À l'heure où, au sein de la communauté scientifique, s'affirme le mouvement pour les archives ouvertes, le libre accès aux thèses est devenu une préoccupation majeure des universités, écoles d'ingénieurs et autres établissements d'enseignement supérieur et de recherche. La diffusion des thèses électroniques s'inscrit dans ce souci de rendre visibles et accessibles des documents scientifiques validés mais non publiés. Après avoir rappelé les modalités techniques de dépôt, de traitement et de diffusion des thèses, cette étude évoque le projet français STAR de dépôt, signalement et archivage de ce type de document, et insiste sur les enjeux scientifiques d'une politique nationale de diffusion électronique des thèses.

par **DIANE LE HÉNAFF-STITELET** et **CATHERINE THIOLON**

Gérer et diffuser des thèses électroniques : **un choix politique pour un enjeu scientifique**

PEU DE THESES SOUTENUES SONT actuellement publiées en France. Or la thèse est un document scientifique présentant les résultats d'une recherche dont l'ensemble de la communauté scientifique doit pouvoir bénéficier. Il faut donc en faciliter la diffusion. Retrouver une thèse suppose en amont un dépôt systématique et un signalement national. La sensibilisation du doctorant au dépôt de sa thèse, pour permettre sa visibilité et sa valorisation, mobilise les acteurs institutionnels engagés dans la gestion de ce type de document.

Un examen des différentes étapes de la gestion des thèses électroniques, en suivant celles de la chaîne documentaire, nous permettra d'en décrire les enjeux et la technique.

Actuellement, les établissements universitaires et de recherche ne se sont pas suffisamment mobilisés autour de projets de gestion des thèses. Une volonté politique est primordiale pour répondre à cet enjeu scientifique et nous verrons comment le projet national STAR de dépôt, signalement et archivage des thèses peut s'articuler avec les projets d'établissement sans s'y substituer.

Enfin, le regard français vers l'extérieur étant sans doute insuffisant alors que nous pourrions gagner à conduire une réflexion commune avec

nos partenaires et à avoir plus d'échanges avec eux sur les méthodes utilisées, nous examinerons quelques expériences et projets étrangers et internationaux.

1 La thèse, le doctorant et les différents acteurs institutionnels

En France, la mise en place de la réforme LMD[1] de mise en équivalence des diplômes d'enseignement supérieur dans les pays de l'Union européenne, permettant une circulation plus aisée des étudiants, doit aussi s'accompagner d'une large diffusion, ouverte à tous, des résultats d'études. Une thèse[2] est un document résultant d'une étude doctorale. Elle est présentée devant un jury qui décide d'attribuer ou non le diplôme de docteur : la thèse est donc à la fois un document scientifique et un diplôme universitaire.

Le doctorat est délivré par les universités et les écoles normales supérieures, ainsi que par les établissements publics d'enseignement supérieur (comme les grandes écoles) autorisés par arrêté des ministres chargés de l'enseignement supérieur et de la recherche universitaire à le faire seuls ou conjointement[3].

Le doctorant effectue son travail scientifique (étude doctorale) au sein d'un laboratoire de recherche d'accueil (EPST, EPIC, entreprises, universités, etc.). À la rentrée 2003, 314 écoles doctorales accréditées par l'État avaient pour mission d'accompagner le doctorant durant sa thèse du point de vue de sa formation comme de son parcours professionnel.

Tous ces acteurs – universités et autres établissements d'enseignement supérieur, écoles doctorales, laboratoires d'accueil – ont un rôle identifié dans le parcours du doctorant dont l'aboutissement sera la soutenance de la thèse et la délivrance du doctorat. Mais les services de l'université et le ministère chargé de l'enseignement supérieur auront une vision plus administrative de la thèse et se préoccuperont plus de son signalement dans un catalogue collectif alors que les laboratoires d'accueil et les écoles doctorales verront dans une thèse un document scientifique dont l'enjeu est d'abord sa visibilité.

[1] Voir le glossaire des sigles page 279.
[2] Sont exclues de cet article les thèses, dites d'exercice, présentées par des praticiens (médecins, pharmaciens, etc.) dont le travail de recherche ne correspond pas à l'étude doctorale telle que décrite dans l'arrêté du 25 avril 2002.
[3] Arrêté du 25 avril 2002 relatif aux études doctorales (JO n° 99 du 27 avril 2002, p. 7633-7635).

Les thèses, dans leur fond comme dans leur forme, sont très diverses. L'intérêt scientifique d'une recherche est valable à plus ou moins long terme en fonction des disciplines. De même, le volume d'une thèse ainsi que sa structure peuvent varier en fonction des pays et des disciplines. Enfin, les thèses contiennent plus ou moins de signes cabalistiques divers, de formules, procédés, figures, tableaux et autres objets.

Cependant, on remarque qu'en France cette hétérogénéité est compensée par un périmètre connu : dix mille thèses sont soutenues en moyenne par an. L'ampleur en est maîtrisable et peut donc inciter à s'engager dans leur gestion tant au niveau national qu'à celui d'un établissement.

Diane Le Hénaff-Stitelet

est ingénieur systèmes d'information à l'Institut national d'études agronomiques (INRA), Route de Saint-Cyr, F-78000 Versailles (lehenaff@versailles.inra.fr), et enseigne les technologies de l'information et l'IST à l'Université de Versailles Saint-Quentin.

Catherine Thiolon

est adjointe au délégué général aux systèmes d'information à l'INRA (catherine.thiolon@versailles.inra.fr) et vice-présidente de l'ADBS, chargée de la communication.

2 Les étapes de la gestion des thèses

Gérer des thèses, notamment de façon électronique, revient à suivre chacune des étapes de la chaîne documentaire :
- la production par les doctorants ;
- le dépôt par ces doctorants ou la collecte par un professionnel de l'information-documentation ;
- le traitement documentaire plus ou moins complexe ;
- la diffusion des documents.

