Les éditions électroniques de l'École nationale des Chartes : objectifs, principes, outils et perspectives

par Florence Clavaud
Conservateur du patrimoine
Déléguée à la formation informatique à l'ENC*

Le présent article présente sommairement, en dégageant quelques lignes de force et perspectives, le travail réalisé par l'École nationale des chartes (ENC) en matière d'édition électronique depuis six ans. L'ENC utilise cette expression pour désigner en fait une catégorie précise d'éditions électroniques : celles qui comportent en leur cœur une édition critique de sources.

Quelques mots, tout d'abord, pour rappeler le contexte dans lequel ces éditions électroniques sont produites.

L'ENC, créée en 1821, est aujourd'hui un grand établissement public à caractère scientifique, culturel et professionnel. Elle a deux missions essentielles.

D'une part, elle forme à la recherche en sciences humaines et sociales, en particulier en histoire, des étudiants titulaires d'un diplôme universitaire de niveau licence ou plus ; elle forme également sur le plan scientifique les futurs cadres de la conservation et de la valorisation du patrimoine culturel, en particulier les futurs conservateurs du patrimoine et des bibliothèques. Les enseignements concernent les disciplines auxiliaires de l'histoire, une formation globale riche en histoire étant acquise avant l'entrée à l'ENC, et contrôlée, pour les archivistes-paléographes, au moment du concours d'entrée. Les matières enseignées incluent la paléographie, la philologie, la diplomatique, la codicologie, l'archivistique, la bibliographie et l'histoire du livre et des médias, l'histoire du droit, l'histoire de l'art, l'archéologie, en couvrant toutes les périodes historiques du

* Adresse de courriel : florence.clavaud@enc.sorbonne.fr
Ce texte a été relu par MM. Olivier Guyotjeannin, professeur d'archivistique et de diplomatique médiévale à l'École nationale des chartes (ENC), directeur de l'équipe de recherche de l'établissement, et Olivier Canteaut, directeur de la recherche et de l'informatique à l'ENC. Nous les remercions vivement pour cette collaboration.

Moyen Âge à nos jours. L'ENC délivre deux diplômes. Le premier, le plus ancien et le plus connu, propre à l'établissement, s'obtient après trois ans d'études et la soutenance d'une thèse de l'ENC ; il s'agit du diplôme d'archiviste-paléographe, qui ouvre la voie aux carrières évoquées plus haut. Le deuxième, qui est une création beaucoup plus récente puisque l'organisation du cycle d'études correspondant date des années 2006-2008, est un master universitaire, qui sanctionne donc deux ans d'études, et dont la dénomination « Nouvelles technologies appliquées à l'histoire » renseigne bien sur son contenu hybride et sur sa spécificité ; nous y reviendrons plus loin.

D'autre part, l'ENC mène des recherches dans les disciplines historiques et littéraires. Une équipe de recherche de petite taille et pluridisciplinaire, composée notamment des enseignants-chercheurs de l'établissement[1], travaille sur plusieurs axes dans lesquels son expertise et son action sont reconnues. Cette équipe place au cœur de ses préoccupations le document primaire, et en particulier, mais pas seulement, le document textuel. Non seulement le contenu et la forme matérielle des corpus documentaires, leur provenance et leur tradition, leur évolution, leur langue, y sont étudiés, mais ils y font souvent l'objet d'éditions critiques. L'ENC a choisi de diffuser elle-même les résultats de ses recherches, largement et de manière multimodale, que ce soit par la publication d'ouvrages imprimés (l'établissement publie quatre collections sous cette forme), par la participation à des colloques et journées d'études, ou, pour en arriver au sujet qui nous préoccupe ici, par la mise en ligne de ressources diverses.

En effet, c'est pour soutenir cette double mission d'enseignement et de recherche que l'ENC a choisi il y a quelques années de faire du développement des technologies numériques appliquées aux sources historiques une de ses priorités. Après quelques initiatives fondatrices, cette orientation stratégique a été clairement exprimée en 2005 comme un axe politique, lors de la conception du programme quadriennal d'action pour les années 2006-2009. On y lit clairement une volonté forte de poursuivre ainsi trois objectifs :

– diffuser en ligne, librement et gratuitement, des ressources et outils de référence ;
– publier sous la forme électronique, également en libre accès, certains des travaux de l'équipe de recherche et de ses partenaires ;
– plus généralement, contribuer autant que possible à renouveler par les technologies numériques, non seulement les moyens, mais aussi les méthodes de la recherche en histoire et dans les disciplines connexes.

[1] L'équipe d'accueil 3624 « Histoire, mémoire et patrimoine », dirigée par M. Olivier Guyotjeannin.

Le programme d'action quadriennal récemment défini pour les années 2010-2013 s'inscrit dans la droite ligne du programme précédent en ce qui concerne cette stratégie. Cependant, alors que le texte précédent isolait à côté d'activités essentiellement scientifiques certaines activités à forte dimension informatique (en tant qu'axes ou projets), le programme élaboré pour les quatre années à venir introduit les technologies numériques au cœur des axes de recherche scientifique. Cette différence est significative de l'évolution en cours, du passage actuel d'une phase exploratoire à une phase «industrielle» de mise en œuvre de ces technologies numériques.

