les informations factuelles (*vs* textuelles)
TITLE	Title	TIT	Ce titre est celui qui figure par exemple dans le signet que l'on place sur une page ; ce n'est pas celui qui est affiché en grands caractères au début de la page le cas échéant (dans la page on ne dispose que des H1, H2, etc.).
BASE	Base_Url	ARR, BOX	La valeur de l'attribut HREF indique le site Web de base, de référence. En l'enregistrant, on souhaite garder l'indication comme quoi telle et telle page sont issues d'un même site. (Dans les faits, cette information sera peu fréquente et de qualité irrégulière.)
BODY	Body_Side	TEXT	On peut trouver des pages « blanches », sans contenu textuel, mais pourvues d'un titre : ce sont des pages « minimales » mais qui ont quand même un minimum d'informations. Concrètement, cela signifie que l'absence de corps (BODY) dans le document HTML n'entraîne pas nécessairement l'absence de texte (TEXT) dans le document Corpus, le texte pouvant n'être constitué que du seul titre.
H1, H2, H3, H4	Heading	HEAD	La mise en forme effectuée par les navigateurs appose la présentation conventionnelle de titres.
DIV, CENTER, BLOCKQUOTE, FORM, ADDRESS	High_Division	(NLS)	Délimitent des alinéas, à défaut de divisions plus fines. (Les alinéas ne sont pas emboîtés.)
P, PRE, H5, H6	Low_Division	NLS	La présentation en caractères menus des plus petits niveaux de titre est souvent utilisée pour d'autres fonctions que le marquage d'intertitres (copyright par exemple) : on banalise donc H5 et H6 en alinéas.
HR	Dividing	NLS ou HEAD	La division est contextuelle : si une division intervient au milieu d'un alinéa, elle le sépare en deux alinéas ; si la division intervient dans un intitulé, elle le sépare en deux éléments de type intitulé.
BR	Caesura	(NLS ou HEAD)	Un seul retour à la ligne n'est souvent qu'une indication de mise en page ; il faut alors juste rétablir l'espace implicite entre le dernier mot de la ligne et le premier mot de la ligne suivante. Plusieurs retours à la ligne entre deux segments textuels les présentent comme deux zones distinctes, mais de même nature (si la première s'affiche comme un intitulé, la seconde aussi).
UL, OL, DL, DIR,	List	BLOC	

MENU LI	Item	ITEM	
DT	Definiendum	ITEM, HEAD	Le début d'un « terme à définir » (ou *entrée* dans un lexique) marque le début d'un élément d'une liste (le lexique), et la fin de l'élément précédent le cas échéant.
DD	Definiens	NLS	
TABLE	Table	SCOP	Cet élément n'est pas nécessairement utilisé pour insérer des tables de données au sein d'un texte : on le trouve aussi comme organisant la présentation d'une page avec la disposition de plusieurs blocs de texte. Même sans marquage du quadrillage sous-jacent, ces blocs sont perçus comme un ensemble. L'élément SCOP permet de représenter le rapport du tableau à sa légende éventuelle.
CAPTION	Caption	META	On retranscrit le fait que la légende porte sur le tableau.
ROW	Row		On ne donne pas un statut privilégié aux lignes, par rapport aux colonnes.
TH	Header_Cell	(NLS)	S'il n'y a pas d'alinéa précisé, une cellule de tableau correspond implicitement à un alinéa.
TD	Data_Cell	(NLS)	*idem*
APPLET, IMG	Alt_Media		On va chercher l'équivalent textuel donné par l'attribut prévu à cet effet, s'il est présent. Là encore, pour l'insérer dans le texte courant, il faut penser à rétablir les espaces qui entourent implicitement cette formulation.
SCRIPT, STYLE	Code		Les programmes gérant la composition dynamique de la page ne font pas partie de son texte.

Remarque : Dans la lecture actuelle, on ne gère pas la double lecture possible des tableaux :
- en tant que zone que décrit globalement une légende (balises SCOP / META, fait),
- en tant qu'ensemble de cellules en interrelation et qui forment un tout (BLOC / ITEM, non fait).

Cela demanderait d'introduire une certaine complexité dans le traitement ou / et de réviser la DTD, ce à quoi nous avons renoncé dans le cadre limité et exploratoire de la thèse.

c) Interprétation des clics souris : multiplicité du clic, paliers de sélection, et nature de la désignation

Au moment de soumettre un texte à DECID, l'utilisateur peut orienter les recherches sur certains points particuliers, et signaler des termes particulièrement significatifs ou importants, utiles à retrouver dans les textes rapprochés.