Accompagnement de la production de la thèse par les acteurs de proximité

Les écoles doctorales ainsi que les laboratoires d'accueil permettent au doctorant de trouver les ressources informatiques et de formation nécessaires à l'accomplissement de son projet de recherche. Celles-ci pourront l'aider notamment dans la rédaction de sa thèse, en lui indiquant comment respecter les feuilles de style et appliquer les métadonnées nécessaires au traitement ultérieur du document.

Au-delà de ces aspects plus techniques relatifs à la structure du document, la formation du doctorant, souvent dispensée par des professionnels de l'information-documentation, est aussi l'occasion de le sensibiliser aux enjeux de la diffusion des résultats scientifiques (visibilité, valorisation, notoriété), au développement des archives ouvertes, au droit d'auteur – toutes notions qui pourront lui être très utiles par la suite dans sa carrière de chercheur.

Dépôt par le doctorant ou collecte par le professionnel de l'information-documentation ?

Le dépôt du document dans une archive documentaire sollicite l'acteur « auteur » ou l'acteur « représentant de l'établissement ».

Dans le cas où le représentant de l'établissement, par l'intermédiaire du professionnel de l'information-documentation, est chargé de collecter les thèses, cette opération nécessitera d'identifier l'ensemble des travaux en cours ou soutenus, de sensibiliser les doctorants aux avantages du dépôt institutionnel et aux enjeux de diffusion des thèses, de recueillir leur accord préalable dans le respect du droit d'auteur. C'est alors le professionnel de l'I-D qui remplit la notice bibliographique accompagnant la thèse numérique récupérée.

Dans le cas où l'auteur est à l'initiative du dépôt, celui-ci devra importer sa thèse dans un système informatique mais également fournir des données descriptives et administratives relatives à ce document. Ce mode de dépôt tend à se développer pour plusieurs raisons :

- d'une part, le doctorant est la meilleure source pour fournir les éléments descriptifs de la thèse. Notons que ce mode de dépôt confère au professionnel de l'information-documentation un rôle de conseil et de formation en amont, puis en aval de validation dans la chaîne documentaire si le système comporte un *workflow ;*
- d'autre part, dans ce mode de dépôt, le droit d'auteur est naturellement respecté : la responsabilité de la gestion électronique du document incombe au doctorant qui, dans l'acte de dépôt, autorise sa diffusion électronique[4] ;
- enfin, les scientifiques commencent à déposer leurs articles dans la multitude d'archives institutionnelles ou disciplinaires disponibles actuellement, à des fins de visibilité, de diffusion rapide des résultats de la recherche, mais également pour répondre aux recommandations de leur organisme d'affiliation. Le cas des thèses n'est pas différent car il s'agit bien de documents scientifiques résultant de travaux de recherche.

En conséquence, les systèmes documentaires accueillant des thèses ne sont pas forcément indépendants de systèmes accueillant des articles scientifiques. Dans ce cas, le mode de dépôt devient commun ; c'est le cas par exemple du système TEL du CCSD qui est dorénavant intégré dans le système générique HAL.

Le traitement documentaire de la thèse

Il appartient au professionnel de l'I-D de traiter un format d'entrée et de le convertir éventuellement en un ou plusieurs formats de sortie. La plus ou moins grande complexité de cette chaîne de traitement réside dans les étapes de la conversion. S'il n'y en a pas, le format de sortie sera le même que celui d'entrée. Le format PDF est le plus fréquemment utilisé dans ce cas.

Les chaînes de traitement plus complexe utilisées dans le cas des thèses convertissent un document natif que l'on souhaite structuré (DOC, RTF ou SXW) en un autre au format XML respectant la DTD TEI standard ou une DTD spécifique au système. L'utilisation d'une feuille de style XSL permettra de proposer, en sortie, les formats XHTML (traitement XSL-T) et PDF (traitement XSL-FO). [*Voir le schéma ci-contre.*]

La mise en œuvre optimale d'une chaîne de traitement complexe suppose surtout que le doctorant ait respecté la feuille de style liée à l'éditeur de texte Word ou OpenOffice Text et également que le système informatique recoure à la DTD TEI ou TEI-Lite généralement utilisée dans la chaîne de conversion des thèses. Les métadonnées sont alors extraites automatiquement ; ainsi le doctorant n'aura pas à ressaisir, sous forme de notice bibliographique, une information présente dans le document.

La partie qui précède occulte volontairement le cas d'un document natif en format LaTeX, surtout utilisé en mathématiques et autres sciences exactes. Si actuellement aucun traducteur du format LaTeX vers XML n'est disponible sur le marché, des traducteurs existent[5] et une solution reposant sur un script de conversion est utilisée dans l'édition scientifique française[6].

L'intérêt de diffuser le document PDF tel qu'il a été importé réside sans aucun doute dans la mise en ligne très rapide de celui-ci. Mais la structuration du document, si elle est contraignante à l'étape de production, permet de séquencer le contenu et d'identifier des objets scientifiques spécifiques contenus dans une thèse. La chaîne de conversion XML décrite offre de réels avantages : d'une part, l'extraction des métadonnées de gestion et de certaines métadonnées descriptives est automatique et ne nécessite pas de saisie particulière de la part de l'auteur ; d'autre part, elle optimise la recherche d'information à l'usage de la

