Annotation collaborative d'un corpus de documents médiévaux : outils pour l'analyse de la structure et du contenu des sermons de Jacques de Voragine

Marjorie Burghart

Annotation collaborative d'un corpus de documents médiévaux : outils pour l'analyse de la structure et du contenu des sermons de Jacques de Voragine

Résumé : Cet article présente les travaux réalisés pour informatiser un corpus de sermons médiévaux[1], au sein de l'équipe internationale dirigée par Nicole Bériou[2]. Pour permettre la mise en valeur et l'exploitation d'une source caractérisée à la fois par la richesse de son contenu et par l'importance de sa structure, nous avons opté pour la mise en œuvre des technologies liées à XML. Confrontés à diverses questions techniques (quels outils choisir ?) et méthodologiques (comment faire primer la réflexion scientifique sur les contraintes de la modélisation informatique ?), nous avons été amenés à développer une plate-forme articulée en deux volets. Le premier, orienté vers la production de documents XML, permet aux membres de l'équipe d'intervenir sur le corpus pour l'annoter selon des angles d'approche différents. Le second, orienté vers la diffusion des documents ainsi préparés, permet au lecteur de tirer le meilleur parti de l'encodage XML.

Sommaire :

I – L'objet du corpus

a) Les sermons latins du XIII^e siècle

Si cette littérature, extrêmement abondante[3], peut sembler ennuyeuse ou être rébarbative au premier abord, elle s'avère regorger d'informations sur les méthodes de communication, les représentations mentales et la culture médiévales. Les prédicateurs dans leurs sermons usent d'un langage métaphorique, qui prend appui sur le savoir et sur l'expérience des diverses catégories sociales pour transmettre le message religieux. Ils recourent aussi à de nombreux *exempla*, petits récits anecdotiques par lesquels ils illustrent leur propos, le font plus sûrement mémoriser, et cherchent ainsi à modeler les comportements de leurs auditeurs[4], ce qui contribue encore à faire de ces textes une clef indispensable à la compréhension des processus d'élaboration, de diffusion et de réception de la culture commune des hommes du temps.

---

[1] L'édition électronique du Thesaurus des sermons de Jacques de Voragine est un projet pilote de l'UMR 5648 – Histoire et archéologie des mondes chrétiens et musulmans médiévaux (Université Lumière Lyon 2 / CNRS/ EHESS), développé avec l'appui financier de l'Institut universitaire de France, dont Nicole Bériou est membre senior. Il dispose d'un site web : http://www.sermones.net. Il a fait l'objet de deux présentations publiques : l'une, le 26 juin 2003 à Turin, au cours de la rencontre « Medioevo in rete. Studi medievali e iniziative digitali fra i due versanti alpini », coordonnée par Reti medievali et par le Département d'Histoire de l'Université de Turin ; l'autre à Leeds, le 16 juillet 2003, lors de l'International Medieval Congress, session 1206 : « Indexing and Encoding the Sermons of Jacopo da Varazze ». Il a obtenu un appui spécifique du CNRS, depuis la fin de l'année 2003, dans le cadre du programme interdisciplinaire « Histoire des savoirs », comme un élément du projet intitulé « Écrits pragmatiques et communication au Moyen Âge ».

[2] Je remercie Nicole Bériou d'avoir bien voulu relire cet article, et de m'avoir fait bénéficier de ses remarques.

[3] Dans le répertoire de Schneyer, plus de 100 000 textes de sermons rédigés en latin sont mentionnés pour la période entre 1150 et 1350, dont 60 000 environ sont attribués (J.-B. SCHNEYER, *Repertorium der Lateinischen Sermones des Mittelalters für die Zeit von 1150-1350*, Münster, Westfalen, 1971-1990, 11 vol.)

[4] Voir Cl. BREMOND, J. LE GOFF, J.-C. SCHMITT, *L'exemplum*, Turnhout, Brepols, 1982, réimpression avec mise à jour 1996 (Typologie des sources du Moyen Âge occidental, 40) ; *Les exempla médiévaux : nouvelles perspectives*, J. BERLIOZ et M.A. POLO DE BEAULIEU, dir., Paris, Champion, 1998.

Mais plus encore, c'est une nouvelle manière de structurer le discours qui caractérise les sermons latins du XIII^e siècle. Leur construction, héritière des techniques élaborées pour l'exégèse, repose sur un plan plus strict qu'aux siècles précédents. Il est fondé sur le commentaire méthodique d'un *thema*, c'est-à-dire une phrase généralement tirée de la Bible, prétexte à développements. Le discours est également structuré par l'usage de la *distinctio*, technique rhétorique et outil didactique consistant à « distinguer » les propriétés d'un élément pour les ordonner en schéma et les commenter. Ce qui permet d'affirmer que « comme les cathédrales, les sermons latins du XIII^e existent par la densité, la fermeté et la richesse de leur architecture »[5].

b) Les sermons de Jacques de Voragine

Le corpus particulier des sermons de Jacques de Voragine comporte environ 700 textes, distribués en quatre séries composées après 1267, et peut-être seulement à partir de 1277 (*sermones de sanctis, de tempore, quadragesimales* et *Liber Marialis*) couvrant l'ensemble du cycle liturgique[6]. L'œuvre de ce Dominicain génois, surtout connu pour être l'auteur de la *Legenda aurea*[7], est nourrie de sa propre expérience de prédicateur et de dignitaire chargé de diverses responsabilités au sein de son ordre. À travers son action dans les fonctions importantes qu'il a été amené à exercer[8], on devine l'attention qu'il portait au recrutement et à la formation des frères. Les collections de sermons modèles qu'il a laissées constituent, avec la *Légende dorée*, sa contribution aux instruments de travail sur lesquels s'appuyaient les frères pour composer leurs prêches. Le succès très large et durable de ces collections témoigne de leur intérêt, de l'influence qu'elles ont pu avoir sur la formation des prédicateurs[9].

Le projet de Thesaurus des sermons de Jacques de Voragine se veut distinct d'une entreprise d'édition critique : il s'agit d'équiper d'outils d'interrogation un texte qui a été établi à partir d'une édition moderne accessible et dont les erreurs manifestes (le rendant incompréhensible ou anachronique) ont été corrigées à l'aide de quelques manuscrits fiables[10], s'inspirant en cela de la méthode recommandée par L. J. Bataillon[11]. L'édition critique, résultant de la collation systématique d'une sélection de manuscrits et publiée sous la forme traditionnelle d'une impression sur papier, est d'ailleurs menée parallèlement, à son rythme, grâce au concours de G. P. Maggioni, membre actif de l'équipe pour tout ce qui relève des questions philologiques.

Une grille d'analyse systématique des sermons a donc été élaborée par l'équipe. Elle prévoit la saisie, pour chaque texte du corpus, de méta données permettant de situer le sermon dans son contexte de production (auteur, occasion liturgique, référence scripturaire du *thema*, collection d'appartenance, etc.) et d'analyse (témoins utilisés pour l'établissement du texte, responsabilités, date, etc.). D'autre part, la grille prévoit le repérage et la description d'une série d'éléments structurels, rhétoriques ou narratifs dans le corps même du texte : éléments de plan, *distinctiones* et leurs membres, *exempla*, *figurae* – dont font partie les métaphores –, passages concernant la liturgie, matière biblique, noms propres, interprétation des noms, et sources déclarées, c'est-à-dire *auctoritates* sur lesquelles l'auteur appuie son discours.

La volonté de l'équipe est donc d'offrir, à travers ce Thesaurus, un outil pratique pour la recherche sur la prédication médiévale, qui soit un instrument d'investigation à l'usage de tous ceux qui s'intéressent à la culture médiévale dont nous sommes encore, à bien des égards, les lointains héritiers. Et c'est tout naturellement que, dès les premiers pas du projet, l'option de l'édition électronique a été choisie, afin de tirer le meilleur profit possible du texte numérique d'une part, et des différentes couches d'analyse apportées par l'équipe d'autre part.

c) Les questions de l'édition électronique

---

[5] N. BERIOU, « Les sermons latins après 1200 » dans *The Sermon*, B.M. KIENZLE, dir., Turnhout, Brepols, 2000 (Typologie des sources du Moyen Âge occidental, 81-83), p. 379.

