

Laurent Romary et Henri Hudrisier.

Le balisage normalisé des concepts et documents en liaison avec les normes de l'EAD.

Résumé:

Dans ce papier, nous tenterons d'analyser comment peut se concevoir à distance la quasi-totalité des activités d'enseignement-recherche qui caractérisent un troisième cycle y compris sa phase de production doctorale. Nous nous attacherons à montrer comment certaines normes et standards des technologies de l'information se complètent pour répondre aux fonctionnalités de la gestion, d'aménagement de travail pédagogique sur des documents, des concepts, des références appartenant à des disciplines d'une faculté d'art, lettres et sciences humaines. Nous nous appuyerons sur les travaux entrepris à l'université Paris-VIII, à la MSH Paris Nord et au Loria en concordance avec nos activités d'expert en normalisation.

Abstract:.....
.....

Dans ce papier, nous tenterons d'analyser comment peut se concevoir à distance la quasi-totalité des activités d'enseignement-recherche qui caractérisent un troisième cycle y compris sa phase de production doctorale. Nous nous attacherons à montrer comment certaines normes et standards des technologies de l'information répondent très complètement aux fonctionnalités de la gestion voire de l'aménagement des concepts et des références dans les disciplines d'une faculté d'art et sciences. L'appréhension des concepts et des références de sa discipline d'abord de façon extensive, puis selon des choix de plus en plus ciblés en fonction des hypothèses de recherche du futur doctorant (ou DESS) constitue en effet l'essentiel du travail que nous considéreront comme fonctionnellement de 2^{ème} cycle avant la phase fonctionnelle de 3^{ème} cycle : l'écriture de la thèse proprement dite¹. Nous nous situons dans la dynamique des travaux coordonnés au niveau d'un groupe de travail de l'Université Paris 8 animé par la présidente de son conseil scientifique. Nous choisissons aussi de prendre pour corpus les écrits et enregistrements des cours de Deleuze qu'une équipe a pour projet d'organiser en bibliothèque virtuelle et dictionnaire électronique des concepts. Ce projet implique en synergie des étudiants et enseignants du département Philosophie mais aussi d'Informatique, d'Hypermédia et de Documentation en liaison avec l'AUF, la Maison des Sciences de l'Homme de Paris Nord et le Loria. Ce dernier intervient notamment pour l'intégration de la TEI (Text Encoding Initiative²) des normes d'ingénierie linguistiques notamment l'ISO-TC37 qui interagiraient avec les normes de l'EAO en cours de développement (ISO/IEC-JTC1-SC36). Il est prévu d'utiliser le même environnement technique et normatif dans deux autres projets : « Études berbères » et « Étude comparée des traductions de Rimbault ».

Le contexte de développement d'une telle proposition

Ces journées de réflexion initi@tives 2003 à Versailles sont nées d'un double constat. Nombre d'universitaires, notamment, et c'est bien normal, ceux qui appartiennent aux sciences humaines, aux arts et aux lettres, évaluent souvent mal l'importance de la montée en puissance de la normalisation dans le nouveau système technique qui se

¹ Nous donnons ici une acception très fonctionnelle aux trois cycles universitaire (voir plus bas paragraphe consacré aux « Trois cycles fonctionnels de l'étudiant »).

² On peut lire en français sur ce sujet le numéro 24 des Cahiers Gutenberg, IRISA, Rennes, France.

met en place actuellement, notamment dans les TIC. Symétriquement, nombre de techniciens ou d'industriels, acteurs du développement de ces techniques, ou même certains les experts impliqués dans la définition de ces normes et standards ignorent, ou feignent quelquefois d'ignorer, les difficultés de compréhension, donc d'appropriation, des nouveaux paradigmes du système technique émergent qui caractérise les TIC (et donc les TICE pour ce qui nous concerne plus particulièrement).

Cette incompréhension réciproque est très dommageable pour le développement des dernières générations de NTIC, mais elle s'explique assez facilement. Le système technique des TIC est actuellement dans une phase de transformation radicale que l'on pourrait résumer comme étant celle de l'émergence du Web sémantique et de l'information numérisée, balisée, structurée, normalisée et convergente (donc interopérable et multimédia).

Or, ce renouveau technique exigerait des efforts d'appropriation importants de la part des utilisateurs et des efforts de vulgarisation de la part de l'offre industrielle et normative. Cependant, ni les industriels, ni les experts en normalisation ni d'un autre côté, la majorité des utilisateurs ne semblent avoir amorcé un rapprochement significatif.

De là découle cette incompréhension réciproque et la non-appropriation des techniques (et plus grave le ralentissement du développement d'applications) dans le secteur de l'information structurée et normalisée.

Cependant ce qui rend cette révolution technologique particulièrement intéressante c'est qu'elle implique que les contenus pédagogiques et les processus d'application s'intègrent et se développent technologiquement. C'est aussi ce qui rend la dernière phase de progrès technologique des TICE particulièrement difficile à s'approprier, parce que les développements techniques ne se feront plus désormais qu'en synergie avec une instrumentation numérique raisonnée des objets, des textes et des documents pédagogiques ou à vocation pédagogique.

Cette spécificité des documents à vocation pédagogique est importante : beaucoup de décideurs pédagogiques ont bien compris que les cours en lignes devaient être numérisés et qu'il était important que ce travail doive se faire avec un souci de « ré-utilisabilité » donc de normalisation.

Néanmoins, ils paraissent moins bien comprendre, ou plutôt feignent de ne pas comprendre que les bibliothèques, mais aussi les musées, les encyclopédies, etc.... sont aussi des ressources sinon pédagogiques, à tout le moins «à vocation pédagogique »³.

C'est précisément dans cette jachère relative que nous prétendons œuvrer : la plupart des spécialistes de l'EAD raisonnent fondamentalement sur des ressources pédagogiques spécifiques. Il nous semble cependant, et cela semble correspondre à la réalité statistique mondiale,⁴ qu'il est en effet très difficile de dépasser le niveau de la phase fonctionnelle propédeutique (1^{er} cycle) avec de l'enseignement à distance constitué de ressources exclusivement spécifiques : cela tient au fait évident que, à partir du 2^{ème} cycle et a fortiori en 3^{ème} cycle on demande à l'étudiant d'être inventif donc imprévisible en matière de demande des ressources pédagogiques. Seule la totalité extensive des références de la discipline qu'il aborde (et bien souvent aussi d'une, ou de plusieurs autres discipline(s)) peuvent répondre à son besoin

³ Sur ce point les bibliothécaires et les spécialistes de l'e-learning partagent le même point de vue.

⁴ La canadienne P8 \$\$\$

d'exploration. Les bibliothèques, les musées, la littérature grise, les résultats bruts des chercheurs, les archives, les interviews, les recueils d'enquêtes, les mass-médias, mais aussi nombre de ressources et corpus originaux et inattendus constituent ainsi la matière pédagogique fondamentale d'un étudiant de 2^{ème} et 3^{ème} cycle.