Nous allons maintenant évoquer plus précisément, au travers d'une analyse à la fois historique et prospective, les activités de l'établissement dans ce domaine.

Il faut pour commencer souligner le fait que si les activités d'informatique appliquée aux documents primaires ont émergé à l'ENC il y a environ six ans, c'est parce que le contexte était favorable à cette émergence.

En premier lieu, le contexte technologique lui-même était réellement propice. Le champ des possibilités offertes par les technologies numériques était enfin très ouvert. Les techniques de numérisation avaient atteint un niveau de qualité et de maturité permettant de traiter tout type de document pour obtenir de vrais substituts numériques; la vitesse de transmission des données sur Internet était suffisamment importante, et les langages de développement d'applications Web suffisamment stables et puissants, pour permettre des réalisations efficaces et accessibles au sens propre du terme. Des standards internationaux existaient depuis quelques années et avaient fait leurs preuves dans différents secteurs, au premier plan desquels il faut placer le métalangage XML[2] et les grammaires XML applicables au texte et à ses métadonnées; ainsi un socle de normes pouvait servir de référentiel pour arpenter des territoires neufs. Beaucoup de choses pouvaient donc être faites; cependant dans le domaine de la mise en ligne de sources et travaux de chercheurs, très peu avait été tenté.

De plus il se trouvait à ce moment-là dans l'établissement quelques personnes, des chercheurs, pour avoir envie de tenter l'aventure, et quelques autres personnes, connaissant les technologies clé, pour avoir la volonté et les moyens d'avancer dans ce domaine. On ne dira jamais assez combien, pour de telles activités et à de tels moments, la présence et l'action de quelques-uns peuvent être déterminantes.

---

[2] *EXtensible Markup Language*: un métalangage simple et flexible pour le balisage, donc la structuration du texte, dérivé de SGML. Les spécifications de la version 1.0 ont été publiées par le W3C en 1998. XML a donné naissance à un grand nombre de normes et de langages, comme XSLT; son utilisation s'est répandue dans tous les domaines informatiques, notamment ceux où l'échange et la transformation de données ou documents sont essentiels; voir <http://www.w3.org/XML/>.

Les premières réalisations portent donc la marque de leurs initiateurs et auteurs. Elles concernent principalement les sources médiévales. Elles ont été menées à bien rapidement, en ne ménageant pas les efforts, et en combinant ambition et rigueur.

Elles ont en outre été conceptualisées. Dès le début une idée force a consisté à apporter avec l'édition électronique autre chose que l'édition imprimée, d'où notamment la mise en place de moyens de recherche textuelle dans les ressources mises en ligne. La volonté de ne renoncer en rien, même si on publie des ressources sur Internet et non pas sous la forme d'un livre imprimé, au niveau élevé d'exigence scientifique atteint par l'établissement, a trouvé sa traduction informatique dans le choix du respect de standards tels que TEI[3] – il est en effet plus difficile et plus long de s'approprier un modèle de document structuré générique, conçu par d'autres, et les technologies associées, que de définir ses propres règles ou de produire un site Web en quelques jours ; mais le résultat est exploitable, réutilisable et peut passer les années.

Toutes ces caractéristiques ont fait des premières réalisations des travaux pilotes, c'est-à-dire à la fois relativement isolés, exploratoires et fondateurs. Décrivons rapidement les principales réalisations de cette période, telles qu'on les trouve encore aujourd'hui sur Internet :

– un site Web à vocation pédagogique, THELEME[4], qui propose des ressources destinées essentiellement à un public d'étudiants de l'enseignement secondaire ou de niveau licence, ou à leurs enseignants. Outre des cours complets et des bibliographies mis à jour par les enseignants-chercheurs, ce site comprend en son cœur une application documentaire hybride, les dossiers THELEME[5], au sein de laquelle on peut consulter, autour de l'image numérique d'un document primaire ou d'un extrait de ce document (charte médiévale, document d'archives de l'époque moderne, ou page d'un manuscrit littéraire ancien), l'édition de ce document, associée le cas échéant à sa traduction en français, à des commentaires paléographiques, diplomatiques, linguistiques

[3] *Text Encoding Initiative*, au départ (1987) un projet de recherche en *digital humanities* porté par trois associations professionnelles anglo-saxonnes de chercheurs en sciences humaines et sociales, qui a produit un modèle XML pour l'édition structurée de tout type de texte ; voir <http://www.tei-c.org/>.

[4] THELEME (Techniques pour l'Historien en Ligne : Études, Manuels, Exercices) : voir <http://theleme.enc.sorbonne.fr/>.