Les éditeurs de texte ont répandu l'usage de clics-souris pour la sélection de zones de texte de différents paliers. Conventionnellement, plus le nombre de clics est grand, plus large est la zone. Pour Word, au fil du texte, un clic positionne le curseur (point dans la chaîne de caractères), deux clics sélectionnent le mot, trois le paragraphe ; dans la marge, un clic sélectionne la ligne, deux le paragraphe, trois le document. Visuellement, le texte sélectionné apparaît comme surligné par un fond coloré. Le même effet est rendu en glissant le curseur, maintenu appuyé, sur le texte.

L'interface de DECID, en s'inscrivant dans cette logique, offre un mode de sélection intuitif. Le surlignage intervient sur deux paliers. En sélectionnant des mots (deux clics souris), l'attention est attirée sur telle formulation précise. Si c'est le thème abordé par un passage qui doit être retenu, alors un clic supplémentaire permet la saisie rapide de l'ensemble du paragraphe.

Dans notre modèle, ces deux types de surlignage ne peuvent pas se combiner de n'importe quelle manière : des expressions peuvent être mises en valeur au cœur d'un paragraphe exprimant une idée importante, mais les chevauchements et l'inclusion inverse ne sont pas permis. Le fait de saisir d'une part des mots, d'autre part des paragraphes, garantit la conformité au modèle.

En situant le surlignage de l'expression au niveau du mot et celui de l'idée au niveau du paragraphe, on fait une approximation, suffisante dans un premier temps pour notre système. D'une

certaine manière, on pose le mot (graphique) comme unité élémentaire de formulation, et l'on suppose généralement admissible la règle « un paragraphe, une idée ». Ces choix, évidemment discutables pour une linguistique un tant soit peu rigoureuse, pourraient être affinés ou revus en fonction des besoins des utilisateurs et des possibilités de traitement.

La combinaison dynamique des deux surlignages introduit quelques subtilités. Observons notre utilisateur : parcourant son document-requête, il décide de sélectionner un mot, puis un autre, puis en fait tout un regroupement de notions importantes. Il s'aperçoit ce faisant –aidé par la fatigue de multiplier les double-clics– que c'est plutôt ce passage dans son ensemble qui l'intéresse. Il triple-clique : le surlignage au niveau du mot s'efface pour faire place au surlignage du paragraphe. Il y avait cependant le nom d'une technique précise qui lui paraît crucial : il double-clique dessus, et le terme est surligné d'une couleur plus vive à l'intérieur du paragraphe. En résumé donc : le surlignage d'une idée efface le surlignage des expressions ; le surlignage d'expressions apporte un relief supplémentaire au surlignage d'une idée[33].

Il est important de guider ainsi l'utilisateur vers l'utilisation du surlignage vertical (idée) plutôt qu'horizontal (expression). En effet, le second est beaucoup plus restrictif quant aux rapprochements concernés.

3. Programme de lecture d'un fichier SGML : l'approche par niveaux

a) Niveaux fondamentaux

L'étude de plusieurs DTD de (corpus de) documents textuels fait apparaître six niveaux fondamentaux :

0. le niveau racine, d'ensemble, que décrit la DTD.
1. le découpage en documents.
2. les frontières de chaque texte : en effet, le document peut être muni d'autres informations textuelles, mais qui ne sont pas son contenu à proprement parler.
3. la structuration interne du texte. Elle comprend le découpage de base : un segment de texte est toujours délimité par une et une seule balise faisant partie des éléments de base. Autrement dit, les balises de base effectuent un découpage sans reste du texte. Les autres balises sont facultatives et servent à ajouter une structure au découpage de base ; elles portent donc sur le découpage de base, non directement sur le contenu textuel lui-même.
4. les bribes textuels. Ce sont des éléments particuliers au fil du texte, en deçà du découpage de base.
5. un point dans la chaîne du texte, sans contenu textuel (cela correspond aux éléments vides SGML).

Les uns et les autres ont des affinités particulières avec les facettes que nous avons retenues pour la description des textes : les niveaux 0 et 1 jouent au plan de l'intertextualité ; les niveaux 2 et 3 articulent la structuration interne de chaque texte ; le niveau 4 s'ancre dans la matière linguistique.

Les deux niveaux extrêmes sont un peu particuliers : le niveau 0 n'est inclus dans aucun élément ; et le niveau 5 n'est celui d'aucun élément.