[4] L'auteur ou les acteurs de la thèse doivent donner leur accord de diffusion pour chaque support (papier, électronique, microformes, etc.). Cette autorisation est révocable.
[5] Le logiciel TRALICS crée par l'équipe Apics (Analyse et problèmes inverses pour contrôle et le signal) de l'INRIA (www-sop.inria.fr/apics/tralics) et le logiciel HERMES développé par l'université de Brême (Allemagne) traduisant le LaTex vers Unicode XML et MathML (http://hermes.aei.mpg.de).
[6] Voir : Jean-Paul Jorda, Marie-Louise Chaix, Ahmed Mahboub « LaTex et XML dans la chaîne éditoriale d'EDP Sciences », *Cahiers GUTenberg*, mai 2001, n° 39-40.
[7] Étude de la mise en oeuvre du dispositif national de diffusion des thèses par voie électronique dans les établissements d'enseignement supérieur et de recherche : note de synthèse. www.sup.adc.education.fr/bib/Acti/These/rapportV5.doc
[8] Voir : Cédric Dumas, « CASTORE, une plate-forme de bibliothèque numérique de littérature grise », *Biblioacid*, octobre 2005, vol. 2, n° 3, http://biblioacid.typepad.com/ba/pdf/Bav2n3.pdf

communauté scientifique. En effet, les thèses sont volumineuses et l'usager appréciera d'accéder directement au procédé scientifique recherché décrit dans une thèse plutôt que de devoir la parcourir entièrement pour trouver cette information. Enfin, l'usage d'XML garantit la pérennité des données et l'indépendance vis-à-vis des fournisseurs informatiques, même si les garanties sur le format PDF ont été renforcées par la disponibilité de son code source et par sa large utilisation.

L'inconvénient majeur de la chaîne de traitement complexe est son coût humain, qui a aussi pour conséquence l'allongement du délai de mise en ligne. Les thèses produites étant de qualité insuffisante en ce qui concerne le respect des feuilles de style (15 % des thèses déposées sous forme électronique en tiendraient compte), il faut une intervention lourde de professionnels de l'information-documentation, évaluée entre un et trois jours en moyenne pour une thèse normalement stylée. Les erreurs les plus fréquentes rencontrées lors de la vérification préalable ou identifiées par échec de la chaîne de traitement sont la mauvaise utilisation des styles (titre, texte, image, tableau, équation, etc.), la création manuelle de table des matières, le poids excessif ou la non-conformité des images, l'assemblage de plusieurs fichiers pour l'impression, l'affichage disparate des objets Word et également des problèmes avec des caractères spéciaux... Ces erreurs peuvent parfois prolonger la durée d'intervention sur la thèse à une semaine de travail à temps plein. La Sous-Direction des bibliothèques et de la documentation a estimé dans une récente note de synthèse[7] que le traitement par la chaîne de conversion XML des dix mille thèses soutenues en France chaque année nécessiterait l'emploi de cent cinquante documentalistes à temps plein (ETP) !

Notons qu'il existe en France trois outils intégrant une chaîne de traitement des documents structurés par conversion XML : Sparte gérée par l'ABES, Cyberdocs maintenue par une communauté d'utilisateurs de toutes origines, principalement francophones, sous l'impulsion de l'Université Lyon 2 [*voir page 280*], et le dernier-né, Castore, développé par l'École des mines de Nantes[8].

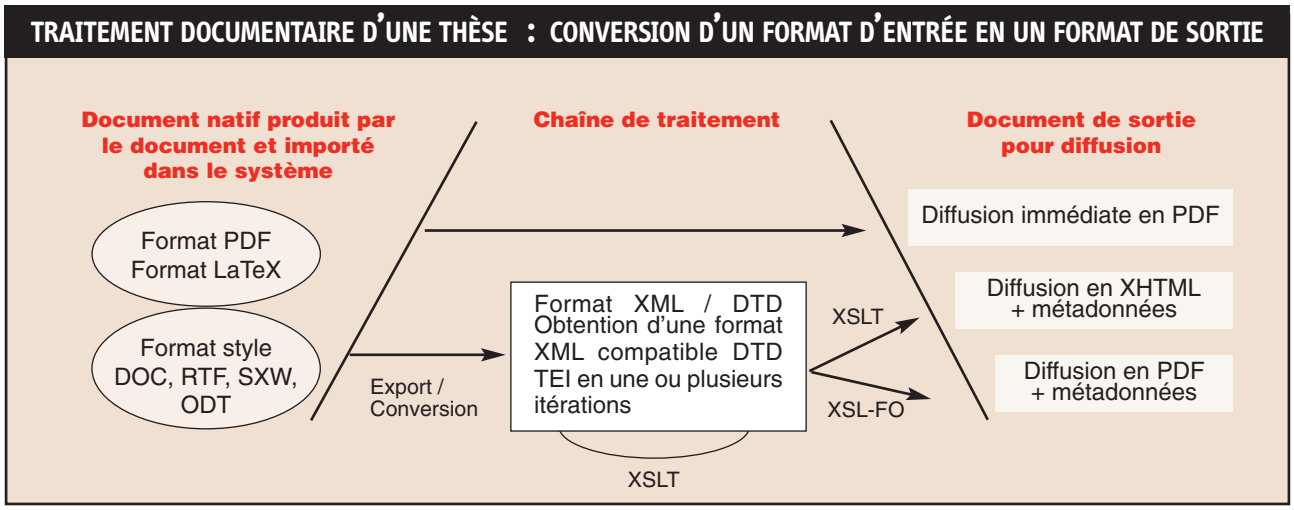
La diffusion des thèses : un enjeu scientifique

Actuellement quelques bases de données ou catalogues en ligne permettent le signalement de certaines thèses en France. Ces catalogues permettent d'identifier une thèse par ses éléments administratifs et de localiser l'établissement de soutenance, mais le téléchargement de la version électronique du document est rarement proposé. L'usager doit donc effectuer une demande de reproduction auprès de cet établissement et il obtiendra une version papier de la thèse… au bout d'un certain temps et souvent moyennant une contrepartie financière. Les performances actuelles des technologies de l'information ainsi que la production des documents en numérique devraient permettre une diffusion plus large des thèses et un accès plus aisé pour les usagers.

Les enjeux de la diffusion électronique des thèses dépassent le caractère technologique et administratif. En effet, la recherche scientifique originale décrite dans une thèse doit être valorisée et accessible à l'ensemble de la communauté, car le jeune chercheur, qui effectue ici ses premiers travaux, a besoin de notoriété.