[5] <http://theleme.enc.sorbonne.fr/dossiers/>. Cette application utilise à la fois une base de données relationnelles pour stocker les métadonnées des documents publiés (ce qui permet de faire des recherches sur ces métadonnées, voir <http://theleme.enc.sorbonne.fr/dossiers/formulaire.php> et des fichiers XML/TEI pour le stockage sous une forme structurée des éditions et commentaires scientifiques des textes (un document TEI par dossier, dont chaque section est transformée par des programmes XSLT en HTML pour montrer le fac-similé interactif, l'édition du texte et sa traduction, les commentaires). Voir par exemple les dossiers n^{os} 80 (acte seigneurial de 1177, <http://theleme.enc.sorbonne.fr/dossiers/notice80.php>) et 100 (extrait du Didascalicon de Hugues de Saint-Victor, <http://theleme.enc.sorbonne.fr/dossiers/notice100.php>).

et historiques. La fonctionnalité d'alignement de zone du document numérisé avec l'édition du texte écrit sur cette zone, mise en place pour obtenir le « fac-similé interactif » a beaucoup attiré l'attention sur cette application, qui continue de s'enrichir. Les étudiants contribuent à ce projet, puisque certains des dossiers documentaires qui sont présentés ont été préparés par eux, sous la direction d'un enseignant.

– une véritable collection d'éditions électroniques destinées à un public de chercheurs, ELEC (Éditions en Ligne de l'École des chartes)[6], au sein de laquelle on trouve aujourd'hui quatre publications qui consistent en bases de données de référence, et onze autres qui consistent en éditions critiques diplomatiques de corpus complets de textes, principalement de l'époque médiévale. Les contenus de ce dernier sous-ensemble ne sont pas totalement homogènes sur le plan technique ; mais leur format est principalement XML/TEI et les applications correspondantes ont été construites entre 2006 et 2008 sur une plate-forme informatique unique, elle-même développée principalement par l'ENC à partir de logiciels libres, baptisée TELMA[7]. TELMA est une réalisation conjointe de l'ENC et de l'IRHT[8], développée dans le cadre du Centre de ressources numériques (CRN) du même nom, TELMA[9], créé en 2006 par le CNRS en même temps que d'autres Centres de ressources numériques pour servir de centres d'expertise, de pilotes et d'opérateurs dans le domaine de la mise en ligne et de l'exploitation informatique des sources et travaux de recherche. La plate-forme a donc accueilli entre 2006 et 2008 des applications développées par le CRN pour publier des corpus préparés par d'autres entités que l'ENC et l'IRHT. Le CRN TELMA a aussi joué d'autres rôles, notamment celui de conseil auprès d'équipes de recherche démarrant un projet, ou de contributeur à des formations.

Dans le même temps le contexte évoluait, de sorte qu'aujourd'hui la donne a changé par rapport à celle qui a conduit à ces premiers travaux exploratoires. Parmi les indices de ce changement, nous en relèverons deux.

En premier lieu, en partie grâce aux bons résultats obtenus par ces travaux et par ceux d'autres opérateurs, de nombreux chercheurs français en sciences humaines et sociales expriment aujourd'hui des besoins

[6] <http://elec.enc.sorbonne.fr/>.

[7] TELMA (Traitement Électronique des Manuscrits et des Archives). On peut retrouver la liste complète des applications publiées à l'aide de TELMA, et utiliser ses fonctionnalités de recherche et de consultation, en allant à la page : <http://www.cn-telma.fr/corpus/>. On peut aussi retrouver ces fonctionnalités en accédant directement à chacune des applications d'édition électronique qui l'utilisent, par ex. pour l'ENC, les comptes consulaires de Montferrand, ici <http://elec.enc.sorbonne.fr/montferrand/>.

[8] IRHT : Institut de Recherche et d'Histoire des Textes (UPR 841 du CNRS), voir <http://www.irht.cnrs.fr/> (site Web consulté le 3 avril 2010).

[9] Le site Internet du CRN TELMA est au 3 avril 2010 à l'adresse : <http://www.cn-telma.fr/>.

en matière d'informatisation de leurs recherches, qu'ils n'exprimaient pas il y a trois ans. Les sollicitations directes en matière de conseil ponctuel ou plus continu, de formation sommaire ou avancée, de développement d'applications et d'outils, sont très nombreuses pour l'ENC, qu'elles proviennent de l'extérieur de l'établissement ou de son équipe de chercheurs. Ce constat est fait aujourd'hui par tous les acteurs du domaine en France. Un bon indice, parmi d'autres, de cette montée en charge, est le succès obtenu en octobre 2008 par l'école thématique organisée par le CNRS à Fréjus, sur le thème « *Préservation et diffusion numérique des sources de la recherche en sciences humaines et sociales* »[10].