[6] Voir S. BERTINI GUIDETTI, *I sermones di Iacopo da Varazze : Il potere delle immagini nel Duecento*, Florence, SISMEL, edizioni del Galluzzo, 1998; C. CASAGRANDE, « Iacopo da Varazze », dans *Dizionario biografico degli Italiani*, vol. sous presse.

[7] Iacopo da Varazze, *Legenda aurea*, éd. critique par Giovanni Paolo Maggioni, Florence, SISMEL, edizioni del Galluzzo, 2 vol., 1998. Une traduction française annotée, sous la direction d'Alain Boureau, vient de paraître en mars 2004 aux éditions Gallimard, coll. de la Pléiade ; voir aussi : *De la sainteté à l'hagiographie. Genèse et usage de la Légende dorée*. Études réunies par B. Fleith et Fr. Morenzoni, Genève, Droz, 2001.

[8] Il a été probablement *lector* du couvent de Gênes, puis certainement prieur de la province de Lombardie en 1267, et il a assumé la charge de vicaire général de son ordre de 1283 à 1285. C'est en 1292 qu'il devient, par nomination pontificale, archevêque de Gênes, jusqu'à sa mort en 1298.

[9] S. Bertini Guidetti a compté, dans les seuls répertoires de Schneyer et de Kaeppeli, 253 manuscrits des *Sermones de sanctis*, 522 des *Dominicales*, 327 des *Quadragesimales*, et 76 du *Liber Marialis*. Cela sans parler des éditions imprimées, qui se sont poursuivies jusqu'au XIX^e siècle.

[10] Le texte qui a été saisi, puis contrôlé sur manuscrits, est celui de l'édition faite par Robert Clutius au XVIIe siècle, dans la version réimprimée au XVIII^e siècle sous le titre : *Iacobi de Voragine ordinis praedicatorum Sermones aurei*, Augsbourg et Cracovie, 1760, 2 tomes en 4 vol.

[11] L. J. Bataillon « Les problèmes de l'édition des sermons et des ouvrages pour prédicateurs au XIII^e siècle », dans *The Editing of Theological and Philosophical Texts from the Middle Ages*, M. ASZTALOS, dir., Stockholm, 1986, p. 105-120

Richesse du contenu, importance de la structure : pour l'édition électronique de ce corpus, le choix du format XML[12] et des technologies qui lui sont liées s'est d'emblée imposé[13]. S'engager dans une telle entreprise soulève néanmoins un certain nombre de questions : les choix préalables consistaient à définir une méthode de travail et à choisir des outils adaptés, puis à s'orienter vers une stratégie d'encodage.

Un éditeur XML adapté ?

Tout d'abord, la recherche d'un éditeur XML adapté aux besoins de l'équipe s'est révélée problématique. Les outils existants se sont avérés inadaptés aux besoins. Il nous fallait permettre aux membres d'une équipe européenne, donc géographiquement très éclatée, d'intervenir sur les textes du corpus pour les annoter et les analyser, selon un aspect particulier dont ils sont spécialistes (*exempla* pour les uns, *distinctiones* pour les autres, etc.). Nicole Dufournaud, après avoir travaillé en collaboration avec un informaticien (J.-D. Fekete) sur le codage XML d'un corpus de lettres de rémission du XVI^e^ siècle, reconnaissait : « Nous avons eu énormément de difficulté à trouver un environnement de travail convenable pour SGML/XML et TEI. Nous avons fini par utiliser un environnement de développement informatique presque identique à celui utilisé par des gros projets de développements informatiques (Emacs, CVS, validation, transformation de programmes, etc.) Si des historiens doivent utiliser de tels environnements, une formation spécifique leur est indispensable. »[14] On imagine les difficultés qu'aurait représenté la mise en place d'un tel système, qui suppose un lourd apprentissage et la mise en place d'une chaîne de circulation des documents contraignante, au niveau d'une équipe de recherche complète…

Les choix d'encodage

Pour l'encodage du corpus, nous nous sommes d'emblée intéressés à la DTD[15] proposée par la *Text Encoding Initiative* (ou TEI)[16]. La TEI est en quelque sorte une grammaire, qui propose des jeux de balises et des recommandations méthodologiques sur leur usage (les *guidelines*), appuyées sur l'expérience partagée par d'autres utilisateurs de la DTD. Elle offre des facilités liées à la capitalisation d'expériences diverses[17]. Mais le fait de s'appuyer sur la TEI n'exonère bien évidemment pas d'une réflexion sur la stratégie et les choix d'encodage. Au sein même de la TEI, les *guidelines* peuvent donner le choix entre différentes méthodes pour l'approche d'un même problème, chacune présentant un jeu différent d'avantages et de contraintes[18]. Certains choix peuvent modifier en profondeur la structure des fichiers XML produits : ainsi l'adoption de l'encodage *standoff*[19], utilisé pour la représentation des éventuelles hiérarchies concurrentes[20]. Deux démarches d'analyses scientifiques comparables, portant sur des corpus de même nature, pourraient donc aboutir à des modélisations informatiques différentes. Comment alors se doter d'outils communs, et éviter que des contraintes liées à des choix techniques ne viennent gêner l'interopérabilité de textes correspondant aux mêmes questionnements pour les historiens ?

La nécessité de répondre à ces questions nous a rapidement amenés à envisager le développement d'outils informatiques spécifiques, pour accompagner l'étape de constitution et d'enrichissement du corpus, puis pour permettre la diffusion et l'exploitation de celui-ci.

II – ScolastiX : Application collaborative pour la préparation du corpus

---

[12] BRAY T., PAOLI J., SPERBERG-McQUEEN C. M., eds, *Extensible Markup Language (XML) 1.0*, Cambridge, World Wide Web Consortium, 1998. [**En ligne**] http://www.w3.org/TR/REC-xml.

[13] XML est particulièrement adapté aux documents semi-structurés : ce format permet de croiser les bénéfices de l'interrogation *full text* et d'une base de données classique.

[14] Jean-Daniel FEKETE et Nicole DUFOURNAUD, « Analyse historique de sources manuscrites : application de TEI à un corpus de lettres de rémission du XVI^e^ siècle », *Document Numérique*, 3, 1-2, 1999, numéro spécial « Les documents anciens », p. 117-134

[15] DTD : *Document Type Definition* ou *Définition de Type de Document*. Une DTD définit, pour une classe donnée de documents XML, quels sont les entités, éléments et attributs pouvant être utilisés, et fixe les règles de leur agencement.

[16] Lou Burnard, C. M. Sperberg-McQueen, eds, *TEI P4 : guidelines for Electronic Text Encoding and Interchange (XML-compatible edition)*, 2 vol., University of Virginia Press, 2002. Site web : http://www.tei-c.org/

[17] De nombreux projets impliquant des sources médiévales utilisent déjà la TEI : le *Canterbury Tales* Project (http://www.cta.dmu.ac.uk/projects/ctp/), ou le projet Charrette (http://www.mshs.univ-poitiers.fr/cescm/lancelot/index.html). Certains s'en sont fait l'écho dans le Médiéviste et l'Ordinateur, par exemple : Anna Mette HANSEN, « Text encoding of manuscripts : Danish prayer books from the 16th century », *Le Médiéviste et l'Ordinateur*, 41, 2002. [En ligne] http://lemo.irht.cnrs.fr/41/mo41_09.htm

[18] Pour l'encodage de l'apparat critique, par exemple, les *TEI guidelines* exposent trois méthodes différentes pour lier l'apparat au texte. V. Lou Burnard, C. M. Sperberg-McQueen, eds, *TEI P4: guidelines*…, chap. 19.

[19] Henry S. THOMPSON and David MCKELVIE. « Hyperlink semantics for standoff markup of read-only documents », dans *Proceedings of SGML Europe '97*, Barcelona, Spain, May 1997.