Il se trouve que depuis une quinzaine d'année l'organisation référentielle, structurelle et conceptuelle des corpus de documents (au départ essentiellement textuel) est le but unique de la TEI. Il se trouve aussi que l'approfondissement de tels objectifs a profondément influencé les concepteurs des systèmes d'informations ce qui fait qu'un certain nombre de fonctionnalités d'XML de normes de l'ingénierie linguistiques, de structure de MPEG7 ou de fondements du Web sémantique s'appuient sur les principes de la TEI ou des principes similaires.

C'est en raisonnant selon la même logique nous prétendons qu'il est faisable et souhaitable d'organiser les étudiants de 2^{ème} et 3^{ème} cycle (surtout en sciences humaines mais pas seulement) sur des logiques d'EAD⁵ en demandant à ces étudiants de travailler sur des ressources numériques originales et en leur demandant de les référencer, de les structurer et de les baliser sémantiquement (en organiser les concepts, les acteurs, les lieux, la temporalité...).

Les universitaires de sciences humaines les technologies de l'EAD et la normalisation.

Il semble bien normal que nombre d'universitaires (sauf ceux qui sont directement impliqués dans la recherche en informatique ou en IST) ne soient pas complètement informés des derniers développements en matière de technologies d'information et en particulier d'EAD. On peut même constater que chez certains d'entre eux, notamment dans les matières littéraires à première vue moins facilement formalisables dans des langages formels, nous assistons souvent un véritable déni d'usage de toutes les techniques de médiation informatique. Il faut cependant observer que la proportion des ressources d'EAD en sciences exactes et expérimentales est sans commune proportion avec les quelques exemples d'application aux sciences humaines.

Cependant si nous quittons le monde explicite de l'EAD pour celui de l'information en général nous pouvons alors disposer de ressources littéraires ou de sciences humaines déjà très considérables et cependant très inégalement exploitées.

De nombreuses archives textuelles se mettent en place (Gallica, OTA, WWP, Frantext, etc.). De très nombreux CD-ROM ou sites Internet structurés et de recherche ont été développés dans des secteurs historiques, littéraires, philosophiques, sociologiques, psychanalytiques, dans le secteur des langues (y compris anciennes⁶)

Les bibliothécaires sont eux aussi de très anciens usagers des technologies de l'information et notamment ce sont des promoteurs de normes. Ils savent depuis des millénaires que la bibliothèque ne se construit pas sans normes et on sait tous que ce sont les documentalistes de l'aéronautique qui sont à l'origine de SGML et des principes mêmes de l'information numérisée, balisée, structurée et normalisée.

⁵ Essentiellement une logique plate-forme collaborative.

⁶ Citons par exemple le « Thesaurus linguae graecae » qui contient la quasi exhaustivité de la littérature grecque classique.

De son côté, la TEI qui fonde les grandes lignes de notre démarche compte parmi ses membres actifs nombre de conservateurs de bibliothèques⁷.

Des démarches spécifiques en sciences humaines

En effet, nombre d'activités académiques, notamment dans les disciplines littéraires (sciences sociales et humaines, recherche littéraire, histoire, histoire de l'art...), mais aussi, à un certain niveau, dans toutes les sciences, ont pour fondement premier de travailler sur des corpus de documents (des textes, mais aussi des images, des notations musicales, des audiovisuels ou des langages formels mathématiques par exemple) d'en analyser la forme, le style, les références, les concepts fondamentaux, la sémantique pour conforter, déduire ou opposer les hypothèses qui constituent le premier stade de la production d'une publication scientifique.

Chacun sait en effet que ce qui distingue la publication scientifique académique des activités ordinaires de communication (essais, vulgarisation scientifique journalistiques...) c'est précisément l'acceptation d'un contrôle paritaire de conformité des concepts, des références et d'exhaustivité (quand cela est possible) du corpus de documents qui sous-tendent les hypothèses et les raisonnements et thèses soutenues. On pourrait, en ce sens, les qualifier de production scientifique normalisée.

Il nous semble dès lors pertinent, que, nous appuyant sur cette qualité normative des productions de la littérature scientifique, nous puissions en profiter pour les assimiler à des ressources pédagogiques potentielles. Pour y parvenir, il faudra cependant repenser les interfaces habituellement mises en œuvre dans l'EAD et avant toute chose, proposer des scénarios de mise en œuvre et d'usage.

Les trois cycles académiques fonctionnels de l'étudiant

Malgré des différences notables entre les universités de différents pays, un étudiant qui suit un cycle universitaire complet passera nécessairement par 3 cycles d'études successifs.

1. Une phase propédeutique préalable destinée à lui faire acquérir par imitation des maîtres et des condisciples des savoirs, des méthodes, des savoir-faire... L'étudiant doit aussi se familiariser avec le catalogue élémentaire de concepts et le corpus minimum des documents de référence d'une discipline. C'est le 1^{er} cycle qui relève plus de la consommation organisée de cours et exercices. *Cette phase peut s'accommoder parfaitement de ressources pédagogiques traditionnelles organisées préalablement sans attention particulière à telle ou telle catégorie d'étudiant autre que leur niveau ou leur degré de spécialité par rapport au sujet traité.*
2. Une phase d'appréhension approfondie des concepts et des références de la discipline (ou des disciplines connexes) de façon d'abord extensive. C'est en s'investissant dans ce type d'exercices que l'étudiant de 2^{ème} cycle acquiert graduellement une aptitude à la recherche⁸ (ou pour le moins une aptitude à

⁷ ainsi que des informaticiens documentalistes impliqués dans la mise au point de SGML, qui se sont ensuite remobilisés dans le projet TEI et qui sont à l'origine du développement d'un langage aussi important que XML (Michael Sperberg McQueen, Steve de Rose, Henry Thompson etc.). Ce sont encore les mêmes qui inspirent les grandes lignes de ce qui se développe sous le nom très prometteur de Web sémantique.