En second lieu, le nombre d'équipes capables d'agir dans ces domaines en France a augmenté. D'une part, les savoir-faire technologiques se sont diffusés et le modèle de référence TEI est un peu mieux connu. D'autre part, même si la publication électronique ne bénéficie pas encore du même statut et n'apporte pas le même niveau de reconnaissance au chercheur que la publication papier, les chercheurs et leurs organismes de rattachement ont commencé à prendre la mesure des avantages de ce canal, plus souple puisqu'on peut publier sans attendre d'avoir terminé et mettre à jour la publication facilement, et souvent plus efficace puisqu'on peut atteindre plus facilement un cercle plus large. L'édition électronique des revues scientifiques s'est également organisée, ce qui a pu orienter vers des solutions en ligne pour l'édition électronique de sources. Diverses entités ont donc financé et organisé des équipes techniques dédiées aux activités de publication électronique de sources historiques[11]. Les modalités d'action de ces équipes sont variées, leurs moyens restent limités, leur vie au quotidien est souvent difficile ; cependant tous leurs travaux vont dans le même sens et il devient utile et possible, même si ce n'est pas toujours facile, de travailler ensemble hors des cadres institutionnels. Finalement, une infrastructure réseau, portée par le TGE ADONIS[12], est en cours d'organisation, pour coordonner et aider ces activités, pour fédérer les compétences et offrir des solutions mutualisées lorsque cela est

[10] Voir le wiki réalisé pour réunir les informations, contributions, commentaires et bilans de l'école : <http://www.digitalhumanities.cnrs.fr/wikis/ecole-sources-num/index.php?title=Accueil> (site consulté le 3 avril 2010). Une école de formation nationale (ANGD) du CNRS sera organisée en octobre 2010 pour poursuivre dans la même voie.

[11] Citons par exemple : MUTEC, un dispositif créé en 2007 à Lyon, par le Service d'Ingénierie Documentaire de l'ISH et l'Unité Numérique de l'ENS LSH en partenariat avec d'autres laboratoires de recherche, pour accompagner les chercheurs de la région Rhône-Alpes dans leurs projets d'éditions critiques et de constitution de corpus numériques, et leur permettre de partager leurs méthodes, outils et réflexions sur ce type de projet (<http://www.mutec-shs.fr/) ; les activités d'édition électronique et d'informatique appliquée au document du Centre d'Études Supérieures de la Renaissance (une UFR de l'université François-Rabelais de Tours, UMR du CNRS : <http://cesr.univ-tours.fr/>).

[12] ADONIS (Accès unifié aux Données et Documents Numériques des Sciences humaines et sociales) est un très grand équipement du CNRS créé en 2007 ; voir <http://www.tge-adonis.fr/>.

applicable[13], avec comme principal objectif pour l'année 2010 d'ouvrir un portail permettant d'avoir un accès unifié aux sources et travaux des chercheurs.

Comme un miroir de ce nouvel état des choses, on note que le concept d'humanités numériques – traduction du concept anglo-saxon de *digital humanities* –, qui sert à dénommer cette activité au carrefour des sciences humaines et sociales et de l'informatique, est sorti des cercles étroits de convaincus, le plus souvent à l'initiative de ces mêmes cercles, pour s'exposer[14] et être discuté, à défaut d'être, pour l'instant, véritablement utilisé par les décideurs.

Cette évolution a un impact important sur les travaux d'édition électronique de l'ENC.

Alors que ce qui a déjà été réalisé et mis en ligne doit nécessairement être maintenu en état de marche et ses contenus mis à jour, il faut aussi faire face à des besoins croissants, dans un environnement où les interlocuteurs se multiplient. Il faut également souligner que les acquis de la première phase ne suffisent pas pour traiter sans effort les nouvelles questions qui se posent. En effet, les contenus des futures éditions électroniques sont de nature plus variée, cette hétérogénéité documentaire concerne parfois un seul et même corpus[15]; de ce fait les modèles de structuration des textes précédemment élaborés et les solutions techniques mises en œuvre jusqu'ici ne sont pas forcément applicables ou suffisantes. Les attentes fonctionnelles des chercheurs impliqués sont plus pointues. Rien que de naturel et de très stimulant dans cette évolution; il est néanmoins clair que les projets informatiques sont de plus en plus complexes. Un autre facteur de complexité, inhérent à ces travaux, est un niveau assez élevé d'incertitude: il s'avère assez souvent que les corpus, donc le périmètre des projets informatiques, sont eux-mêmes à «inventer» sinon mouvants, que les besoins fonctionnels ne peuvent être définis entièrement au début des projets, mais sont formulés au fur et à mesure que les chercheurs cernent les problématiques, définissent leur méthodologie, formulent des hypothèses à vérifier.

Une telle situation a conduit à réaffirmer les principes fondamentaux définis pendant la première phase, car ceux-ci demeurent d'autant plus valables pour cette deuxième phase qu'ils sont la condition nécessaire

[13] C'est notamment le cas de l'offre d'archivage pérenne qu'ADONIS met actuellement en place sur le modèle international de l'Open Archival Information System (OAIS) en liaison étroite avec le Service Interministériel des Archives de France du ministère de la Culture, et en collaboration avec le Centre informatique National de l'Enseignement Supérieur (CINES).

[14] Voir par exemple le portail des *digital humanities* au CNRS, <http://www.digitalhumanities.cnrs.fr/> (consulté le 1er avril 2010).