[20] Dans un document XML, les données sont structurées sous la forme d'un arbre unique, ce qui pose problème pour représenter des niveaux d'annotation n'ayant pas vocation à être organisés de manière strictement hiérarchique, et susceptibles de se superposer. Diverses méthodes permettent de contourner cette difficulté. L'encodage externe au texte (voire au document), ou *standoff markup*, est l'un d'entre elles.

a) Les lignes directrices

Un outil ouvert et souple

La volonté[21] première est de proposer une application permettant l'annotation collaborative d'un corpus de textes : les membres de l'équipe doivent pouvoir intervenir en parallèle sur les documents, sans avoir à imposer une chaîne de traitement contraignante. Nous souhaitions également offrir un outil qui offre en quelque sorte une « couche d'abstraction » entre la problématique scientifique posée par les historiens et la (ou les) modélisations informatiques pouvant en découler.

La solution adoptée se rapproche du *standoff*. Elle repose sur un stockage séparé des textes sources et des informations relatives à leur analyse scientifique (v. fig. 1) :

- Les textes du corpus subissent un premier encodage en XML lors de leur chargement dans l'application. Cet encodage, très basique, est réalisé automatiquement : il consiste en un balisage des phrases et des mots, basé sur la reconnaissance de la ponctuation. Chaque phrase et chaque mot reçoivent un identifiant unique, incrémenté séquentiellement. Ces textes sont stockés en lecture seule sur un serveur central.
- Les annotations, entrées par les utilisateurs via un client web, sont stockées dans une base de données relationnelle. En plus de l'analyse, du commentaire et des éléments divers, chaque annotation est rattachée à un point ou à un passage du texte source, grâce à l'identifiant du début et de la fin de l'extrait concerné. Le travail simultané et parallèle sur le corpus est ainsi rendu possible, et il est également très facile pour les utilisateurs de reprendre leur travail pour le compléter ou le modifier.
- Une fois le travail sur un texte achevé, il est exporté : un parseur fusionne le texte source (débarrassé du premier encodage automatique) avec les informations issues de la base de données. Cette fusion peut être réalisée selon des règles paramétrables (nom des balises et attributs correspondant à chaque type d'information, règles d'agencement des éléments,…), correspondant donc à des DTD différentes. Ce système permet de donner une certaine indépendance à la grille d'analyse scientifique et à sa traduction en XML, simplifiant à la fois l'évolution d'une DTD pendant le processus d'enrichissement du corpus, et l'adoption d'une application identique par des projets à la problématique similaire mais travaillant avec des DTD distinctes.

[21] ScolastiX : ce nom, choisi pour désigner cette application, correspond à l'acronyme de *Système Collaboratif Libre pour l'Annotation Scientifique de Textes et d'Images en XML*.

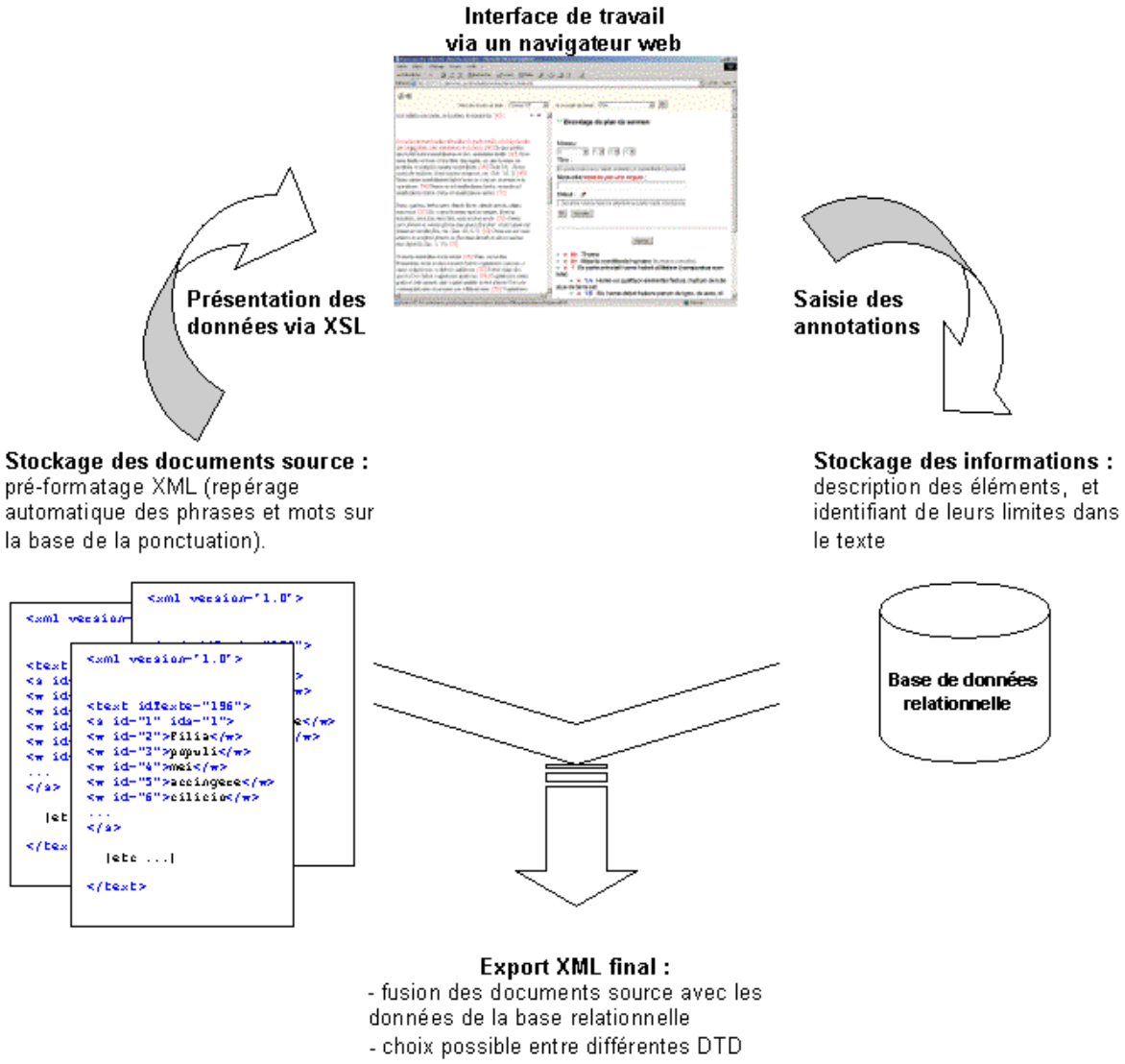


Figure 1 –Schéma général de l'application

Un outil adapté aux pratiques

En plus de ces caractéristiques techniques, nous souhaitions développer une application qui corresponde aux pratiques propres aux équipes de recherche.

Il fallait avant tout une interface de travail ergonomique, permettant une prise en main rapide sans un lourd apprentissage. Notre choix s'est porté sur une interface « tout web » : la saisie des annotations se fait via un navigateur, au travers de formulaires familiers aujourd'hui à tous les utilisateurs d'Internet. Le chercheur n'a donc pas à se préoccuper de la façon dont sont organisées et représentées les données.

Pour répondre au caractère international de l'équipe, nous avons voulu intégrer le support du multilinguisme. Tous les textes de l'interface ont donc été préparés en trois langues (français, anglais et italien), mais séparés de la présentation HTML elle-même. À chaque langue correspond un fichier de variables, contenant l'ensemble des expressions utilisées dans l'interface. L'ajout d'une nouvelle langue est donc simplifié : il suffit de traduire ce fichier dictionnaire.

Un outil libre

Nous souhaitions enfin pouvoir partager le plus largement possible le fruit de nos développements.