⁸ Dans le meilleur des cas ces premières explorations et structuration de domaine sont très correctement menées et aboutissent à des publications qui peuvent être éditées pendant la phase administrative du 2^{ème}

l'autonomie intellectuelle par rapport à la discipline et aux métiers qui y sont attachés). Le travail de l'enseignant consiste précisément à guider l'étudiant dans cette prise progressive d'autonomie d'abord en suggérant des thèmes d'exploration dans des sujets réputés inexplorés. Pour les disciplines de sciences humaines qui ont la langue et l'écriture comme modalité quasi-unique de rassemblement des sources, raisonnement et communication, le territoire à explorer et à s'approprier est fondamentalement un univers de textes⁹. Il existe un moyen très efficace (largement partagé par nombre d'enseignants-chercheurs) pour guider les étudiants et les inciter à s'approprier cette nécessaire autonomie. L'enseignant-chercheur distribue aux étudiants des sous-ensembles des sujets de sa propre recherche ou de celles du laboratoire auquel il appartient. Il peut aussi suggérer des sujets inédits qu'il aimerait voir explorés. *La médiation électronique et le potentiel de fonctionnement à distance de ce 2^{ème} cycle est au cœur de ce qui nous apparaît innovant dans notre approche. L'étudiant sous la conduite collégiale de ses professeurs doit progressivement s'impliquer dans des exercices pratiques de structuration des textes (ou des documents). Notre projet consiste précisément à ce que les exercices cumulés de plusieurs promotions d'étudiants puissent constituer les grandes bibliothèques virtuelles structurées en TEI dont toutes les matières de sciences humaines auraient besoin mais qui ne trouveront jamais des financements spécifiques*¹⁰. (nous reviendrons sur la méthode)

3. Une troisième phase fonctionnelle (mais aussi académique) qui est le 3^{ème} cycle. Progressivement, l'étudiant a affiné ses choix, élaboré des hypothèses originales, sélectionné un sujet d'étude le plus original et le plus inédit possible. Il est prêt pour la rédaction d'une thèse¹¹. Cette rédaction constitue dès lors l'essentiel du travail de l'étudiant-chercheur. Bien sûr pendant sa phase de rédaction l'étudiant continue des activités que nous classons comme relevant du 2^{ème} cycle fonctionnel. Il peut (il doit), affiner la structuration de son propre corpus de texte et documents. Cependant, il est vraisemblable que contrairement à la phase de 2^{ème} cycle il diffèrera la mise en ligne des travaux d'approfondissement de son corpus pour préserver l'effet de nouveauté de son travail de thèse. Ces détails (qui ne sont pas sans importance) ne changent pas fondamentalement l'organisation fonctionnelle générale du projet. Il est cependant évident qu'un tel dispositif d'EAD destiné au 2^{ème} et 3^{ème} cycle universitaire devra bien sûr prévoir des dispositifs de sécurité et de confidentialité permettant de réserver ces espaces et ces phases de communication restreintes (confidentialité réservée dans le temps du corpus, confidentialité possible du dialogue entre le doctorant et son directeur de recherche par exemple). *La médiation électronique et le potentiel de*

cycle académique : par exemple en France un mémoire de maîtrise. Dans notre conception fonctionnelle des cycles nous appelleront déjà ces publications précédant la thèse du 3^{ème} cycle fonctionnel.

⁹ Tous les raisonnements que nous tenons peuvent être élargis (sans grands changements méthodologiques) aux disciplines du visuel et de l'audiovisuel : histoire de l'art, étude cinématographiques, sciences du spectacle... Les normes et techniques de structuration et balisages existent (citons MPEG 7 notamment.)

¹⁰ Les plans de numérisation systématiques tels qu'ils sont développés par de grandes institutions (en France Ministère de la culture et ministère de l'éducation) sont fondamentaux pour élargir la masse critique des documents numériques disponibles, mais fondamentalement, ils ne peuvent en aucun cas approfondir le traitement structurel ni surtout conceptuel des ressources numérisées. Ils répondent à l'adage du bibliothécaire qui dit que «Si la bibliothécaire devait lire tous les livres, la bibliothèque n'ouvrirait jamais.

Par contre, des ressources numériques accumulées selon les principes que nous proposons seront bien évidemment très largement balisées tant sur le plan structurel que sur le plan conceptuel.

¹¹ Ou par exemple en France d'un mémoire de DESS qui bien que non obligatoirement soumis à l'obligation de recherche devra développer un point de vue autonome et personnel sur le sujet d'étude.

fonctionnement à distance du 3^{ème} cycle fonctionnel fait aussi partie de notre projet mais nous semble beaucoup plus classique. On se retrouve dans un dispositif de contrôle éditorial relativement conventionnel. Le doctorant se voit proposé une feuille de style et des instructions aux auteurs. Il serait aussi utile que les enseignants (notamment le directeur de recherches) puissent disposer d'un environnement logiciel sophistiqué (mais dont il existe aujourd'hui de nombreuses briques pas forcément intégrées) lui permettant de consulter les états successifs d'avancement de la thèse, de disposer d'outils «d'annotations suggestions» à destination du doctorant (voire de d'annotations qu'il voudrait garder confidentielles pour lui-même). L'enseignant (mais aussi le doctorant) pourra ainsi se concentrer plus facilement sur le contrôle d'avancement de telle ou telle partie de la thèse, n'avoir à corriger que les ajouts ou modification explicitement mises en évidence, etc.... On peut aussi de cette manière envisager plus rationnellement une cogestion de thèse qui partage entre deux directeurs de recherche deux facettes distinctes d'une même rédaction.

Plus précisément, les 3 cycles fonctionnels ou académiques correspondent ainsi à 3 modes pédagogiques et collaboratifs, ainsi qu'à trois types de mise en œuvre logicielle.

Pour le 1^{er} cycle : un type d'EAO relativement traditionnel dans lequel c'est l'enseignant qui distribue à distance des curriculum de cours correspondant aux savoirs qu'il veut transmettre. On peut mettre en œuvre tous les types d'exercices, de contrôle et de validation d'acquis que l'on jugera nécessaire.

Pour le 2^{er} cycle : un type d'EAO fondamentalement différent. Les étudiants sont impliqués à titre individuel dans la maîtrise progressive d'un univers de textes (et document) classiques¹², connus ou à découvrir. A ce titre ils doivent :

A) **Individuellement ils doivent structurer ces textes** à différents niveaux¹³ :

- aménager des métadonnées (auteurs, titre, édition, concepts abordés, thématiques),
- organiser structurellement (divisions, paragraphes, appréhender les notes, l'index, la bibliographie, les avertissements, préfaces et postfaces),
- réaliser des annotations de surface (concepts, personnes, lieux, dates, structures stylistiques ou argumentatives, mise en évidence du lexique employé...)

B) **Collectivement les étudiants (et solidairement les enseignants-chercheurs)** sont aussi collectivement les acteurs d'un plus vaste projet, et ils participent à ce titre à l'avancement de l'instrumentation numérique de leur discipline¹⁴. En élargissant la collégialité des participants de la TEI organisé en vaste campus virtuel qui devra être instrumenté sur des plates-formes d'EAO paramétrées à cet effet, on saura se

¹² Pour des raisons évidentes liées notamment à l'efficacité de l'apprentissage par l'imitation, il sera plus pertinent de confier des textes classiques en début de 2^{ème} cycle et réserver la structuration des corpus à découvrir pour les périodes terminales du 2^{ème} cycle voire la phase de travail sur corpus propre à la phase de thèse (3^{ème} cycle).

¹³ Niveaux qu'ils pourront personnellement mais aussi collectivement découvrir progressivement.