Le tableau ci-après illustre la composition des différents niveaux pour plusieurs DTD. Chaque balise fait partie d'un niveau, mais aussi détermine le niveau des éléments qu'elle peut contenir : c'est le chiffre indiqué entre parenthèses. (Les balises non mentionnées sont celles qui développent la structure des parties comportant des informations factuelles qui ne nous intéressent pas.)

[33] Cette superposition des deux modes de surlignage rend compte de l'*acte annotatif* '*contextualiser'*, tel que l'analyse Jacques VIRBEL :
« *contextualiser* : repérer des termes caractéristiques (mots clés, terminologies et formules propres à l'auteur, etc.) et créer un passage pertinent, du point de vue linguistique, pour l'appréhension sémantique, syntaxique, ou autre, de ce terme ou de cette expression ». (Virbel 1994, §III, p. 97)

	Niveau 0 : ensemble des données	**Niveau 1 : document**	**Niveau 2 : texte**	**Niveau 3 : structuration interne**	**Niveau 4 : segment remarquable**
Corpus	CORPUS (1)	DOC (2)	TEXT (3) ARR, BOX (4)	SCOP, META, BLOCK, ITEM (3) TIT, HEAD, NLS (4)	IDEA, EXPR (4)
HTML		HTML (2)	HEAD, BODY (3)	DIV, CENTER, BLOCKQUOTE, FORM, ADDRESS, UL, OL, DL, DIR, MENU, LI, DD, TABLE, ROW, TH, TD (3) TITLE, H1-H6, P, PRE, DT, CAPTION, SCRIPT, STYLE (4) BASE, HR (5)	APPLET, IMG (4) BR (5)
GraalDoc	GRAALDOC (1)	TEXTUNIT (2)	LISTKW (3)	SECT, FIG, FIGCAP, LIST, ITEM, P, TITLE, (3) SEQ, KW, FACT[34] (4) FIGBODY (5)	SEQATT (5)
Livre Electronique des Actions EDF-DER	DOSSIER (1)	ANNEE, SERVICE, DEPARTEMENT, GROUPE (1) ACTION (2)	SERV, DEP, GRP, ADMINIS, RESPONS, DESCRIPT	CODE, AN, SIGLE, LIBELLE, NOM, TITRE, TITDESC, ENTELIST, PARA[35] (4) *liste*, ELEMLIST (3)	REFERACT (4)

b) Articulation des niveaux

A partir du moment où sont donnés, pour chaque balise, son niveau et les niveaux des éléments qu'elle peut contenir, des lois régissent l'agencement des balises. Chaque niveau suppose la présence du (des) niveau(x) précédent(s) ; et les premiers niveaux, jusqu'à une balise du niveau de base, sont toujours présents dans le contexte de chaque segment de texte. Les balises de ces niveaux peuvent ne pas être marquées, mais alors elles sont explicitables : elles prennent une valeur par défaut, dépendant éventuellement du contexte précédent et de la balise à atteindre. En ce qui concerne les balises de base, par définition elles correspondent à la charnière entre la structuration qui organise tout le texte, et le contenu (les mots) : elles sont donc d'un niveau strictement inférieur à 4, et leur contenu est d'un niveau strictement supérieur à 3.

Les niveaux servent d'indicateurs pour rétablir des balises implicites. Quand une balise est trouvée, elle doit être introduite dans le contexte courant. Si son niveau ne lui permet pas d'être incluse dans ce contexte, les balises du contexte doivent être implicitement fermées jusqu'à revenir à un niveau admissible pour la nouvelle balise. La balise est introduite dans le contexte après avoir inséré les balises implicitement ouvertes pour avoir un contexte complet et cohérent : passage par chacun des niveaux, et enchaînements valides (un élément de liste suppose une liste, une cellule de tableau suppose un tableau, une information documentaire se trouve dans l'entête). Quand on a affaire à du contenu textuel, alors le contexte doit contenir une balise de l'ensemble de base. Si ce n'est pas le cas, la balise de base par défaut est introduite ; si nécessaire, les maillons manquants du contexte sont rétablis (comme pour toute nouvelle balise trouvée).

[34] L'élément FACT est une inclusion à partir du niveau racine, il échappe à la division en niveaux. Il a été placé arbitrairement en niveau 3, mais en fait son comportement sera décrit comme un cas particulier dans les fonctions du programme.