Les expériences actuelles de diffusion montrent que la consultation des thèses en ligne est élevée et répond donc à une réelle demande : plus de 1 000 téléchargements par jour sont comptabilisés par l'archive du CCSD qui comprend 4 000 thèses, une thèse de l'archive Pastel pour le réseau ParisTech qui en comprend 636 est téléchargée huit fois par mois en moyenne… Ces chiffres sont sans aucune mesure avec les demandes de consultation des thèses sur papier, évaluées en moyenne à une fois tous les dix ans !

TRAITEMENT DOCUMENTAIRE D'UNE THÈSE : CONVERSION D'UN FORMAT D'ENTRÉE EN UN FORMAT DE SORTIE



Exemples de logiciels libres de création d'archives

Les plus connus

DSpace, développé conjointement par les bibliothèques du MIT (États-Unis) et la société Hewlett-Packard

Eprint développé par l'université de Southampton (Royaume-Uni)

Greenstone, projet de la bibliothèque numérique de Nouvelle Zélande, développé et distribué en partenariat avec l'Unesco

Cyberdocs, plate-forme de conversion et de production de documents au format XML, intégrant le système SDX pour la diffusion des documents sous différents formats. Cyberdocs a reçu le soutien de l'Agence intergouvernementale de la francophonie [voir aussi pages 280-282]

À noter aussi

CDSS (CERN Document Server Software), le logiciel utilisé par le CERN à Genève

Archimède, une initiative plus récente de l'Université Laval (Québec)

Une production française : **HAL,** produit par le CCSD, un logiciel spécifiquement adapté pour permettre le transfert automatique des données vers l'archive ArXiv basée à l'Université de Cornwell (États-Unis)

Castore, autre production française développée par l'École des mines de Nantes.

Les textes réglementaires

L'arrêté du 25 septembre 1985 relatif « aux modalités de dépôt, signalement et reproduction des thèses ou travaux présentés en soutenance en vue du doctorat » (JO du 21 novembre 1985, p.13496-13497) a été suivi par une circulaire d'application du 6 novembre 1985. Il est suivi par la circulaire d'application du 6 novembre 1985.

À noter que, contrairement à un arrêté, les circulaires n'ont pas de portée juridique obligatoire.

Si la circulaire du 11 mars 1996 concernant le signalement des thèses ne modifie pas en profondeur les préconisations de l'arrêté de 1985, celles du 21 septembre 2000 et du 29 mars 2005 (circulaire n° 05-094, « Dépôt, signalement, diffusion et archivage des thèses sous forme électronique », www.sup.adc.education.fr/bib/) introduisent la notion de thèses électronique.

Comme les circulaires ne doivent pas comporter d'éléments nouveaux ni de modifications par rapport à un arrêté précédemment paru, le ministère délégué à l'Enseignement supérieur et à la Recherche prévoit un nouvel arrêté en 2005 destiné à remplacer le précédent.

Il existe d'autres textes réglementaires relatifs aux thèses ; nous n'indiquons ici que ceux qui ont un lien avec le sujet de cet article.

En savoir plus sur la technique

TEF (Thèses électroniques françaises)

Les notices bibliographiques comportent actuellement des éléments de description et de gestion de la thèse que l'on nomme communément métadonnées. Ces métadonnées peuvent être écrites informatiquement en plusieurs formats.

TEF est un jeu de métadonnées qui concerne la description des thèses électroniques*. La recommandation TEF de l'Afnor a une valeur normative. La première version TEF 1.0 ne couvre que les métadonnées descriptives ; les métadonnées de gestion sont en cours d'élaboration (leur publication est annoncée pour la fin 2005).

Les logiciels actuels utilisent les formats Dublin Core (DC), ETD-MS recommandé par le NDLTD et DC-OAI.

TEF est un format national qui sera notamment utilisé dans l'application STAR de l'ABES. Des feuilles de style permettront la transformation des métadonnées format DC-OAI et ETD-MS en métadonnées TEF ; ainsi, les échanges de notices entre systèmes pourront se faire par conversion.

OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting)

Lors de la convention de Santa Fé, en 1999, l'Initiative pour les archives ouvertes ou OAI a mis au point un protocole OAI-PMH permettant l'échange de données entre deux applications : un entrepôt et un moissonneur.

Le moissonneur envoie des requêtes auxquelles l'application cible répond en terme de métadonnées. Ces métadonnées traduisent le contenu de l'entrepôt c'est-à-dire le fonds documentaire de l'archive. Plusieurs jeux de métadonnées sont possibles mais le format Dublin Core est requis (DC-OAI).

Une archive qui est compatible OAI-PMH est dite « ouverte »**.

Les spécifications sont disponibles sur le sitewww.openarchives.org/

* Voir : *Documentaliste – Sciences de l'information*, avril 2005, vol. 42, n° 2, p. 92.

** Voir aussi : Christine Aubry et Joanna Janik (dir.), *Les Archives Ouvertes : enjeux et pratiques*, ADBS Éditions, 2005, 332 p.

Le problème des versions

Si certains projets incluent, au moment du dépôt, une vérification de la version finale de la thèse validée par le jury de soutenance (ie : comportant les modifications éventuelles), comme c'est le cas actuellement pour l'Université Lyon 2 [*voir page 280-282*], il semble difficile de généraliser cette démarche à tous les établissements mettant en place un projet de diffusion électronique des thèses. En effet, les établissements de soutenance bénéficient de la proximité des services centraux alors que les laboratoires d'accueil revendiquent plutôt une proximité avec le doctorant. Dans ce dernier cas, ce sera plutôt celui-ci qui effectuera le dépôt du document et endossera alors la responsabilité de la version déposée.

D'un point de vue scientifique, la démarche de dépôt par le doctorant est similaire aux dépôts d'articles scientifiques version « préprint » et version « postprint » qui, sauf dans certaines disciplines, ne semblent pas un frein à l'auto-archivage ni à la diffusion des articles au sein d'une communauté.

Cependant, la multiplicité des versions accessibles d'une même thèse (établissement de soutenance, laboratoire d'accueil et école doctorale sont susceptibles de diffuser le même travail dans des versions différentes) peut créer une confusion pour un usager en recherche d'information.