Les choix technologiques ont été guidés par ce désir : l'application est basée sur le langage de script côté serveur PHP[22] et la base de données MySQL[23]. Les compétences nécessaires à l'intégration de ces outils sont aujourd'hui très largement répandues parmi les informaticiens, ce qui assure un transfert de compétences aisé. Il est également très facile de trouver un hébergement supportant PHP et MySQL, aussi bien dans le monde académique que chez des hébergeurs privés.

Nous avons enfin souhaité nous inscrire dans une démarche Open Source : dès qu'elle aura atteint un niveau de stabilité suffisant, l'application sera offerte au téléchargement sur Internet, sous licence GPL[24]. Nous espérons d'une part partager des expériences avec d'autres équipes scientifiques, mais également bénéficier de l'apport de la communauté des développeurs du libre, pour les futures évolutions de la plate-forme.

b) L'interface de travail : une visite guidée

Voyons, au travers de quelques exemples, comment s'organise le travail au sein de l'application.

Un environnement de travail en commun

Les utilisateurs sont rattachés à des groupes disposant de droits définis : chaque groupe est responsable de 1 à n angles de travail sur le corpus (par exemple, pour l'analyse des sermons de Jacques de Voragine, un groupe est chargé des plans et des méta données, un autre des *distinctiones*, un troisième des *exempla*, etc.). Chaque groupe est lui-même géré par un ou plusieurs administrateurs. La **figure 2** montre la page d'accueil présentée à un administrateur de groupe, après qu'il se soit identifié. Le cadre du bas est occupé par un système très basique de messagerie. Le cadre du haut est une sorte de barre de menu ; les trois boutons du bas n'apparaissent que pour les administrateurs de groupe, ceux du haut sont communs à tous les membres.

Figure 2 : écran d'accueil

En tant qu'administrateur, on peut utiliser le système de messagerie pour envoyer des informations aux membres de son propre groupe ou à tous les collaborateurs. On est autorisé à gérer les membres de son groupe en modifiant des comptes existants, en invitant des utilisateurs d'autres groupes ou en créant un nouveau compte. Enfin, on a la possibilité de choisir les angles de travail accessibles aux membres du groupe.

Si l'on s'intéresse aux trois boutons du haut, accessibles à tous les utilisateurs, on remarque qu'ils sont disposés selon le cycle de vie du document au sein de l'application :

[22] http://www.php.net
[23] http://www.mysql.com
[24] http://www.gnu.org/copyleft/gpl.html

- « Gérer les textes » : un sermon est tout d'abord ajouté au corpus, chargé au format texte brut, et converti dans un pré-encodage XML. Si nécessaire, l'ajout du texte peut être validé par un administrateur[25]. Lors du chargement d'un nouveau texte, la saisie des méta données essentielles est rendue obligatoire.
- « Étudier un texte » : il s'agit du cœur de l'application. L'utilisateur, lorsqu'il entre dans cette section, est invité à choisir un texte, et un angle de travail. Pour simplifier la prise en main, l'interface d'annotation a été ramenée à des éléments bien maîtrisés par tous les utilisateurs d'Internet (**figure 3**).

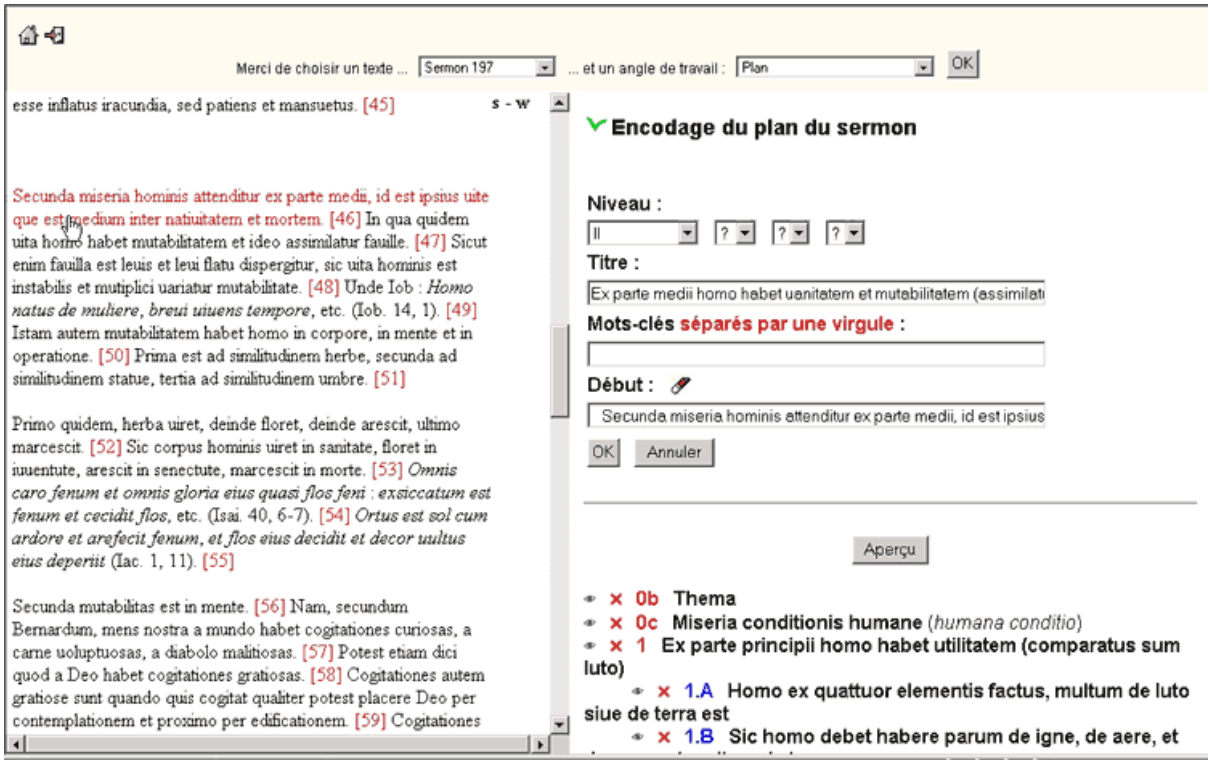


Figure 3 : un angle de travail

- L'écran est divisé en trois parties. En haut, des menus permettent de passer d'un texte à l'autre et/ou d'un angle à l'autre. À gauche, le texte de base est présenté grâce à une transformation XSL. À droite, un formulaire, différent pour chaque angle. Un simple clic de souris sur les mots ou phrases du texte permet d'attacher l'annotation à une partie du document (renseigne les champs « Début » et « Fin »).
- Dès ce stade, le contrôle de l'annotation est possible. Tout d'abord, chaque nouvelle entrée génère une ligne au pied du formulaire, pour rappeler ce qui a été saisi. L'utilisateur peut à partir de là supprimer les entrées erratiques, ou en demander un aperçu. À travers un jeu de vues dynamiques, on peut accéder soit à un aperçu général, donnant l'ensemble des entrées insérées à leur place dans le corps du texte, soit à un aperçu individuel (**figure 4**)

[25] Cette fonctionnalité, non implémentée pour notre équipe, a été prévue pour le cas d'équipes impliquant par exemple à la fois des chercheurs et des étudiants.