[35] Cette définition de l'élément PARA déforme légèrement la vision de la structure. En effet, elle empêche qu'une liste soit incluse dans un paragraphe. Un paragraphe qui originellement contient une ou plusieurs listes est découpé : chaque zone de texte de part et d'autre des listes est redéfinie comme un paragraphe indépendant. Le paragraphe est donc remplacé par une succession de listes, entre lesquelles s'intercale un paragraphe le cas échéant.

c) Mise en œuvre : informations à apporter pour la description d'une DTD

Concrètement, les données qui permettent au programme de « lire » le fichier SGML, d'une façon robuste et en saisissant une part de l'information implicite, sont donc :

- les balises à reconnaître (les autres seront ignorées et seront bien mise hors du contenu textuel) ;
- pour chaque balise, son niveau et le niveau des éléments qu'elle peut contenir ;
- des enchaînements constants remarquables (par exemple, un *item* est nécessairement inclus dans une *liste*).

A partir de cela, on définit :

- l'ensemble des balises du niveau de base, et celles d'un niveau inférieur ;
- une fonction qui indique si telle balise peut contenir telle autre ;
- une fonction qui donne les balises implicites en un point donné.

Le programme est alors capable d'interpréter le fichier, même s'il contient une instance non strictement conforme à sa DTD. Il perçoit les balises manquantes ou implicites, et effectue une lecture plausible des passages dont la structuration est déficiente.

d) Conception du texte sous-jacente

Toutes les DTD destinées à structurer des données textuelles ne s'analysent pas avec le cadre que nous venons de tracer. La TEI par exemple s'inscrirait avec difficulté dans cette grille, et de petits ajustements ont été nécessaires pour GraalDoc et le Livre Electronique. La possibilité de penser la structuration d'un corpus selon ces six niveaux correspond à plusieurs propriétés.

Quatre paliers de division régulière et complète

Chaque inclusion d'un élément dans un autre correspond à une croissance (stricte ou non) du niveau. Or le niveau initial est constant (élément racine : c'est ici le corpus, l'ensemble des documents), et le niveau du texte y est strictement supérieur. Le contexte de chaque portion de texte traverse donc des points de croissance stricte : ici, les passages du niveau 0 au niveau 1, du niveau 1 au niveau 2, du niveau 2 au niveau 3, et du niveau 3 au niveau 4. Chaque passage d'un niveau à un autre, jusqu'au niveau 4, correspond à une division régulière et totale du corpus.

Ainsi, le passage du niveau 3 au niveau 4 fournit le découpage le plus fin à l'intérieur du texte : il peut s'agir de paragraphes, de séquences, d'intertitres, etc. selon le modèle envisagé. Ce découpage correspond aux unités de traitement des logiciels qui font une analyse séquentielle, aux voisinages élémentaires pour certains calculs distributionnels, etc.

Le passage du niveau 2 au niveau 3 présente le document comme formé de composantes de diverses natures, le texte étant l'une de ces composantes. C'est la transition du niveau 2 au niveau 3 qui sépare les données textuelles des données factuelles, le document primaire du document secondaire, l'ensemble des aspects du document du texte qu'il contient.

Le passage du niveau 1 au niveau 2 n'est autre que la division d'un corpus en documents. Le document représente le plus souvent l'entité extérieure que l'on cherche à caractériser, et que l'on analyse grâce au découpage le plus fin (transition du niveau 3 au niveau 4).

Le passage du niveau 0 au niveau 1 est un cas limite : en général, on ne considère qu'un seul corpus dans un traitement. La prise en compte de plusieurs corpus, et donc la répartition en différents corpus de l'ensemble de textes initial, rend significative cette transition.

La modélisation trace donc ces quatre partitions. Si l'on s'en tient à ces quatre passages nécessaires, on a donc une vision très morcelée de la réalité. En effet, certains textes ont des affinités entre eux, certains segments d'un texte se regroupent, etc. Heureusement, ces informations peuvent trouver leur place à l'intérieur de chaque niveau : ce sont par exemple pour la DTD Corpus les relations intertextuelles transcrites par les rangements et les boîtes ; et les structures d'organisation des alinéas, en liste par exemple. La présence de ces divisions n'exclue donc pas une représentation plus nuancée des relations qui se tissent entre les divers éléments.

Non seulement les divisions ne sont pas nuisibles, mais elles sont utiles et opérationnelles dans plusieurs cas. Elles permettent un repérage simple et précis des localisations. Elles associent

systématiquement à chaque point du texte une série de contextes de différents niveaux. Elles préparent la matière pour des analyses distributionnelles.

Une différenciation de nature entre les niveaux

Chaque balise est porteuse d'un contexte implicite, dû à son niveau. Cela permet de considérer une partie d'un corpus, localement, tout en se calant toujours sur une vision d'ensemble. C'est traduire une forme d'anticipation sur le rôle et la portée de l'élément considéré. Cela reflète bien aussi le contexte somme toute très réduit que l'on a littéralement en mémoire quand on lit un texte sans le survoler, ou quand on le rédige linéairement : l'attention est portée sur un point du texte, le reste constitue un arrière-plan.