Pour garantir la pérennité du document, il semble nécessaire de mettre en place un système permettant l'archivage de sa version finale en plus de son signalement. Le projet STAR actuellement en cours de réalisation au sein de la Sous-Direction des bibliothèques et de la documentation va dans ce sens ; ce projet est évoqué plus loin.

3 Les logiciels

Le logiciel ou système informatique constitue la brique technique du projet de mise en place d'un système d'information visant à gérer les thèses ou un fonds de documents scientifiques. Une large place est faite dans cet article aux systèmes d'information spécifiques aux thèses, mais il est important de considérer qu'un grand nombre d'établissements mettent en place des systèmes d'information plus globaux. Le choix d'un système spécifique aux thèses ou plus global a des conséquences sur le choix du logiciel [*voir ci-contre*].

Des logiciels d'archives ouvertes tels que DSpace ou Eprint sont utilisés directement pour l'archivage et la diffusion de documents scientifiques, mais ils ont également inspiré la création de systèmes informatiques plus complexes et adaptés aux besoins de certaines institutions, en intégrant notamment des *workflows* spécifiques. Ces logiciels sont libres d'utilisation et d'adaptation, ils respectent des normes d'échanges de données et sont interopérables.

Cette interopérabilité est essentielle car les thèses sont disséminées dans de multiples archives. Grâce au protocole OAI-PMH (pour les archives compatibles), les éléments descriptifs d'une thèse sont moissonnés et deviennent disponibles dans des bases de référence, elles-mêmes indexées par les moteurs de recherche usuels.

A contrario, lorsque la gestion des thèses (et autres documents scientifiques) s'inclut dans un projet plus global nécessitant une refonte du système d'information institutionnel, les établissements procèdent en général à un appel d'offre auquel répondent beaucoup de sociétés commercialisant des logiciels propriétaires. Dans ce cas, l'interopérabilité avec d'autres systèmes, fondée sur des standards comme le protocole OAI-PMH, ou la compatibilité Open-URL n'est, en général, pas proposée ou l'est très peu. Il est à souhaiter que les sociétés commercialisant des logiciels propriétaires intègrent rapidement les derniers usages en documentation.

4 Les projets d'établissement émanent d'un choix politique

La mise en place d'un projet d'établissement concernant les thèses électroniques relève avant tout d'un choix politique. L'implication de la direction de l'établissement dans ce type de projet est une garantie de réussite pour sa mise en place.

Avant tout choix d'une solution technique, il faut répondre à un certain nombre de questions de type stratégique, organisationnel puis fonctionnel. Les principales sont les suivantes :

- quel est l'objectif principal visé par la mise en place d'un système de gestion des thèses ? Souhaite-t-on faciliter le signalement des thèses et le fonctionnement administratif ou/et améliorer la notoriété et la visibilité scientifique de l'établissement ?
- ce projet s'inscrit-il dans une refonte globale d'un système d'information institutionnel ou est-il spécifique aux thèses ou aux documents scientifiques ?
- un projet de partenariat avec d'autres établissements est-il envisagé ?
- quels sont les moyens humains disponibles et mobilisables à court, moyen et long termes ?
- quelles sont les priorités en terme d'image de ►

l'établissement : une mise en ligne rapide et exhaustive des thèses ou un accès de qualité à des documents structurés ?

Gérer les thèses d'un établissement, c'est également suivre les différentes étapes de la chaîne documentaire :
- participer à l'étape de production de la thèse suppose la mise en place de programmes de sensibilisation et de formation des doctorants ;
- le choix du déposant, doctorant ou professionnel de l'information-documentation, n'est pas anodin car il impliquera éventuellement la mise en place d'un *workflow* et le paramétrage de l'interface de dépôt ou bien la mobilisation d'un professionnel visant l'exhaustivité de la collecte ;
- le choix de traiter un document structuré et converti peut paraître actuellement philosophique ou militant, mais les expériences en cours ont montré les avantages et inconvénients dont il est sage de s'imprégner. Il semble que la durée nécessaire au traitement d'une thèse ne soit pas compatible avec les besoins rapides de mise en ligne du document : le choix de la chaîne de traitement complexe doit donc être appréhendé en termes de moyens humains et de compétences disponibles ;
- la diffusion est un enjeu majeur dans un projet d'établissement. La visibilité du fonds suppose que l'archive ait atteint une masse critique permettant d'être connue par les usagers et reconnue par le doctorant qui souhaitera y déposer sa thèse. Cette visibilité augmente la notoriété d'une archive et *a fortiori* de son établissement.

5 Le signalement national des thèses en France : le projet STAR

Le ministère délégué à l'Enseignement supérieur et à la Recherche est responsable de l'application des textes réglementaires relatifs aux thèses, qui sont des documents administratifs soumis à des règles, notamment en ce qui concerne leur diffusion. Ainsi l'avis conjoint du jury et du président de l'établissement de soutenance, ainsi que celui de l'auteur, sont requis pour la diffusion de toute thèse.

Ce cadre réglementaire n'a pas de lien avec les enjeux scientifiques de la thèse décrits dans cet article, mais il permet de comprendre le souci de dépôt de la version validée de ce document, de son archivage pérenne, ainsi que de son signalement dans le catalogue collectif SUDOC de l'ABES. Un tel projet est nécessaire pour la centralisation de l'information au niveau national, utile pour la recherche d'information et rassurant quant à l'existence d'au moins une version officielle de chaque thèse, disponible et accessible.

Ce projet sera mis en application grâce à l'outil STAR, géré par l'ABES qui le lancera fin 2005[9]. Un *workflow* pour les établissements sera implémenté afin d'assurer la prise en compte de thèses validées. Les formats d'archivage prévus sont PDF et/ou XML. Une version pour diffusion sera créée, archivée sur le serveur du CINES et, le cas échéant, le document sera également accessible sur le serveur de l'établissement via l'URL. La notice bibliographique sera signalée et indexée dans le SUDOC.