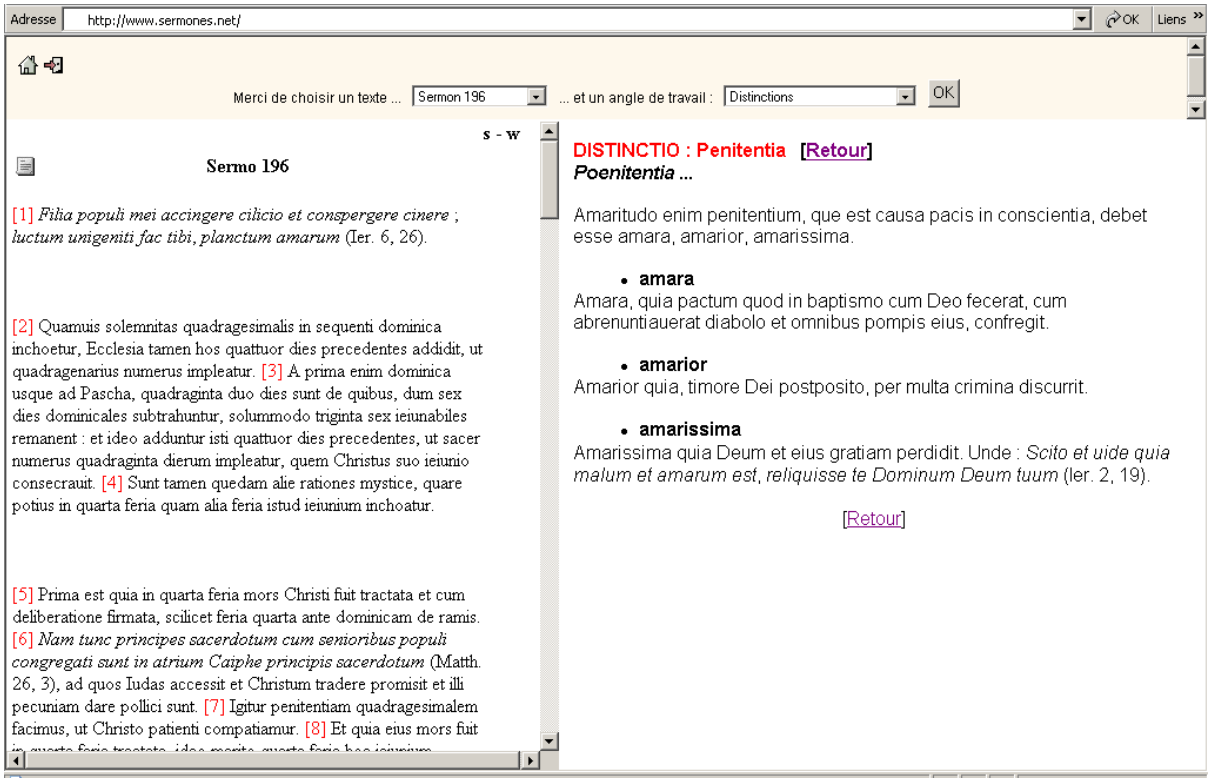


Figure 4 : l'aperçu individuel dynamique d'une distinctio

- Enfin, les utilisateurs disposent d'outils pour le suivi du travail. Un bilan est proposé pour chaque texte, listant sur une seule page les annotations saisies pour chaque angle de travail. Un autre bilan donne une vue d'ensemble du traitement du corpus ; il se présente sous la forme d'un tableau à double entrée montrant quels angles de travail ont été complètement traités pour chaque texte.
- « Exporter un texte » : l'utilisateur choisit un texte, et un schéma d'encodage parmi ceux proposés. Il peut ensuite télécharger le document XML produit

Une fois le corpus constitué et annoté, on dispose donc d'une série de documents XML. Il est alors nécessaire de proposer des outils pour tirer parti de l'encodage.

III – Un exemple d'application pour la diffusion du corpus

L'application présentée ici est avant tout un prototype exploratoire, préparé afin de montrer les possibilités ouvertes par XML. Sa réalisation définitive ne sera achevée qu'après une campagne de développement. Son fonctionnement illustre néanmoins une partie des attentes quant à l'exploitation des textes.

a) Quelques points techniques

Comme pour ScolastiX, l'application d'annotation collaborative, le choix de l'Open Source s'est imposé pour la diffusion et la valorisation du corpus.

Sans préjuger des choix lors de la prochaine campagne de développement, un prototype très sommaire a été réalisé en utilisant la plateforme SDX[26]. Il s'agit d'un outil libre, basé principalement sur Tomcat[27], Cocoon[28] et Lucene[29], et initialement développé sous l'impulsion de la mission de la recherche et de la technologie du Ministère de la Culture. Son principal avantage est une approche très documentaire : de nombreuses fonctions facilitent l'indexation, la recherche avancée, et permettent de trier les résultats par pertinence, etc. Il permet surtout d'indexer d'une façon uniforme des documents aux DTD différentes, et donc des corpus potentiellement différents.

[26] http://adnx.org/sdx/
[27] http://jakarta.apache.org/tomcat/
[28] http://cocoon.apache.org/
[29] http://jakarta.apache.org/lucene/docs/index.html

En revanche, l'indexation réalisée par SDX « aplanit » les documents XML : contrairement aux possibilités offertes par les bases de données XML natives telles qu'eXist[30] ou Xindice[31], on ne peut donc pas interroger la structure des documents (i.e. par exemple rechercher tous les documents présentant une certaine imbrication d'éléments, dans un ordre donné). Le couplage de SDX avec de telles bases pourrait néanmoins offrir ces possibilités dans l'avenir.

Le caractère très exploratoire du prototype actuel n'appelle guère plus de précisions techniques, le propos étant une simple démonstration de l'usage possible d'un corpus de textes XML.

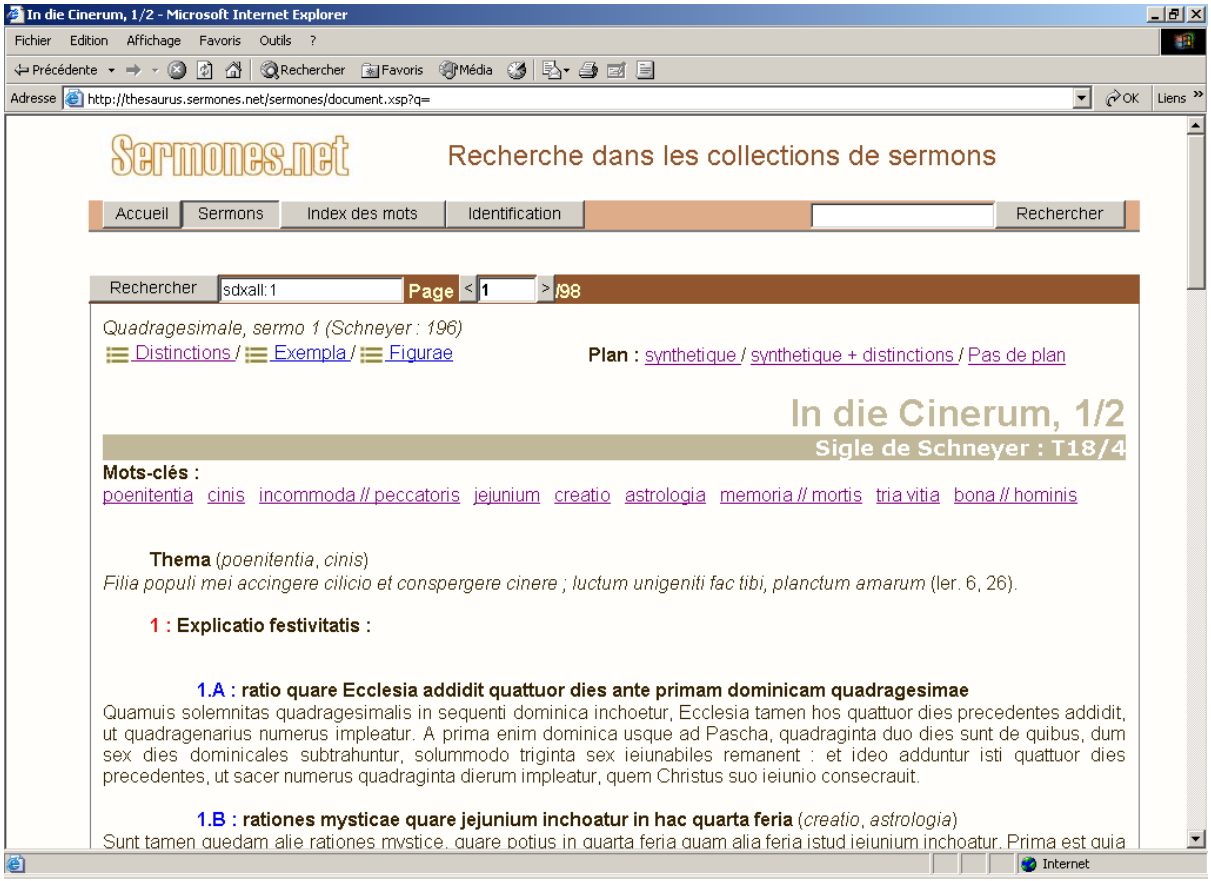
b) L'exploitation des textes

L'utilisateur dispose de modes d'accès aux textes :