[15] Ainsi, pour un seul et même projet en cours, le corpus à éditer, à mettre en ligne et à outiller inclut un inventaire d'archives du XVIIe siècle et un cartulaire médiéval. Pour un autre projet, un nombre important de lois, décrets, circulaires, mémoires et rapports du XIXe siècle devront être numérisés et encodés.

pour que les investissements faits produisent des résultats intéressants, de qualité et relativement durables :

- publier en ligne des documents primaires inédits ou des documents dont il n'existe qu'une édition ancienne, en mode texte de manière à permettre leur exploitation par un moteur de recherche ;
- pour que les corpus numériques produits soient réutilisables et puissent être pérennisés, utiliser des standards partagés par des communautés larges et des formats publics (principalement XML et la grammaire TEI, dans sa dernière version majeure, la TEI P5) ;
- veiller à ce que les ressources ainsi publiées soient faciles d'accès et retrouvables (notamment en assurant l'univocité, la précision et la pérennité des URLs) ;
- offrir des services différents de ceux qu'offre une édition imprimée (par exemple en proposant des moyens de navigation hypertextuelle avancés, sans parler des moyens de recherche par critères ou dans le texte ; ou encore, en mettant en ligne la totalité du corpus de documents textuels primaires à la fois sous la forme d'une édition critique et sous la forme d'images numériques, afin de « valider » l'édition et de donner un accès direct à l'original).

Pour relever les nouveaux défis dans le respect de ces principes, une équipe un peu plus nombreuse que la première équipe a été constituée en 2007-2009. Comme dans le cas de la première équipe, les personnes recrutées ont toutes un profil mixte : elles ont toutes reçu une formation de niveau universitaire en sciences humaines et sociales et une formation en ingénierie système ou documentaire, ou à tout le moins elles sont très intéressées par la recherche en sciences humaines et sociales. Il est en effet nécessaire que cette équipe comprenne les besoins des chercheurs pour lesquels elle travaille, bien plus, qu'elle soit en mesure de les aider à les formuler, ou encore d'orienter judicieusement la recherche de solutions informatiques et d'évaluer correctement l'intérêt de certains outils. En ce qui concerne les compétences informatiques de cette équipe, grâce à des efforts importants d'apprentissage et de formation, tous partagent un socle commun de savoir-faire fondamentaux, et certains sont en outre spécialisés dans certains domaines (tels que l'ingénierie système, le « Web design », la lexicographie ou le traitement automatique des langues). Il est ainsi possible de constituer, en fonction des projets, de petites équipes très soudées. Certains des acteurs de cette équipe ont en outre une expérience de la gestion de projet informatique, qui s'avère très précieuse et permet de privilégier les approches flexibles et collaboratives et la réactivité.

Cette équipe a besoin, pour répondre correctement aux défis actuels, de renforcer son efficacité, de capitaliser ses connaissances et de travailler avec d'autres équipes. C'est pourquoi, aujourd'hui plus qu'auparavant, elle formalise autant que possible ses relations avec les équipes de chercheurs pour lesquelles elle travaille. Cela peut se traduire, si le projet

concerne des entités extérieures à l'ENC, par une convention précisant les droits et responsabilités de chaque partie en présence. La planification et la coordination des taches sont par ailleurs devenues une activité cruciale ; il faut parvenir à évaluer correctement les travaux à accomplir, veiller à ne pas multiplier, pendant la même période et pour la même personne, les projets à suivre, à rapprocher les activités comparables et à mutualiser ce qui peut l'être. En outre, depuis environ un an, l'équipe utilise des outils de développement collaboratifs, en particulier un système de gestion de version de fichiers, ce qui permet à chacun de développer sa part puis de charger ses fichiers dans un entrepôt commun. Elle s'efforce aussi de documenter ses travaux. Enfin, elle a commencé à nouer des contacts fructueux avec d'autres équipes françaises ou étrangères travaillant dans le même domaine[16], pour réutiliser et partager des outils libres, échanger sur les expériences, méthodes et modèles documentaires, agir en commun lors de rencontres et séminaires, voire réfléchir à des réalisations communes permettant de généraliser des approches institutionnelles ou nationales.

Nous allons maintenant présenter rapidement quelques-unes des pistes de travail actuelles de l'ENC en matière d'éditions électroniques.