À noter qu'une exportation des thèses vers le système HAL du CCSD serait prévue, ce qui augmenterait considérablement le nombre de thèses accessibles via cette archive.

Il semble cependant que la réflexion sur les processus associés à la gestion des thèses pour la mise en place du projet national n'ait pas bien pris en compte les systèmes locaux d'établissements déjà existants qui s'appuient sur les jeux de métadonnées internationales, notamment ETD-MS et DC-OAI.

En effet, STAR a été conçu comme le point d'entrée des données sur les thèses : saisie directement dans l'outil ou fonction d'import (fonction prévue mais non encore disponible) qui ne supporte que la norme TEF et oblige ainsi les systèmes d'établissement à convertir les métadonnées dans ce nouveau format pour permettre l'échange des données. Cette contrainte est importante car elle est multipliée par le nombre de systèmes existants qui, pour la plupart, n'utilisent pas la nouvelle norme TEF. Il aurait été préférable pour que les établissements acceptent cette nouvelle application STAR que la conversion (de ETD-MS vers TEF, par exemple) se fasse au niveau de l'application nationale cible. Espérons que ces possibilités de conversion seront implémentées dans STAR lors de la livraison de la fonction d'import.

6 Regard vers l'international

Si les projets en place ou actuellement en cours en France sont d'envergure nationale, deux ouvertures françaises vers l'international existent ou ont été expérimentées :
- le projet d'origine francophone Cyberthèses, dont l'Université Lyon 2 a pris l'initiative, implique actuellement différents établissements de plusieurs pays : outre la France, l'Université Cheik Anta Diop à Dakar, la faculté de médecine de Bamako au

[9] Voir : Yann Nicolas, « STAR, carrefour des thèses électroniques françaises », *Arabesques*, oct.-nov.-déc. 2005, n° 40, p.4-5.

Liste des sigles cités

ABES Agence bibliographique de l'enseignement supérieur
CCSD Centre pour la communication scientifique directe
CINES Centre informatique national de l'enseignement supérieur
DTD document type definition
EPIC établissement public à caractère industriel et commercial
EPST établissement public à caractère scientifique et technologique
ETOL Electronic Theses On Line
ETP équivalent temps plein
FAIR Focus on Access to Institutional Resources
HAL Hyper article en ligne
INRIA Institut national de recherche en informatique et en automatique
IST information scientifique et technique
JISC Joint Information Systems Committee
LMD licence, master, doctorat
MIT Massachusetts Institute of Technology
NDLTD Networked Digital Library of Theses and Dissertations
OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting
PDF Portable Document Format
STAR Signalement des tThèses, a Archivage et Recherche
SUDOC Système universitaire de documentation
TEF Thèses électroniques françaises
TEI Text Encoding Initiative
TEL Thèses en ligne
URL Uniform uniform Resource resource Locatorlocator
XHTML eXtensible HyperText Markup Language
XML eXtensible Markup Language
XSL-FO eXtensible Stylesheet Language Formatting Objects
XSLT eXtensible Stylesheet Language for Transformation

Mali, le réseau des bibliothèques de Madagascar, l'Institut national agronomique d'Algérie et plusieurs universités au Chili ;
- le CCSD est la seule institution française membre du comité international de réflexion NDLTD chargé de promouvoir l'adoption, la production, l'utilisation, la diffusion et l'archivage des thèses électroniques.

On notera que, contrairement aux articles scientifiques, écrits le plus souvent en langue anglaise, les thèses sont majoritairement rédigées dans la langue du pays de soutenance. Cette spécificité pose le problème de la création d'archives multilingues. L'équipe du CCSD a expérimenté un partenariat avec l'Allemagne à travers le projet ETOL, mais celui ci n'a pas atteint les objectifs visés notamment du fait des problèmes d'indexation multilingue. Cette initiative mérite tout de même d'être soulignée et saluée.

Il est souhaitable, au moment où se construit le paysage des thèses électroniques en France, de s'intéresser aux modèles de nos partenaires internationaux : du 28 au 30 septembre, a eu lieu à l'Université de New South Wales à Sydney, en Australie, le huitième colloque international sur les thèses électroniques[10].

Pour exemple, le schéma au Royaume-Uni n'est pas fondamentalement différent du nôtre. Susan Copeland, de l'Université Robert Gordon, a décrit, lors d'une récente journée d'étude organisée par l'ADBS[11], la mise en place des différents projets de diffusion des thèses électroniques qui ont conduit à une véritable politique nationale en la matière. Prenant le relais d'une réflexion datant des années quatre-vingt-dix, le JISC, organisme britannique chargé de la mutualisation des systèmes d'information, a instauré le programme FAIR destiné à financer et à suivre des projets dans le domaine des thèses électroniques.

Ainsi, pour la période 2002-2004, une étude sur la diffusion électronique des thèses a été menée par l'Université Robert Gordon au sein d'un consortium de cinq universités. Cette étude a abouti à un certain nombre de recommandations : utilisation par les universités d'un système compatible OAI-PMH (Eprint), dépôt par les doctorants et liberté du choix de format pour la rédaction des thèses et leur diffusion (le PDF est généralement utilisé)...

De 2004 à 2005, un projet, mené par l'Université de Glasgow au sein d'un consortium de dix universités, avait pour but de mettre en pratique ces recommandations. Le système résultant est un modèle d'infrastructure hybride qui repose sur l'utilisation de logiciels libres, du protocole d'échange de données OAI-PMH et de métadonnées respectant les normes internationales : mise en place d'archives institutionnelles au sein des universités et création d'un catalogue collectif géré par la British Library, qui, à travers une interface unique d'interrogation, permet de retrouver les thèses disséminées.