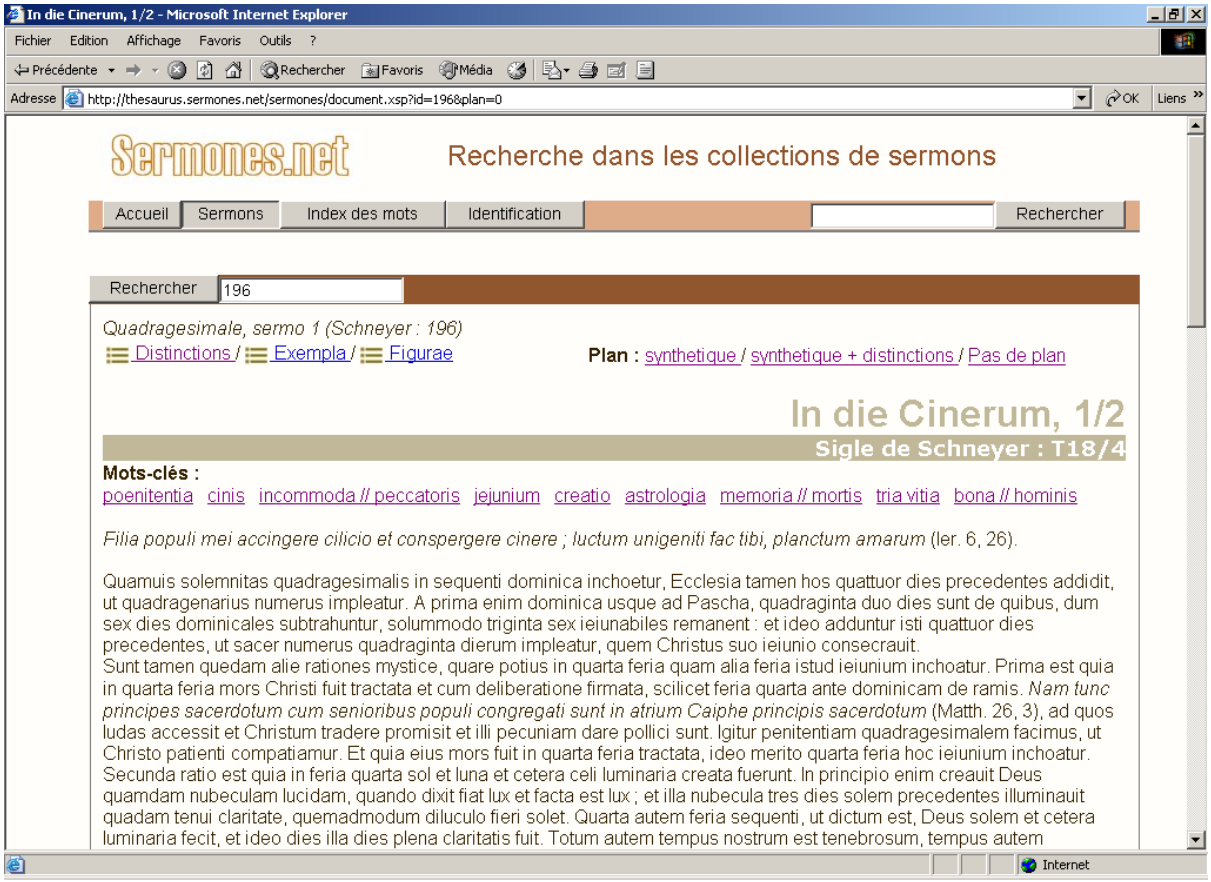
- Le premier et le plus simple est la lecture systématique, en « feuilletant » le corpus d'un document à l'autre, comme on feuilletterait un livre. Les méta données attachées à chaque texte permettent de proposer des parcours de lecture diversifiés, à travers une collection donnée par exemple, ou selon des axes thématiques.
- Un jeu d'index permet également d'approcher le corpus par les éléments qu'il contient. Pour les sermons, ce sont des index des références scripturaires, des *exempla*, des *distinctiones*, des *figurae*, une liste d'*incipit*, etc.
- Enfin, l'utilisateur dispose d'une recherche multicritère avec opérateurs booléens certes classique, mais aux possibilités étendues par l'encodage XML des documents : on peut bien entendu chercher un mot ou une expression dans le texte complet, mais on peut également cibler sa requête, en cherchant un mot ou une expression au sein d'un élément particulier (on peut imaginer par exemple une recherche de tous les documents comportant un *exemplum* contenant les mots « auarus » et « infernus »). Pour faciliter la construction des requêtes et l'usage efficace des caractères génériques, on peut consulter une liste des formes de mots utilisées dans le corpus. Il s'agit bien évidemment d'un pis aller, dans l'attente de la mise en place d'outils linguistiques adaptés à la langue latine.

À l'échelle du document individuel, le lecteur dispose de divers moyens de personnaliser la forme du document. Grâce à XML, ce n'est pas le créateur d'un corpus qui en fixe arbitrairement la présentation ; l'utilisateur a la liberté de l'adapter, pour chaque texte consulté, afin qu'elle corresponde au mieux type d'accès à l'information qui l'intéresse. Le plan, par exemple, est affiché par défaut dans le corps des sermons, pour en faciliter la lecture. Un utilisateur souhaitant avoir un accès plus direct à la source peut demander un affichage du texte seul du sermon, sans aucune mention de plan ajoutée. À l'inverse, il peut demander un affichage donnant une version synthétique du plan en en-tête du texte (une sorte de table des matières), et même une version où s'intercalent les éléments de plan et les distinctions, pour avoir une approche schématique de la structure du document (**figure 5** a, b, c et d).

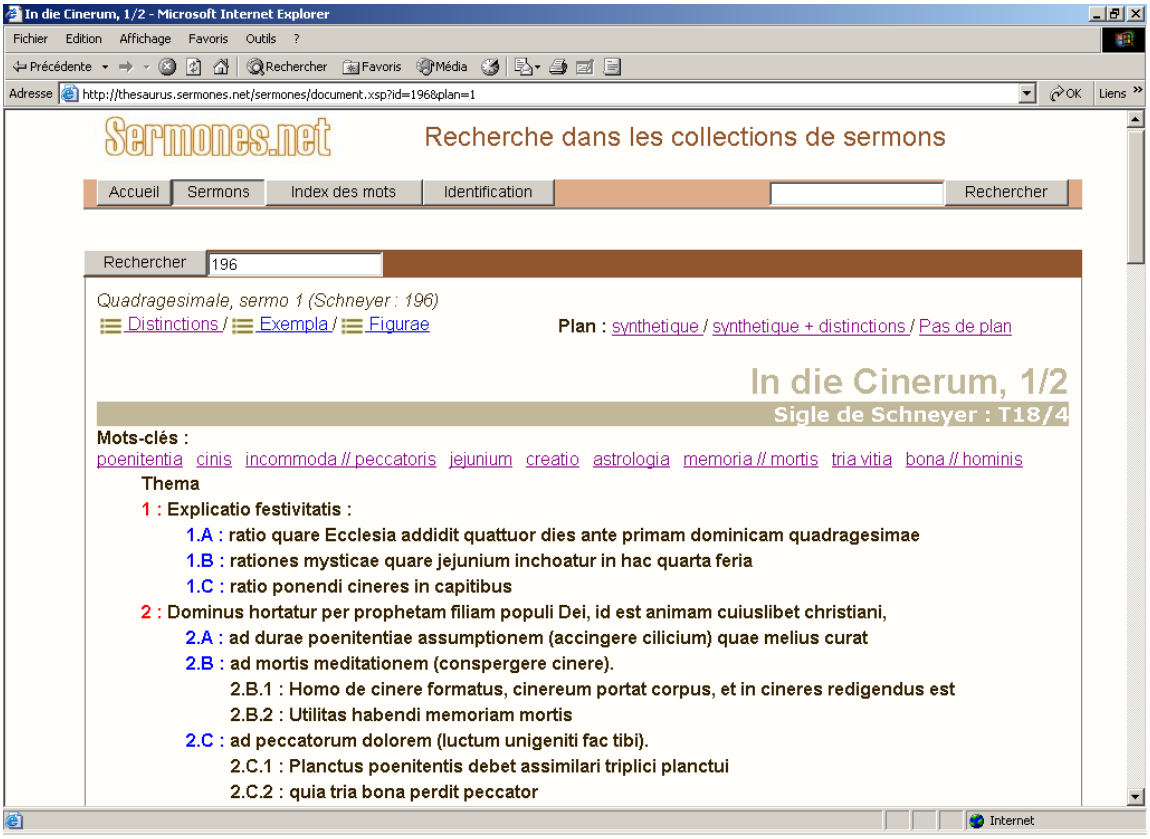
[30] http://exist.sourceforge.net/
[31] http://xml.apache.org/xindice/