Un des objectifs à moyen ou long terme serait, en quelques mots, de faire entrer réellement les chercheurs de l'ENC dans le cercle des humanités numériques, pour qu'ils en soient des acteurs à part entière. Cet objectif n'est pas encore atteint, au sens où la plupart des chercheurs de l'ENC (ou plus largement des unités de recherche en SHS en France) ne peuvent pas produire eux-mêmes les fichiers XML qui seront publiés. Ce sont les « informaticiens » de l'ENC qui font ce travail, soit, dans les cas les plus difficiles et pour les corpus de petite taille, en encodant manuellement les éditions critiques et études historiques préalablement produites au traitement de texte, soit en les convertissant par programme en XML, soit en combinant les deux approches. Cela présente deux inconvénients : d'une part une relative perte de temps en particulier pour l'équipe informatique qui pourrait consacrer ce temps aux tâches de modélisation, de développement ou de formation, d'autre part un prisme d'interprétation du travail du chercheur, qui perd la maîtrise de son travail scientifique. Nous pouvons imaginer que, si les chercheurs étaient équipés d'outils adéquats, abordables pour les non-informaticiens et efficaces, pour produire eux-mêmes leurs éditions critiques et études en XML, cela pourrait au moins les accompagner dans cet exercice rigoureux, voire changer leur regard sur le corpus traité, ou faire surgir plus facilement de nouvelles idées sur l'utilisation des modèles XML existants ou sur

[16] Parmi les équipes avec lesquelles des partenariats fructueux existent ou sont en cours d'organisation, on peut citer l'équipe informatique du projet ARTFL (*Department of Romance Languages and Literatures, Division of the Humanities*, université de Chicago, États-Unis), et le *Center for Computing in the Humanities* (*King's College*, Londres, Royaume-Uni).

leur enrichissement. Actuellement, pour atteindre cet objectif, il manque surtout des outils à la fois suffisamment génériques pour être paramétrables en fonction de la nature de l'édition et de la discipline scientifique, et suffisamment puissants et précis pour «libérer» le travail et permettre qu'il soit collaboratif. Cependant divers travaux sont en cours, à l'ENC et ailleurs, qui nourrissent cette réflexion.

A plus court terme, l'ENC travaille depuis plus d'un an au développement d'une nouvelle plate-forme de recherche et de consultation de corpus numériques, destinée à remplacer, au moins pour les éditions électroniques qu'elle produit en TEI et publie, la plate-forme TELMA mentionnée plus haut. La première version de la plate-forme est actuellement en cours de finalisation. Tirant les leçons de la réalisation de la première plate-forme, le travail, fondé sur l'utilisation des mêmes standards techniques (XHTML, CSS, JavaScript, XML/TEI, XSLT), a abouti à une architecture modulaire, dans laquelle les contenus (fichiers XML/TEI, images numériques), le moteur d'indexation et de recherche textuelle, les moyens de consultation et de navigation, sont clairement séparés, de façon notamment à permettre à chacun d'entre eux d'évoluer séparément, ou encore de manière à pouvoir remplacer un des modules par un autre réalisé avec d'autres composants logiciels si c'est nécessaire pour un corpus donné. Le choix en matière de moteur d'indexation et de recherche s'est pour l'instant porté sur PhiloLogic™[17], un moteur de fouille de texte libre, écrit par l'université de Chicago avec laquelle l'ENC a noué des relations dès 2008. Un autre fondement de cette architecture est la constitution de jeux de scripts (XSLT, PHP) génériques, susceptibles de resservir pour chacun des corpus à publier dans la mesure où d'autres scripts propres au corpus peuvent les surcharger et les inclure.

A l'heure où ces lignes sont écrites, deux nouvelles éditions électroniques ont déjà été publiées sur cette plate-forme. On les trouve listées à la page d'accueil de la collection ELEC, dans la rubrique Nouveautés; il s'agit du *Glossarium mediae et infimae latinitatis* de Ch. du Cange (1610-1688) dans son édition par L. Favre en 1883-1887[18] et du *Sanctoral du lectionnaire de l'office dominicain (1254-1256)*[19]. De nouvelles éditions suivront.

Les éditions électroniques précédemment publiées sur TELMA sont progressivement migrées vers la nouvelle plate-forme. Il s'agit bien d'une véritable opération de migration pour ce qui est des contenus: les fichiers XML/TEI anciens, encodés en TEI P4, sont convertis en TEI P5, restruc-

[17] <http://philologic.uchicago.edu/> (site Internet consulté le 3 avril 2010).

[18] L'adresse directe au 3 avril 2010 est <http://ducange.enc.sorbonne.fr/>.

[19] Édition d'après le ms. Rome, Sainte-Sabine XIV L 1, par Anne-Élisabeth Urfels-Capot; l'adresse directe de l'application au 3 avril 2010 est <http://elec.enc.sorbonne.fr/sanctoral/>.

turés et enrichis, en particulier pour isoler et indexer les dates et les entités nommées.

Chaque item (un acte, une charte, un édit, une division intellectuelle de petite taille dans un texte littéraire) d'un nouveau corpus numérique se voit affecter une URL pérenne et signifiante, et si le corpus avait déjà été publié auparavant, l'URL initiale des items est conservée.

Enfin, pour aller jusqu'au bout de la logique de partage de ressources, l'ENC a décidé de rendre téléchargeables et donc réutilisables par d'autres, sous licence *Creative Commons*[20], les contenus (les fichiers XML/TEI, accompagnés de leur schéma documenté) de ses éditions électroniques, au fur et à mesure qu'elles seront mises en ligne.