¹⁴ Il n'y a rien là d'exceptionnel (ni même trace d'exploitation comme on a pu nous le dire parfois), les étudiants des sciences expérimentales sont «naturellement intégrés» dans des projets collectifs d'intérêt général pour l'avancement de leur discipline. Les étudiants en biologie génétique peuvent être mobilisés collectivement pour la cartographie du génome, des étudiants en physique pour explorer systématiquement des proches de bombardement d'électrons, etc....

partager et planifier la cartographie cohérente et structurée de l'exhaustivité extensive des textes et documents de référence pour une discipline donnée. A partir de cette première évaluation des grandes masses, on convient en consensus de normes de structuration et de conceptualisation (dictionnaire électronique de la discipline). Ce travail de consensus est déjà très avancé chez les participants de la TEI. Dès lors on peut se situer au niveau plus appréhendable d'un département universitaire (par exemple le Département de philosophie de Paris 8¹⁵) et envisager plus directement la gestion EAD des travaux de numérisation et structuration des documents par les étudiants. On peut donc distinguer :

- 1) *trois méta-fonctions cognitives qu'il convient de partager* :
 - Une cartographie générale du domaine (il existe des outils graphiques classiques).
 - Un environnement fonctionnel dédié à la structuration normalisée des textes et documents qui doit être évidemment paramétrée en fonction de la discipline voire de l'école de pensée dans laquelle on se situe (c'est le domaine d'excellence de la TEI).
 - Un environnement fonctionnel dédié à la mise en œuvre d'un dictionnaire électronique du domaine qui permettra de centraliser les concepts, voire d'autre item comme des structures, des patterns linguistiques argumentatifs spécifiques, etc.... (c'est le domaine d'excellence des normes ISO-TC37.¹⁶)
- 2) *Des fonctions de partage des tâches de "numérisation-structuration", de contrôle collaboratif par les étudiants, de contrôle scientifique par le collègue enseignant.* Ces fonctions sont dans le droit fil d'un mode de travail collaboratif. Elles devraient être assez facilement mises en œuvre sur une plate-forme d'enseignement telles que nous les connaissons aujourd'hui qui peuvent être paramétrée pour permettre :
 - 1) La gestion des travaux de chaque étudiant (et la capacité d'auto-contrôle) pour évaluer et former à une appropriation par niveaux de «profondeur de la structuration » puis guider l'étudiant dans la concaténation des phases successives de numérisation puis de balisage d'un texte (ou document).
 - 2) La gestion d'ensemble d'une promotion d'étudiants qui collaborent à un « grand œuvre de numérisation d'un domaine » par accumulation des contributions individuelles.
 - 3) La gestion des procédures d'auto-contrôle, de contrôle réciproque par les étudiants en double ou triple-aveugle, de contrôle scientifique par les enseignants chercheurs.

Ces points sont fondamentaux. Il faut absolument imaginer à moyen terme des procédures permettant d'activer tous les modes possibles de contrôle de conformité. Pour le niveau des formelles d'un livre il ne devrait pas y avoir trop de problèmes et le contrôle réciproque en double ou triple-aveugle devrait donner d'assez bons résultats (c'est déjà beaucoup). Le niveau de balisage conceptuel est plus sujet à caution. La aussi, le contrôle croisé par les étudiants, préalable au contrôle scientifique par le collègue enseignant serait d'un grand recours. On se rend compte sur ces points qu'on

¹⁵ Ou bien sûr d'un réseau de département de philosophie réunis thématiquement ou fédérés pour des raisons linguistiques.

¹⁶ Ou d'un outil comme HYDRO développé au LORIA, outil de terminotique s'appuyant d'ailleurs sur des structures de dictionnaire fondées sur la TEI.

reproduit à 7 à 8 siècles de distance la pratique universitaire médiévale du «disputatio».

Notons que pour ce qui est de la construction collective d'un corpus de textes de référence l'université médiévale appliquait un mode d'implication des étudiants proche du notre. Avant l'invention de l'imprimerie, la force de travail des étudiants était mobilisée pour fournir une part conséquente de la recopie des œuvres classiques. Cette recopie était partagée : les exemplaires servant à la recopie étaient distribués entre plusieurs copistes en autant de petits cahiers rassemblant quelques dizaines de pages ; le stationnaire¹⁷, généralement attaché à une université, reconstituait l'œuvre à partir des recopies collectives après en avoir contrôlé la conformité. Selon cette méthode il était possible de livrer une œuvre complète d'Aristote en quelques jours.

Toutes ces fonctionnalités techniques sont fondamentales à terme. De la sophistication des protocoles de contrôle et des automatismes que l'on saura mettre en œuvre dépendra la viabilité (ou non) d'un tel projet. On perçoit bien que le niveau d'intégration d'ensemble soit assez complexe. Cependant de cette totalité d'intégration de «briques logiques» qui isolément existent déjà ou existent sous des formes voisines, dépendra l'instrumentation sur une large échelle de disciplines qui actuellement ne sont pratiquement pas instrumentalisées numériquement. Quand les ressources numériques existent (parce qu'il y a une croissance considérable de ce type de ressources) elles sont très rarement (et pour cause) balisées conceptuellement. Et pourtant, ce balisage structurel et conceptuel serait très utile à l'avancement (qui peut devenir exponentiel de ce fait) de ces disciplines. La numérisation extensive, rendu possible par la mobilisation cumulée de promotions successives d'étudiants, permettra d'atteindre des masses critiques de documents touchant à l'exhaustivité des documents pertinents pour une discipline. Cet enjeu, on n'aurait jamais osé le rêver pour des disciplines de sciences humaines : la philosophie ne disposera jamais du budget de recherche de la biologie génétique fondamentale.

Pour le 3^{ème} cycle : un type d'EAO qui redevient assez traditionnel, mais très focalisé sur un type de fonctionnement traditionnel et presque unique puisqu'il s'agit de permettre le contrôle par un enseignant particulier (le directeur de recherche¹⁸), d'un seul exercice, la thèse¹⁹, réalisé sur la longue durée (3 ans en principe). Les procédures proprement pédagogiques sont assez simples et nous les avons déjà décrites.

La technicité du dispositif doit se déployer par contre dans l'outil d'encadrement de la rédaction (problème bien maîtrisé par les éditeurs), les dispositifs de mises en évidence et de gestion temporelle et logique des états successifs et des ajouts et retraits, les dispositifs d'annotations à plusieurs niveaux : annotations «pense-bêtes» du doctorant, annotations suggestion du directeur de thèse vers l'étudiant, annotations personnelles privées de l'enseignant (ou par la suite d'un membre quelconque du jury).

Une économie de l'enseignement et de la recherche en sciences humaines

Utiliser les nouvelles technologies dans l'enseignement et la recherche ne peut pas, ne doit pas, se traduire par une augmentation inconsidérée des dépenses d'équipement informatique et de production de cours que ne compenseraient pas des gains de

¹⁷ Nom qui a donné en anglais *stationary* : papetier

¹⁸ Ainsi qu'éventuellement un codirecteur de recherche et in fine un jury.