Si, *in fine*, les solutions techniques sont comparables à celles utilisées en France, le management du processus mis en place par le JISC est différent de l'approche du ministère français. Le JISC finance des projets sélectionnés lors d'appels d'offres : ils émanent d'universités ou de consortiums d'universités volontaires et sont ensuite suivis et évalués suivant le mode « gestion de projet ». ►

[10] 8th International Symposium on Electronic Theses and Dissertations (ETD 2005), Sydney, Australia, 28-30 September 2005.
[11] Cette journée d'étude sur la diffusion des thèses électroniques a eu lieu le 7 juillet 2005 au ministère de la Recherche. Un panorama des politiques nationales française et britanique a tout d'abord été présenté, puis des retours d'expériences de divers établissements - CCSD, Université Lyon 2, INRA, ParisTech (réseau des écoles d'ingénieurs de Paris) - qui ont fait des choix politiques et techniques parfois différents. Le compte rendu de cette journée, ainsi que les différentes présentations des intervenants sont accessibles pour les adhérents ADBS à l'adresse :
www.adbs.fr/site/evenements/journees/journee.php?limit=0&annee-2005&id=73&version=1
La qualité des interventions a été une source d'inspiration et de réflexion pour la rédaction de cet article.

7 Investir dans la gestion des fonds de thèses

L'expérience britannique décrite ci-dessus nous montre que le projet de gestion des thèses à l'échelle nationale du Royaume-Uni émane des universités qui participent à sa construction. C'est la démarche *bottom up* qui garantit que tous les acteurs acceptent ce projet et participent à sa mise en place. En France, malgré une consultation effectuée auprès de quelques acteurs locaux par la représentation nationale, la démarche actuelle est plutôt *top down*.

La gestion et la diffusion des thèses électroniques répondent à un enjeu à la fois administratif et scientifique. Ces deux dimensions doivent être complémentaires et non en opposition : c'est pourquoi l'application nationale STAR n'a pas vocation à se substituer aux systèmes locaux de proximité.

Malgré la diversité des choix politiques et techniques que l'on peut trouver dans les différentes expériences menées tant en France qu'au Royaume-Uni (ou ailleurs, dans le monde académique occidental, certains projets incluant aussi des pays émergeants), il existe tout de même un noyau commun de préoccupations : logiciels libres, respects des normes, souci de l'interopérabilité, acceptation de plusieurs formats avec une diffusion minimale au format PDF, pérennité des accès, archivage, etc., preuve que les bonnes questions ont reçu des réponses adéquates sur lesquelles tout projet d'établissement peut s'appuyer.

Trop longtemps le choix technique (diffusion PDF ou/et conversion XML d'un document nativement stylé) a suscité des polémiques, mais la tendance nouvelle est d'être pragmatique. On apprend en faisant, et l'obtention d'une masse critique convainc les derniers réticents, d'autant plus qu'un choix n'est pas irréversible : il est toujours possible de compléter la technique avec les nouvelles normes et usages et/ou de proposer d'autres formats après une première mise en ligne d'un document en PDF.

L'enjeu scientifique de gestion et surtout de diffusion d'une thèse en vue de sa valorisation et du libre accès aux résultats de la recherche est tellement important que les établissements doivent prendre conscience du rôle qu'ils ont à jouer auprès des doctorants et du bénéfice qu'ils peuvent en tirer ensemble en termes de visibilité et de notoriété. Il est donc nécessaire que, en France, les établissements universitaires et de recherche fassent en plus grand nombre ce choix politique d'investir dans la gestion de leurs fonds scientifiques.

SEPTEMBRE 2005

Un outil pour la gestion des

L'archivage et la diffusion électroniques des thèses nécessitent une organisation qui possède plusieurs caractéristiques : il s'agit d'un travail sur une longue période, en plusieurs étapes, et qui réunit différents acteurs1 autour de différents documents, en format numérique ou imprimé. Le nombre de thèses à traiter varie selon les universités, mais il s'agit toujours d'un fonds dont le volume augmente régulièrement tous les ans.

À l'Université Lumière Lyon 2, dans le cadre du programme Cyberthèses, les procédures de dépôt électronique des thèses sont en vigueur depuis septembre 2000. Elles font intervenir différents services de l'établissement à différents moments du parcours : le Service de la recherche pour l'organisation administrative des soutenances, le Service de reprographie pour l'impression des exemplaires de thèses, le Service commun de la documentation (SCD) pour le référencement et le Service d'édition électronique pour la production, l'archivage et la diffusion des thèses électroniques.

Nous avons mis en place à Lyon 2 un outil qui facilite et optimise les relations entre ces différents services. Cet « Outil de Gestion Électronique des Thèses » (OGET) est à la fois une base de données, un système de GED et un système de workflow. Il s'agit d'un outil libre et ouvert, développé dans une architecture PhP-MySql, avec une interface html. Installé sur un serveur, il est accessible depuis n'importe quel poste via le réseau. Notre objectif est de favoriser la gestion dématérialisée des thèses numériques et des documents qui leurs sont liés, et de faire d'OGET un outil de travail collaboratif. Il s'agit d'un système d'information qui permet la centralisation, la diffusion et la recherche des informations et des documents, ainsi que l'analyse de l'état du fonds de thèses. On peut, par exemple, afficher toutes les thèses soutenues dans une période donnée, ou toutes les thèses dont l'auteur a autorisé la diffusion ; on peut aussi trouver une thèse à partir du prénom de son auteur ou d'un mot contenu dans son titre.

OGET dans les différentes étapes du circuit de diffusion d'une thèse électronique

1 **Le dépôt de la thèse** est une étape primordiale, puisqu'il s'agit de garantir que la thèse sera archivée et diffusée dans sa version intégrale et conforme à la version de soutenance. Ce dépôt a lieu au plus tard un mois avant la soutenance. Lors de ce dépôt, un répertoire dédié à la thèse déposée est créé, et toutes les informations disponibles à ce moment du processus y sont saisies : les coordonnées de l'auteur, les titres, les résumés, les mots clés (en plusieurs langues), l'autorisation de diffusion fournie par l'auteur, etc. Ces informations seront ensuite restituées automatiquement dans le bordereau de métadonnées, évitant ainsi une double saisie.