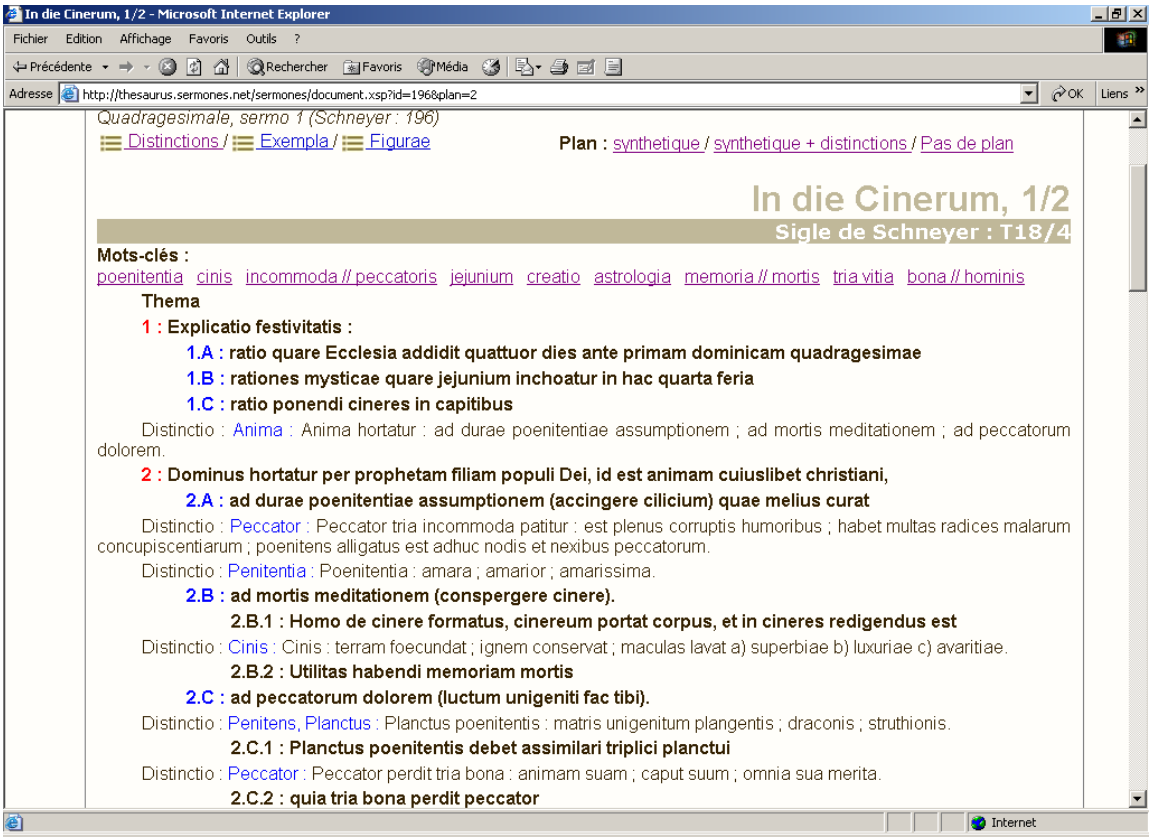
5. a : vue par défaut, plan dans le texte



5. b : le texte seul



5. c : aperçu synthétique du plan



5. d : plan synthétique croisé avec les *distinctiones*

Figure 5 : un document et son plan, quatre choix de présentation

L'utilisateur peut choisir de mettre en valeur d'autres éléments. Une liste des *distinctiones*, des *exempla* ou des *figurae* du texte courant lui est aisément accessible. Le lecteur peut alors se rendre directement à l'emplacement d'une *figura*, marqué ici par une manicule et un cartouche en marge (**figure 6**).

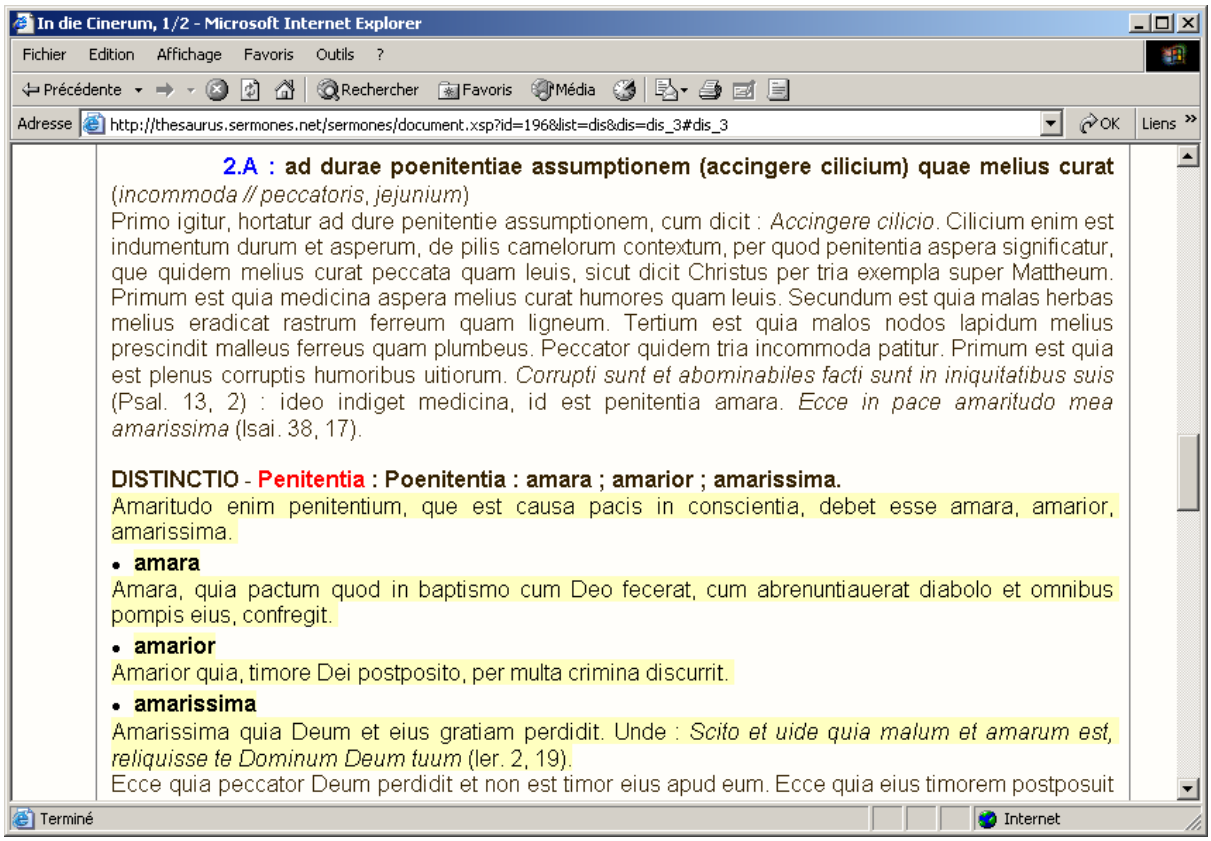


Figure 6 : mise en valeur d'une distinctio

S'il désire avoir les détails d'un *exemplum* ou d'une *distinctio*, il peut faire apparaître cet élément dans son contexte, assorti d'annotations et surligné pour une lecture plus facile (**figure 7**).