¹⁹ Mais aussi bien d'autres catégories de mémoires (maîtrise, DESS, DEA pour la France)

productivité pédagogique ou de recherche. La productivité peut s'entendre à plusieurs niveaux :

- Soit, on augmente significativement le potentiel d'acquisition de savoir ou la qualité des acquisitions
- Soit, on participe du potentiel d'auto-apprentissage sur le *net* comme « bien public universel »
- Soit, on participe d'un avancement significatif de la discipline en créant un environnement numérique qui rend « calculable » un certain nombre de paradigmes et de facettes d'appropriation de la discipline. Notons sur ce point qu'en rendant numérique, tangible et collaborative la « paillasse de laboratoire²⁰ » d'un centre de recherche en sciences humaines on pense faire évoluer significativement la dynamique synergique de la recherche en sciences humaines.
- Soit, on augmente significativement le potentiel de pédagogie et de recherche à distance.

Ce dernier point est particulièrement important pour les disciplines comme les sciences humaines pour lesquelles (contrairement aux sciences exactes ou expérimentales), il est exceptionnel (en France tout au moins) de pouvoir financer les années de thèses d'un doctorant. Cela a pour conséquence la proportion importante d'abandons qui ne frappent pas obligatoirement les moins bons doctorants mais frappent d'abord les étudiants étrangers ou ne résidants pas dans la ville universitaire considérée.

Notre projet nous paraît respecter ces règles d'équilibre fondamentales.

Dès lors nous disposons du système d'information nécessaire et suffisant pour concrétiser un enseignement à distance au niveau des 2^{ème} et 3^{ème} cycle pour toute discipline travaillant essentiellement sur des textes (les images, sons, audio visuel et quantité d'autres documents pouvant être traités en utilisant des méthodes dérivées des mêmes principes).

Nous disposons aussi, ce faisant, d'un environnement de ressources permettant de produire de façon beaucoup plus efficace des synthèses pédagogiques (cours électroniques) destinées au 1^{er} cycle. En développant quelques aménagements il doit être aussi possible d'impliquer des étudiants dans les tâches préparatoires de la fabrication de cours, ainsi d'ailleurs que dans les phases de tests de ces mêmes cours.

La mise en place d'un tel dispositif peut paraître complexe au premier abord. En réalité la plupart des briques logicielles existent déjà. La TEI qui constitue le cœur du dispositif est à la fois une pratique collégiale existant depuis plus de 10 ans et une norme technique d'organisation des textes. Un certain nombre de postes de « lecture-écriture savante », de dispositifs d'alignement multilingue et de pointage de vocabulaire sont d'ors et déjà disponibles.

Dans ces conditions il faut savoir évaluer à leur juste prix ce que coûte un enseignement doctoral (notamment pour des étudiants loin de leur lieu - ou pays- de résidence), évaluer le nombre des doctorants qui abandonnent leur thèse parce qu'ils se sentent isolés dans le rassemblement de leurs documents de référence, ou parce qu'ils n'ont pas complètement intégré les méthodes de l'exercice doctoral ou même parce qu'ils perdent concrètement contact avec leur directeur de thèse.

²⁰ La mise en place d'une telle paillasse virtuelle aura ainsi des effets sur la convivialité des départements de sciences humaines. La paillasse est en effet le point focal, l'endroit privilégié où se travaille la science, c'est donc le premier lieu de convivialité de la recherche en sciences expérimentales.

Il faut savoir aussi évaluer les facilités qu'entraînera un tel dispositif dans l'enseignement présentiel ou à distance.

Analyser formellement et conceptuellement les grands auteurs²¹ d'une discipline, baliser la structure argumentative des textes, le style, l'enchaînement des parties, appréhender la provenance et la pertinence des citations, relever les variations autour de tel ou tel concept... plonge instantanément l'étudiant au cœur du «métier de l'écrivain en sciences humaines». Cette appropriation nécessaire du métier d'écriture advient d'autant plus facilement que l'étudiant dispose d'un poste de travail paramétrable le guidant pour réaliser ce référencage, cette analyse des structures, des concepts, des figures stylistiques et argumentatives ; un outil aussi qui induit naturellement le doctorant à activer nécessairement l'apparat critique d'un ouvrage. Il nous paraît probable que le même dispositif appliqué aux sciences exactes et expérimentales aura des avantages évidents pour la phase d'écriture proprement dite de toute thèse (et du métier de chercheur comme producteur d'articles).

Examinons maintenant les différents niveaux de transmission raisonnée et contrôlée du savoir pouvant être rencontrés dans une discipline littéraire.

A) Un niveau d'appréhension de l'exhaustivité (idéale) des documents : l'ensemble des références autorisées d'une discipline. Il est constitué des références historiques que vient grossir la production actuelle de publications scientifiques du domaine. Remarquons que c'est à ce niveau que vient s'intégrer la production des thèses et autres mémoires de recherche (au minimum de maîtrise pour le système universitaire français).

B) Un niveau de maîtrise conceptuelle : Ce champ s'appuie explicitement (ou implicitement) sur un vocabulaire, ou plutôt des vocabulaires du domaine, rassemblés dans des dictionnaires de spécialité (et/ou des encyclopédies). Ces concepts, ces idées ne sont présentes dans ces dictionnaires que si elles trouvent occurrence dans les textes de référence ce qui permet de pointer depuis les différentes entrées du dictionnaire électronique vers autant de zones de référence qu'il y a d'occurrences et, ce faisant, de disposer à la demande des citations pertinentes confortant une idée ou un concept et lui donnant son épaisseur relative. Notons qu'en liant entre eux à un niveau multilingue (ISO-TC37-SC3) différents dictionnaires réalisés dans différentes langues correspondant à différentes écoles, les sensibilités culturelles propres à la discipline sont facilement mises en évidence.

On doit aussi savoir distinguer deux grandes catégories de productions scientifiques en sciences humaines et pour les thèses en particulier.

1) la publication de découvertes s'appuyant bien sûr à la base sur les grands classiques mais produisant de façon originale de nouveaux concepts, de nouvelles méthodes, de nouveaux champs d'investigation pour la discipline. Ces publications qui participent directement de l'innovation courent plus que les autres le risque de la critique paritaire négative. Le succès d'estime, en cas d'excellence de la réussite, est aussi à la mesure du risque.

2) des publications de compilation traitant de l'histoire d'une école de pensée, comparant plusieurs chercheurs, étudiant la génétique littéraire ou l'exégèse des grands

²¹ Ou même les auteurs plus marginaux et des sources d'information primaires (matériel oral, archives historiques, discours politiques, mass média...)

textes étudiant transversalement une option un concept une idée²² dans leur degré de généralité, d'emprunt croisé etc. ainsi que la critique des sources.