Un contre-exemple nous est donné par les éléments intermédiaires de la TEI. Si l'on considère une citation par exemple, rien ne permet de dire, sans en faire l'analyse, s'il s'agit de quelques mots rapportés, ou d'extraits s'étendant sur plusieurs paragraphes. La consistance du paragraphe laisse perplexe, puisque l'on peut *théoriquement* sans limite imbriquer des paragraphes les uns dans les autres via toutes sortes de structures, notamment listes et tableaux. Les éléments intermédiaires jettent une certaine confusion en permettant d'osciller entre un niveau d'organisation d'ensemble de parties du texte, et un niveau d'éléments inclus au fil du texte.

L'application d'un modèle par niveaux transforme les *inclusions* non compatibles avec les niveaux en *insertions*. Ce faisant, il donne un contexte monotone à chaque élément.

e) Vision orientée traitement vs *orientée archivage*

L'approche par niveaux est pertinente pour guider et développer certains traitements sur les textes ; mais les propriétés qu'elle impose dans sa lecture des textes peuvent être trop restrictives quand il s'agit de donner une représentation descriptive la plus « proche » du texte original. La modélisation apparaît alors comme un certain durcissement. Cette différence d'optique se trouve illustrée par la TEI d'une part, la DTD Corpus d'autre part –sachant que c'est la seconde qui est conforme à une approche par niveaux.

La TEI adopte résolument une attitude descriptive, par opposition à une attitude prescriptive. Cela a pour but de proposer des formes de codage très souples, qui s'adaptent à la plus grande variété de besoins d'expression de telle structure particulière. D'où la multiplication des éléments et le peu de limitations sur leurs combinaisons. La contrepartie est que cela réduit les attentes interprétatives et disperse les parcours de lecture possibles.

> Prescription and Description
>
> An SGML document type definition specifies a set of formal rules which define the set of « valid » documents. The formal definition of document validity is one of SGML's great practical strengths, because it allows automatic mechanical checking for markup errors. It also allows the designer of the document type to make the expected structure of documents much more explicit than would be possible without such a formal specification. [...] Such a formal specification may be regarded as a « document grammar », which accepts a certain subset of all possible documents.
>
> Grammars, however, may be used for two quite different purposes. One may use a grammar to prescribe the legal forms of some language. [...] SGML is frequently used for the specification of document grammar which are prescriptive in just this way.
>
> Grammars may also be used, however, to *describe* some set of independantly existing objects. [...] It is characteristic of descriptive grammars that when an object fails to conform to the grammar, the flaw is usually sought not to the object itself, but in the grammar. [...] The TEI explored new territory in attempting to use SGML to formulate a descriptive grammar for existing documents.
>
> In so doing, it faced a critical problem which often arises in the development of descriptive grammars. When the population being described by the grammar is sufficiently various and complex, it is often the case that no grammar can readily be writen which exactly matches the population. Either the grammar *overgenerates* fr the population, that is, it accepts some items which are not actually present in the target population, or it *undergenerates*, that is, it rejects as invalid some items wich actually occur.
>
> [...] For the TEI, the target population includes, in principle, any document written in any language during the entire span of written history [...]. In this situation, we chose consciously to err on the side of overgeneration, rather than undergeneration, whenever the choice presented itself. An

> overgenerating document grammar has the drawback that it accepts nonsensical documents as valid ones ; it will thus fail to catch some errors of markup. [...] [But] an undergenerating descriptive grammar [would behave] like a prescriptive grammar which has strayed into the wrong arena. [...]
>
> In order to minimize the baleful effects of such overgeneration, the TEI tag sets sometimes define two alternative forms of markup : a somewhat prescriptive form for use when validation is highly desired, and a very loose alternative form for use in transcribing items which simply do not fit the more prescriptive form.
>
> (Sperberg-McQueen, Burnard, 1995, § 1.3.3)

La DTD Corpus vise aussi la prise en compte d'une large diversité de textes. Le parti pris est à l'inverse de se doter d'un petit nombre d'éléments et de structures très générales. Le risque serait de manquer de précision pour un archivage (perte d'informations), mais les éléments et les structures choisis correspondent à l'usage des textes dans DECID. En se conformant à un modèle par niveaux, on applique une lecture particulière, mais dont on sait qu'elle est pertinente et opératoire dans le cadre du traitement visé.