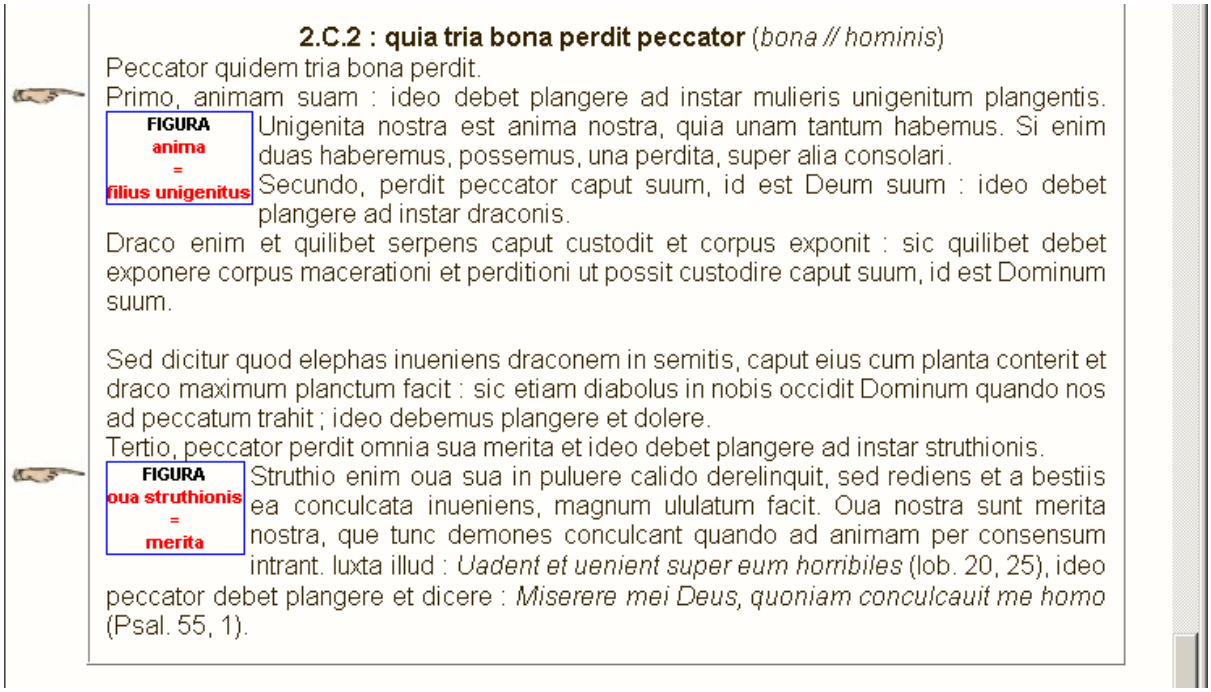


Figure 7 : mise en valeur des figurae

Si la conception des prototypes décrits ici a été initiée autour d'un corpus de sermons, notre souhait est d'ouvrir la réflexion à d'autres types de sources, afin de faire de ScolastiX une plate-forme générique pour l'annotation de sources, accompagnant l'étape de constitution d'un corpus de sources, quelle que soit leur nature et la problématique.

Le soutien apporté par le CNRS depuis la fin de l'année 2003 au projet « Écrits pragmatiques et communication au Moyen Âge », dont les applications décrites ici constituent le volet informatique, a permis cette ouverture. Ce projet vise à mettre au point une méthode de travail scientifique d'analyse et de valorisation de documents écrits appartenant à des champs typologiques divers (en l'espèce, les sermons d'une part, les comptes de châtellenie d'autre part), mais susceptibles d'être soumis à un traitement informatique élaboré sur des bases communes.

Au terme d'une campagne de développement actuellement en cours (février – août 2004), le prototype de ScolastiX doit céder la place à une première version publique en Open Source.

Bibliographie : ouvrages cités

BATAILLON L. J., « Les problèmes de l'édition des sermons et des ouvrages pour prédicateurs au XIII^e siècle », dans *The Editing of Theological and Philosophical Texts from the Middle Ages*, M. ASZTALOS, dir., Stockholm, 1986, p. 105-120.

BERIOU N., « Les sermons latins après 1200 », dans *The Sermon*, B. M. KIENZLE, dir., Turnhout, Brepols, 2000 (Typologie des sources du Moyen Âge occidental, 81-83).

BERLIOZ J. et M. A. Polo de BEAULIEU, dir., *Les exempla médiévaux : nouvelles perspectives*, Paris, Champion, 1998.

BERTINI GUIDETTI S., *I sermones di Iacopo da Varazze : Il potere delle immagini nel Duecento*, Florence, SISMEL, edizioni del Galluzzo, 1998.

BRAY T., PAOLI J., SPERBERG-MCQUEEN C. M., eds, *Extensible Markup Language (XML) 1.0*, Cambridge, World Wide Web Consortium, 1998. [**En ligne**] http://www.w3.org/TR/REC-xml

BREMOND C., LE GOFF J., SCHMITT J.-C., *L'exemplum*, Turnhout, Brepols, 1982, réimpression avec mise à jour 1996 (Typologie des sources du Moyen Âge occidental, 40).

BURNARD L., SPERBERG-MCQUEEN C. M., eds, *TEI P4 : guidelines for Electronic Text Encoding and Interchange (XML-compatible edition)*, 2 vol., University of Virginia Press, 2002.

CASAGRANDE C., « Iacopo da Varazze », dans *Dizionario biografico degli Italiani*, vol. sous presse.

De la sainteté à l'hagiographie. Genèse et usage de la Légende dorée. Études réunies par B. FLEITH et Fr. MORENZONI, Genève, Droz, 2001.

FEKETE J.-D. et DUFOURNAUD N., « Analyse historique de sources manuscrites : application de TEI à un corpus de lettres de rémission du XVI^e siècle », *Document Numérique*, 3, 1-2, 1999, numéro spécial « Les documents anciens », p. 117-134.

HANSEN A. M., « Text encoding of manuscripts : Danish prayer books from the 16th century », *Le Médiéviste et l'ordinateur*, 41, 2002. [**En ligne**] http://lemo.irht.cnrs.fr/41/mo41_09.htm

IACOPO DA VARAZZE, *Legenda aurea*, éd. critique par Giovanni Paolo MAGGIONI, Florence, SISMEL, edizioni del Galluzzo, 2 vol., 1998.

JACQUES DE VORAGINE, *La légende dorée*, nouvelle traduction, introduction et notes, sous la dir. d'Alain BOUREAU avec Monique GOULLET et la collaboration de Pascal COLLOMB, Laurence MOULINIER et Stefano MULA. *La Légende dorée et ses images*, par Dominique DONADIEU-RIGAUT, Paris, Gallimard, 2004 (coll. de la Pléiade, 504).

KAEPPELI T., *Scriptores Ordinis Praedicatorum Medii Aevi*, 1-4 (vol. 4 by Emilio Panella), Rome, 1970-1993.

SCHNEYER J.-B., *Repertorium der Lateinischen Sermones des Mittelalters für die Zeit von 1150-1350*, Münster, Westfalen, 1971-1990, 11 vol.

THOMPSON H. S., MCKELVIE D. « Hyperlink semantics for standoff markup of read-only documents », dans *Proceedings of SGML Europe '97*, Barcelona, Spain, May 1997.