Les conditions de l'interopérabilité de ces contenus avec d'autres applications (possibilité de référencer les contenus dans des portails documentaires, ou de les utiliser à l'aide de Web services) seront aussi progressivement mises en place (création d'un entrepôt OAI-PMH, probablement exposition des métadonnées et des référentiels en RDF, etc.).

Parallèlement, des recherches sont en cours pour repenser les éditions électroniques elles-mêmes. Sans parler d'améliorer l'ergonomie des applications précédemment mises en ligne, voici quelques aspects qui sont l'objet de beaucoup d'attention dans le cadre de certains projets :

– l'intégration aux éditions électroniques des images des documents primaires, lorsque l'édition concerne de tels documents. Ce point a déjà été évoqué plus haut. Peu d'images de documents primaires ont jusqu'ici été publiées dans ELEC ; les projets en cours leur accorderont une place plus importante, ce dès ce printemps pour certaines applications. Il conviendra d'aménager les interfaces en conséquence. Pour une de ces applications au moins[21], l'édition critique étant un travail de très longue haleine, la publication du document primaire inédit, dans son intégralité, par le biais de l'image numérique, rendra un premier service et facilitera les travaux d'édition.
– en plus des pages Web de consultation des éditions critiques proches du modèle canonique et familier de l'édition imprimée, des recherches sont menées pour proposer de nouvelles interfaces, plus éloignées de ce modèle canonique, tirant par exemple parti de JavaScript pour la présentation de l'apparat critique ou l'accès aux index. Ce modèle canonique, d'autre part, n'existe pas toujours pour certaines catégories de textes et il faut inventer des modèles de page. Un autre chantier

[20] Voir <http://fr.creativecommons.org/>. La licence choisie est la suivante : « Paternité – Pas d'Utilisation Commerciale – Pas de Modification ; 2.0 France », elle est disponible en ligne ici <http://creativecommons.org/licenses/by-nc-nd/2.0/fr/> (site consulté le 3 avril 2010).

[21] Il s'agit de l'édition électronique conjointe de l'Inventaire général manuscrit du chartrier de l'abbaye de Saint-Denis, rédigé entre 1680 et 1728, pour la période allant jusqu'à 1302 (Archives nationales de France, LL1189, LL1190 et LL1191) et du cartulaire blanc de Saint-Denis, produit entre 1270 et 1300 pour l'essentiel (*ibidem*, LL1157 et LL 1158).

intéressant concerne les moyens de navigation ou de représentation des corpus numériques. Dans d'autres secteurs professionnels, ou pour d'autres documents, diverses équipes travaillent aussi sur ces problématiques, cherchant des dispositifs de présentation des ressources plus graphiques tels que les lignes de temps[22]. La création de référentiels externes (fichiers d'autorité) ou de modèles de représentation des connaissances (ontologies) pourrait aussi conduire à développer des moyens de navigation liés étroitement aux composants du corpus, tels que des graphes.

Un autre axe de travail concerne la fusion des deux chaînes de production : celle qui permet de fabriquer des éditions électroniques et celle qui permet de préparer des éditions imprimées. Actuellement ces deux chaînes sont distinctes à l'ENC, comme elles le sont encore le plus souvent ailleurs[23]. Cela engendre un double travail et donc une perte de temps importante, ainsi que des risques de divergence, dans le cas où le même texte fait l'objet d'une édition électronique et d'une édition papier ; ajoutons que le caractère propriétaire, dépendant du logiciel utilisé, des fichiers produits pour l'édition papier, les rend difficiles à archiver pour le long terme. Une chaîne de production unique, basée sur XML, serait plus efficace et plus pérenne. Dans le même ordre d'idées, de nombreux travaux d'étudiants (notamment certaines thèses d'archivistes-paléographes) étant aujourd'hui saisis en utilisant LaTeX[24], la conversion de ces fichiers en XML sera à prévoir et effectuer dès lors qu'une édition électronique en sera programmée.

Au-delà de la création des conditions nécessaires à la publication d'éditions électroniques plus performantes et aux contenus ouverts et interopérables, l'ENC souhaite aussi autant que possible contribuer à outiller les chercheurs. Les modules de la plate-forme évoquée ci-dessus seront donc bientôt distribués sous licence libre. Une autre avancée importante sera la mise au point et la distribution de programmes de traitement automatique des textes (incluant des moyens d'enrichissement du balisage, ainsi que des outils de lemmatisation du latin médiéval, dont une

[22] Une des applications libres utiles pour développer de telles interfaces est le *Timeline Wigdet* du *Massachusets Institute of Technology* (<http://www.simile-widgets.org/timeline/>).

[23] En France, les Presses universitaires de Caen (<http://www.unicaen.fr/services/puc/>, site consulté le 3 avril 2010), qui publient à la fois sous la forme papier et la forme électronique des éditions critiques de textes, sont une des très rares entités qui ont réalisé cette fusion.

[24] Un système logiciel sous licence libre pour la composition de documents, souvent utilisé pour rédiger et préparer l'impression de publications scientifiques de contenu complexe, pour lesquels une mise en page normalisée est attendue. Voir <http://www.latex-project.org/> (site consulté le 3 avril 2010).

base lexicale et un corpus de textes lemmatisés de référence[25], des moyens de comparaison de textes...).