C'est la deuxième catégorie qui sera trop souvent conseillée aux étudiants en thèse même si on exige d'eux une certaine dose d'inventivité et d'innovation pour valider leur doctorat.

On pourra aussi classer dans cette deuxième catégorie les œuvres collectives (ou individuelles) que constituent les grandes synthèses thématiques historiques descriptives d'une école, d'un auteur, d'un courant de pensée ainsi que les grands dictionnaires. Ces grands dictionnaires viendront souvent couronner la carrière d'un auteur majeur en sciences humaines.

Scénario d'usage et de construction de la bibliothèque virtuelle des références et du dictionnaire électronique de concepts

La bibliothèque virtuelle des références

Conçue selon les principes de la TEI et en s'appuyant étroitement sur les modèles en cours de normalisation dans le TC36-SC3 & 4, la bibliothèque virtuelle des références de domaine est un des points forts du dispositif. Son originalité repose sur le fait que les étudiants ont pour mission de participer à la construction du référencement, et au balisage structurel et sémantique de cette bibliothèque virtuelle. Les enseignants impliqués dans le projet ont eux l'obligation d'orienter les grandes lignes de conception de ladite bibliothèque virtuelle et d'en contrôler la qualité. Ceci devra dès le début être l'objet d'un contrat explicite entre des promotions d'étudiants et des universitaires volontaires dans le projet. On sait en effet que s'il y a de nombreux projets de bibliothèques virtuelles dans le monde, ils butent sur 2 points faibles : les droits d'auteurs et l'énormité du travail de balisage. Beaucoup de projets se limitent de ce fait au niveau de mise en place des références en ne structurant que très formellement les documents. S'agissant des droits d'auteur (pour ce qui est des auteurs contemporains²³) il faudra autant que possible utiliser de façon mixte des ressources sur CD-ROM édité (dont on peut ainsi maîtriser la redistribution des droits) et «venir pointer» toute la valeur ajoutée TEI à partir du réseau

Pour ce qui est du 2^{ème} point faible, c'est lui qui justifie le recourt d'un tel dispositif pédagogique. De même qu'au Moyen-âge, les étudiants participaient dans une large mesure à la copie normalisée des manuscrits, notre époque doit savoir impliquer les étudiants dans la construction des bibliothèques virtuelles de référence, que l'on sait possible à construire mais que l'on ne sait pas institutionnellement entreprendre par le seul recours au travail salarié.

Il suffit simplement de bien découper fonctionnellement les tâches associées aux phases de développement technique des départements d'informatique, de documentation, de linguistique, de multimédia et de sciences de l'éducation, de sorte à concevoir ladite bibliothèque virtuelle comme un ensemble susceptible de croître par accumulation d'exercices de prise en compte des références bibliographiques et de structuration des dites références.

L'exemple du dictionnaire électronique philosophique

²² On se rapportera ici aux distinctions qu'établit clairement St Bonaventure (cité par Elyzabeth Eisentein, \$\$\$\$ p \$\$\$) entre les différentes catégories d'auteurs : \$\$\$\$\$\$.

²³ On devra aussi privilégier le travail sur des périodes hors droit d'auteur.

Partie intégrante de la bibliothèque virtuelle, un dictionnaire électronique du domaine disciplinaire est mis en place. Il est au cœur du dispositif de balisage, de référencement, de structuration. En effet, les étudiants sont incités à produire la structuration du corpus et à mettre en évidence par des liaisons référentielles ou thématiques, les auteurs, les écoles de pensées, les concepts, les catégories de raisonnement... Ce travail peut prendre pour base les dictionnaires attestés du domaine, par exemple pour le travail entrepris en philosophie à l'université de Paris 8 le dictionnaire de François Chatellet associé à beaucoup d'autres. Mais par ailleurs, ce travail doit partir pragmatiquement d'un travail de « levée de notions » directement alimenté par les textes pris en compte et structuré selon les modes opératoires des terminologies électroniques multilingues normalisées. Dans ce cas de figure, les dictionnaires-papier cités ci-dessus ne seront que les références (certes particulières) mais parmi toutes les autres références constitutives du corpus.

On observera aussi que le développement de ces dictionnaires électroniques participerait de fait au développement des ontologies de domaine qu'exigera la mise en œuvre d'un nouveau système technique de l'information structurée normalisée.

Notons que ces dictionnaires électroniques sont vraiment des outils très innovants au même titre que le sont les dictionnaires électroniques de la traduction automatique. Ces dictionnaires seront tout d'abord naturellement multilingues et multiculturels. Ils sont capables de proposer à un même niveau (ou, au choix, selon des niveaux hiérarchisés) des acceptions voisines, multilingues appartenant à une autre école de pensée... On peut éclairer un tel propos en comparant comme cela a déjà été fait les différentes formulations de la notion de « peuple » en sociologie ou en philosophie à différentes époques, dans différents pays ou langues, chez différents auteurs... Ainsi, *demos* n'équivaut pas à *peuple*, qui ne recouvre pas totalement la notion de *nation*, ni celle de *volk*, qui lui-même n'a pas le même sens pour un philosophe allemand contemporain des lumières, pour Nietzsche, ni bien sûr un auteur nazi.

Ces dictionnaires sont évidemment hypertextuels et hypermédia. Ils rendent très efficace la consultation des liens d'association ou de mise en parallèle des ressources, des fragments de textes, d'images, de structures²⁴, des schémas stylistiques ou argumentatifs, etc....

La publication de thèses collégalement contrôlée.

C'est une fonction importante parce que le suivi de thèse constitue l'activité fondamentale des années doctorales. Cette activité nécessite bien sûr beaucoup de contact présentsiels, mais on se rend compte que le courrier électronique a changé les choses dans une large mesure. Cependant, à l'évidence, le courrier électronique sans aménagement spécifique à cette activité n'est pas parfaitement adapté et, de ce fait, la plupart des enseignants de sciences humaines refusent de l'utiliser à cet effet. Les fonctionnalités qui rendraient le suivi électronique d'une thèse non seulement acceptable, mais aussi plus commode que le « suivi papier » sont facilement schématisables. Ce sont celle d'un poste de lecture-écriture, en fait un traitement de texte aménagé. Les fonctions nécessaires sont les suivantes :

- L'étudiant doit d'abord disposer d'instruction aux auteurs, ou plutôt un environnement convivial, d'un outil d'aide à la mise en page, en fait une DTD qui doit suggérer à l'étudiant non seulement toutes les rubriques et parties nécessaires du texte mais aussi en donner la forme (souvent normalisée dans nombre d'universités), la structure voire le rôle rhétorique. Cet environnement

²⁴ Par exemple en histoire de l'art des schémas de ligne de force d'un tableau, des échantillons graphiques de manière de peindre d'une école picturale, d'un atelier, d'un peintre, etc.

existe, il peut se déduire de Pélleas (Université Marne-la-Vallée), Cyberthèse (Lyon 2 et presses universitaires de Montréal) et du dispositif national relatif au dépôt national des thèses en collaboration avec l'Université Laval au Québec.

- Le suivi d'avancement : C'est une fonction utile à l'étudiant mais surtout aux professeurs qui suivront le travail. On peut en schématiser 2 facettes : a) Le suivi automatique des phases d'avancement et de contrôle d'un exercice (même si cet exercice est exceptionnellement particulier et même si son contrôle est complexe, étendu dans le temps et collaboratif entre plusieurs professeurs et université²⁵). Ce contrôle doit intervenir à de nombreux stades de la rédaction ; il implique la phase terminale le contrôle collégial s'intensifie avec la lecture des 2 rapporteurs préférentiellement un enseignant (le directeur de thèse) mais aussi très souvent d'autres enseignants, voire des experts qui peuvent intervenir sectoriellement, et bien sûr celle des autres membres du jury ; b) La mise en évidence explicite des versions successives (qui doivent être fonctionnellement accessibles pour rendre possible des lectures complémentaires, ainsi que la gestion elle aussi explicite des remarques, suggestions, indications de développement). Il est à remarquer que cette fonctionnalité est indispensable pour les enseignants parce qu'elle facilite et rend beaucoup plus efficaces leurs interventions mais cette fonction est évidemment très utile au doctorant qui n'est pas seulement un rédacteur mais son relecteur le plus assidu (tout au moins il devrait l'être) ;
- L'annotation de suivi. C'est une forme d'écriture collaborative qui doit rendre possible l'annotation ainsi que des suggestions d'ajouts, de suppressions, de modification, annotations ;
- Un niveau de travail collaboratif et de gestion conceptuelle du campus pour ce qui est des travaux et des acteurs est évidemment requis.

Établissons un scénario de balisage

Nous développons quelques pistes à partir de l'édition française de « Être et temps » de Martin Heidegger dans son édition française Gallimard.

Balisage formel

A partir d'une saisie OCR perfectionnée nous pouvons disposer de la mise en page originale. Celle-ci nous permet de générer assez facilement un fichier qui va nous permettre d'utiliser de façon interactive d'une part le corps des chapitres eux-mêmes qui auront été logiquement découpés en sous chapitres (notés § par Heidegger lui-même) et en paragraphes réels : au sens de la mise en page typographique ; nous disposerons aussi de la double pagination : celle de l'édition Gallimard en question et celle de l'édition allemande de référence. Le titrage et sous-titrage nous poseront quelques problèmes parce que Heidegger s'est autorisé à baliser son texte avec des intertitres qui (comme leur nom l'indique) sont des éléments textuels ne s'organisant pas dans une hiérarchie simple avec la structure des chapitres et paragraphes numérotés.

Notons aussi que le livre est hiérarchisé en «grandes parties et sections qui regroupent les chapitres, mais dont le découpage éditorial fait que se retrouve publiée ici une 1^{ère} partie sans 2^{ème} partie ni bien sûr de conclusion : successivement : Introduction, 2 chapitres numérotés 1^{er} et 2^{ème}, puis 1^{ère} partie 1^{ère} section chapitres 1 à 6 vient ensuite

²⁵ Cette disposition, tout au moins en France est obligatoire : les membres du jury appartiennent nécessairement à plusieurs universités

la 2^{ème} section chapitres 1 à 6.

Dans le corps du texte, nous récupérerons aussi les notes : cet appareil de note est très complexe, et pour cause, parce qu'il rend compte de plusieurs niveaux :

- En bas de page, des notes propres à Heidegger (ou plutôt des apostilles) numérotées de «ont » à «n » page par page, des notes propres à l'édition française, numérotées de «1 » à «n » page par page,
- En fin de volume, des notes référantes à la traduction, puis des notes d'accompagnement de la lecture référantes à la pagination de l'édition de référence. L'éditeur nous conseille d'autre part de nous référer à l'index « zu Heideggers, Sein und Zeit » de Hildegardb Feick, qu'il faudrait évidemment activer pour aménager la ressource. Se surajoute de plus une annexe de traduction de concepts grecs présentés sous forme d'index avec des renvois à la pagination (celle de l'édition française), et une autre annexe de traductions de citations latines (renvoyant à la pagination de l'édition française).

Le livre comporte en outre : des notes d'édition diverses précisant la version de référence, une dédicace à Husserl, un avertissement de la 7^{ème} édition (1953), une table des matières détaillée, une préface de l'auteur, puis en fin une postface de F.W. von Herrmann et les annexes.

Au-delà de ces grandes catégories, la structure locale du texte nous apportera encore une importante moisson de points remarquables qu'il faudra baliser : les guillemets, les zones mises en italique, les énumérations ou listes, les parties de phrases écrites en latin ou en grec (encore plus facile à repérer), les appels de notes ou références à d'autres chapitres ou § autant de structures qui se comporteront comme des points remarquables pour appréhender l'œuvre.

On peut noter que réussir un balisage formel d'une telle œuvre, de ses variantes d'édition, d'annotations, de traductions de l'allemand en français ainsi que du grec ou du latin, le lier intelligemment à des appareils d'index, de notes de gloses... est déjà en soit un travail extrêmement utile, car il permet de mettre en évidence la structure explicite non seulement logique mais conceptuelle voulue par Heidegger ou repérée par ses traducteurs éditeurs ou spécialistes critiques.

Balisage sémantique

Grâce à ce travail nous repérerons, et même rassemblerons un matériel conceptuel et sémantique déjà très abondant, que nous pourrons alors redéployer pour qualifier sémantiquement des parties entières et des citations précises renvoyant au vocabulaire lui-même ou à ses équivalents implicites ou explicites. Ce travail devra évidemment être réalisé en utilisant les index classiques comme « zu Heideggers, Sein und Zeit » de Hildegardb Feick, mais aussi d'autres index ou dictionnaires plus généraux ou au contraire spécialisés.

Cependant le style de Heidegger, et l'allemand qui est sa langue maternelle, nous éloigne du travail d'analyse linguistique francophone que nous nous sommes promis de mener dans le cas particulier du projet des philosophes de l'école française liés à l'Université de Paris 8 Vincennes.

Si on examine par exemple « De la grammatologie » de Derrida, ce texte comporte seulement des notes, sans index, ni bibliographie autre que celles que nous pourrons reconstituer à partir des notes. Le premier travail de balisage formel sera assez rapide, mais par contre nous disposerons de moins d'indications pour baliser sémantiquement

le texte.