Enfin, un dernier axe de travail consiste à œuvrer pour que les jeunes chercheurs ou chercheurs de demain aient non seulement une connaissance sommaire des possibilités que leur offrent les technologies employées, mais aussi une maîtrise des fondamentaux de ces technologies, pour s'en servir autant que nécessaire. Il s'agit donc d'élargir le cercle des acteurs des humanités numériques pour préparer les éditions électroniques de demain.

Bien sûr, à l'ENC, cela concerne en premier lieu les étudiants de l'établissement. C'est ainsi que le master « Nouvelles technologies appliquées à l'histoire » donne la possibilité à de jeunes étudiants en histoire qui souhaiteraient, soit poursuivre leurs recherches en doctorat, soit agir comme membres d'équipes de projet d'informatisation et de valorisation de corpus, d'apprendre en master 2 les technologies incontournables pour ce faire, après une première année de master consacrée à l'apprentissage des sciences auxiliaires de l'histoire. Après un semestre d'enseignements comportant un nombre important d'heures de travaux dirigés, les étudiants mettent à profit et complètent leurs connaissances informatiques en effectuant un stage de longue durée, ou bien en travaillant sur un sujet personnel de recherche[26]. Plusieurs stages ont déjà été effectués auprès de laboratoires de recherche pour contribuer à des projets d'édition électronique, en y apportant une compétence à la fois scientifique et technique. Les résultats des stages sont, lorsque la structure d'accueil le peut et que le travail s'y prête, mis en ligne ; les travaux de recherche personnels sont, pour quelques-uns d'entre eux, en cours de publication sur Internet par l'ENC comme autant d'éditions électroniques de petits corpus, explorant diverses voies. L'insertion professionnelle des diplômés de master 2 est très réussie depuis que celui-ci existe.

D'autres actions sont menées en direction de chercheurs plus avancés extérieurs à l'établissement, qu'il s'agisse d'initiatives propres à l'ENC (offre de formation continue[27]) ou plus collectives (participation à des

[25] Travail en cours dans le cadre du projet de recherche OMNIA (Outils et Méthodes Numériques pour l'Interrogation et l'Analyse des textes médiolatins), financé par l'ANR, mené par l'ENC, l'UMR 5594 du CNRS – Archéologie, terre, histoire, sociétés (ARTeHIS, Dijon) – et l'équipe de Lexicographie latine de l'IRHT.

[26] Voir, sur le site Web de l'ENC, les pages consacrées à ce master (<http://www.enc.sorbonne.fr/master-nouvelles-technologies-appliquees-a-l-histoire.html>).

[27] L'offre de formation continue consiste, soit en stages à portée générale organisés régulièrement et faisant l'objet d'un catalogue, soit en formations à la carte. En ce qui concerne l'informatique appliquée au document, le catalogue est en cours de développement, avec une formation de cinq jours à TEI organisée en juin 2009, et plusieurs demandes de formation à la carte ont été, soit émises récemment, soit déjà concrétisées. Voir <http://www.enc.sorbonne.fr/formation-continue.html>.

séminaires ou écoles thématiques[28]). Il est prévu de développer ces actions à l'intérieur de l'établissement, en particulier pour les étudiants archivistes paléographes[29], comme à l'extérieur de l'établissement.

Nous terminerons simplement cet article en exprimant avec confiance le souhait que les lignes ci-dessus, en particulier celles qui décrivent l'état actuel des choses, puissent très vite devenir obsolètes – en d'autres termes, que rien ne vienne arrêter le beau mouvement qui a été initié, que les résultats soient à la hauteur des espérances (quels qu'ils soient ! car il est difficile de prévoir vraiment comment sera organisé et ce que fera ce secteur d'activité dans plus de cinq ans ; ce secteur a d'ailleurs besoin de moyens mais aussi de liberté), et qu'il se trouve à l'ENC et ailleurs de plus en plus de personnes pour œuvrer dans ce domaine.

[28] Parmi les actions auxquelles l'ENC a contribué, nous avons déjà cité l'école thématique du CNRS qui a eu lieu à Fréjus en octobre 2008. Une autre action collective est organisée au mois de juin 2010 à Lyon par MUTEC (<http://www.mutec-shs.fr/la-tei-en-france-pratiques-et-perspectives>).

[29] Pour l'instant, l'enseignement dispensé à ces étudiants ne comporte que peu de cours d'informatique. Quelques ateliers optionnels sur l'édition électronique structurée ont été organisés depuis 2007. L'un d'entre eux a permis à Florine Stankiewicz, dans le cadre de son travail de thèse d'archiviste paléographe soutenu en 2009, d'encoder et de publier sur Internet un répertoire des œuvres de Pierre Gringore (v. 1475-v. 1538), homme de lettres, de théâtre et de cour (<http://theses.enc.sorbonne.fr/Pierre-Gringore/gringore.html>).