Nous pourrions partir plus directement sur du repérage de vocabulaire, de concepts, de locution équivalentes à des notions, des auteurs, des écoles de pensée, explicitement ou implicitement cités. Ce qui est implicitement cité risque fort de nous poser quelques difficultés parce qu'une « certaine élégance stylistique » des auteurs français de sciences humaines dans les années 1970 se caractérise quelquefois par une volonté de s'exprimer de façon allusive. Au fil du travail d'aménagement cumulatif entrepris par les étudiants, il va falloir détecter les tournures stylistiques, le vocabulaire contextuel, les mots outils... qui caractérisent l'apparition d'un concept classique, d'une nouvelle notion proposée, la référence à une école de pensée, à un auteur, à une méthode de raisonnement. Ces environnements significatifs nous permettront de moduler et de relativiser les recherches par partie de mot, mot entier et synonyme que nous pouvons mettre en œuvre.

De toute façon, il est important de bien noter que, dans un environnement même dépourvu de toute intelligence de compréhension automatique du langage, on peut déjà faire beaucoup et vite à condition de disposer d'outils conviviaux de balisage et de pointage référents à des index permettant d'accrocher des définitions, des citations, des références, des alignements multilingues (ou de relation). L'outil de base est avant tout le surligneur électronique²⁶ dont on garde la mémoire et la trace et dont on sait conserver, retrouver et comparer les diverses occurrences. Toute bonne lecture d'un texte (et elles sont indispensables aux étudiants) se caractérise par un repérage des structures et des détails sémantiques de l'œuvre. En cumulant plusieurs « lectures savantes des étudiants », en exigeant éventuellement d'eux qu'ils s'accordent sur un consensus de balisage (activer électroniquement un *disputatio* moderne) puis en faisant contrôler par l'enseignant ce travail d'ensemble, on peut assez rapidement disposer d'un volume conséquent de ressources organisées en bibliothèque virtuelle annotées.

A quoi pourrait ressembler, en terme d'usage, un tel dispositif

Dans un symposium comme celui-ci, qui cherche à établir un dialogue entre des universitaires et des spécialistes des systèmes d'information de l'EAD engagés dans une démarche normative, il est important de pouvoir esquisser le type de dispositifs et les fonctionnalités concrètes dont les étudiants ou les professeurs pourraient disposer. Il est aussi indispensable de bien comprendre qu'au fur et à mesure de la mise en place d'un tel réseau, on devra passer par une phase de maquettage tant pour le dispositif technique, que pour les automatismes linguistiques ou de recopie des métadonnées, que pour bien sûr l'aménagement des corpus et des dictionnaires électroniques associés.

Le dispositif technique sera évidemment un terminal d'ordinateur le plus banalisé possible relié au réseau (aujourd'hui Internet). La logique d'une plate-forme d'enseignement à distance y sera intégrée, ce qui permettra de la façon que nous devons souhaiter la plus normalisée possible de disposer de fonctions de campus, de travail collaboratif et collégial, permettant de gérer maintes fonctionnalités d'interaction entre enseignants et étudiants, de travail associés entre des groupes d'étudiants, de contrôle de tâches, de prise en compte de commentaires, de facilité permettant d'observer l'évolution perfectible d'un aménagement du corpus ou de la

²⁶ On ne peut éviter de citer ici l'outil Annotea, développé dans le cadre du W3C, et qui illustre l'utilisation de RDF à des fins d'annotation.

rédaction d'un mémoire ou d'une thèse.

Au-dessus de cette plate-forme, nous devons disposer de la logique TEI.

Il s'agit en fait de pouvoir baliser des documents que nous aurons préalablement formatés en HTML puis en XML, selon des recommandations qui seront celles de la TEI pour laquelle nous aurons de façon cumulative associé pour chaque discipline (et dans chaque pôle thématique), un ensemble de balises adaptées aux spécificités du projet et des catégories de recherches qui y sont entreprises²⁷.

Par contre, les balises transversales à toutes les catégories de documents, voire celles qui concernent la description référentielle des documents seront le plus normalisé possible, et fondues dans les métadonnées normalisées, celles du Dublin Core, du LOM et du futur LRM. On utilisera aussi au maximum les logiques de balisage du traitement normalisé des ressources linguistiques ISO-TC37.

L'ensemble des ressources de balisage devrait être le plus transparent possible pour l'utilisateur, de même que les phases de balisage et les reprises et contrôles successifs d'un document par une ou plusieurs personnes. Des menus ou sous-menus disponibles à partir de la barre de menu générale doivent y donner accès. Chaque fois qu'il peut être possible de baliser automatiquement en toute sécurité cela doit être fait (par exemple : qui balise un document, à quelle heure et quelles zones ont été remplies ou modifiées).

Les automatismes.

Chaque fois qu'un mécanisme d'aide au balisage peut être proposé, il doit être effectivement implémenté. A ce niveau, la question est plus délicate car il doit toujours être possible de contrôler, d'améliorer ou de déjuger ce que l'automate propose. Ces dispositifs sont très utiles pour baliser très exhaustivement de grands corpus de document.

Les automatismes sont un point crucial qu'il est indispensable de spécifier et de développer soigneusement. Ils concernent :

- a) les recopies de partie de métadonnées dans la génération des références numériques. Chaque fois qu'un chercheur intervient dans le corpus structuré collectif, il doit disposer automatiquement de ces facilités et ne remplir lui-même que les zones et facettes que l'automate n'aurait pas pu remplir seul.
- b) les automates linguistiques de lemmatisation, d'équivalence synonymique, de parenté, de détermination de contexte, de recherche par le contexte.

Ces automates sont en pleine expansion mais ils ne peuvent être utilisables que si les promoteurs d'un projet exigent impérativement qu'ils répondent impérativement aux règles de normalisation et d'interopérabilité linguistique tels qu'ils se définissent dans l'ISO TC37.

En guise de conclusion

Impliqués depuis de longues années dans une démarche synergique de développement linguistique et informatique et d'autre part d'usage, les auteurs de cette communication collaborent depuis 5 ou 6 ans à des niveaux très complémentaires. L'un est linguiste informaticien, formé dans une école d'ingénieur l'autre est documentaliste spécialisé initialement en audiovisuel, historien de formation. Tous deux sont impliqués à des niveaux divers dans l'activité de normalisation.

²⁷ Notons que des DTD spécifiques de la TEI existent déjà et qu'elles seront bien sûr adoptées chaque fois que cela sera possible et pertinent : *TEI-verse* ou *TEI-drama* par exemple.

Depuis presque 10 ans, nous collaborons en synergie pour disséminer dans des disciplines de sciences sociales (littérature, sociologie, muséographie et maintenant philosophie) les principes de la TEI qui préfigurent de notre point de vue une couche fonctionnelle indispensable pour réaliser un Web sémantique réellement efficient dans tous les domaines de la communication.

L'opportunité de la normalisation des systèmes d'EAO ainsi que la volonté qui se fait jour dans de nombreux États, universités ou fondations pour promouvoir des grands projets de bibliothèques virtuelles nous confortent à persévérer dans le développement de notre projet.