Les documents « sources » fournis par le doctorant, généralement produits dans un format de traitement de texte (Word, OpenOffice), sont téléchargés dans OGET, ainsi que les illustrations et les fichiers bureautiques nécessaires à la lecture du document : feuille de style, fichiers de police. Ces documents sources sont en grande partie déjà structurés par l'auteur grâce aux consignes et aux

1 L'auteur : le doctorant, le chercheur ; le(s) chargé(s) de dépôt (accueil, recueil et validation des documents, élaboration des fichiers d'impression) ; le(s) chargé(s) de production (stylage, conversion et diffusion).

thèses électroniques : OGET

formations délivrées par l'université.

OGET permet également d'éditer et d'imprimer les documents nécessaires aux démarches administratives du doctorant : document de validation du dépôt électronique, autorisation d'impression de la thèse par le service de reprographie de l'université. Enfin le chargé de dépôt peut saisir dans OGET toutes les particularités susceptibles d'être ultérieurement utiles pour le traitement de la thèse, comme le logiciel utilisé, la présence de données confidentielles, l'importance des illustrations, leur statut juridique, etc.

2 Après la soutenance, de nouvelles informations sont saisies dans OGET à partir du procès-verbal de soutenance, soit manuellement, soit grâce à une importation depuis une autre base informatique (par exemple Apogée, logiciel de la scolarité, que l'équipe du Service de la recherche utilise pour le suivi administratif). Ces informations vont compléter les métadonnées saisies au moment du dépôt. Il s'agit, entre autres de la discipline, de la faculté, des membres du jury. Il s'agit surtout de l'avis émis par le jury sur la diffusion de la thèse : il peut accorder ou refuser la diffusion ou demander des modifications. Dans ce cas, le doctorant doit effectuer un deuxième dépôt, dans les mêmes conditions que le premier.

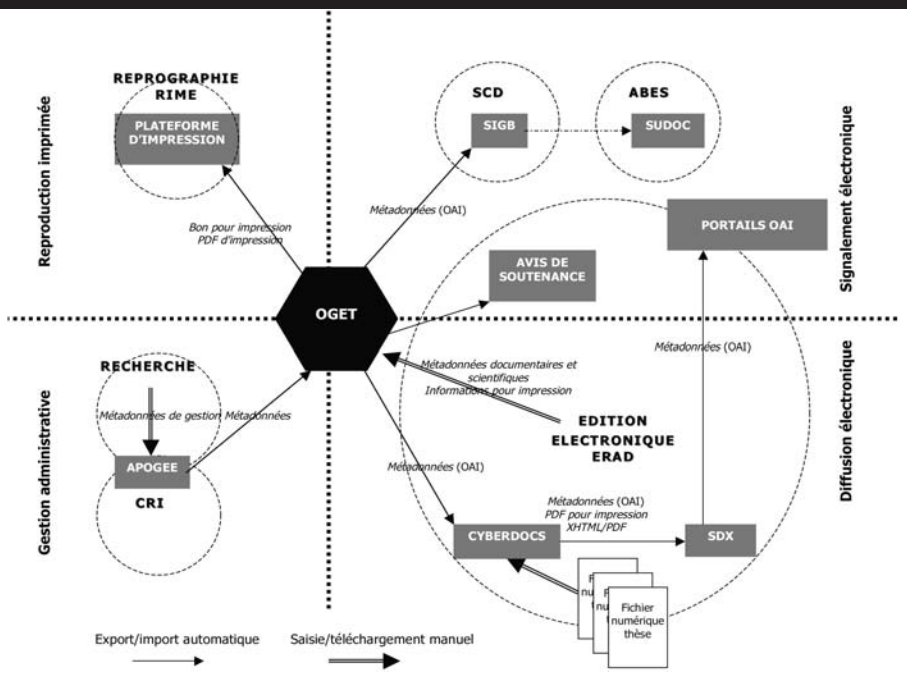
Le dépôt de la thèse doit être clos dans OGET avant de passer à l'étape du traitement, soit après la soutenance, soit après le deuxième dépôt s'il y a demande de modification. Cette procédure garantit que seule la version canonique de la thèse, validée par le jury, sera archivée et diffusée.

3 L'étape du traitement. Une fois la phase de dépôt terminée, on passe au traitement, qui consiste à vérifier la structuration de la thèse, autrement dit le stylage, pour la convertir dans une chaîne de traitement vers un format d'archivage XML, pérenne et intéropérable. OGET permet à la personne chargée de cette étape, appelée le « styleur », de récupérer le document initial, déposé dans l'espace de la thèse, directement du serveur à son poste client. Pendant le stylage, il peut consulter les remarques et les données saisies au moment du dépôt. Il peut télécharger ensuite la version définitivement structurée du document, et saisir à son tour des informations sur le déroulement du traitement de la thèse. Le styleur signale alors dans OGET qu'il a achevé les opérations qui lui incombent.

4 Les étapes de conversion et de diffusion des thèses peuvent alors commencer. Ces opérations sont effectuées dans Cyberdocs2. OGET possède une rubrique conversion et diffusion qui permet de noter quand et comment se

2 Cyberdocs : plate-forme permettant de structurer des documents issus de traitements de texte et de les publier sur Internet, à l'aide de la norme XML et la DTD TEI Lite. Elle est notamment utilisée comme plate-forme de traitement et de diffusion des thèses dans le cadre de Cyberthèses, programme francophone d'archivage et de diffusion électroniques des thèses qui regroupe aujourd'hui de nombreuses institutions francophones ou non à travers le monde. Site de diffusion des thèses soutenues à Lyon 2 : http://theses.univ-lyon2.fr ; site de développement collaboratif de Cyberdocs : http://sourcesup.cru.fr/cybertheses/

LE CADRE FONCTIONNEL DE OGET



APOGEE (Application pour l'organisation et la gestion des étudiants) : application nationale utilisée pour l'organisation et la gestion de la scolarité par les universités
CRI : centre de ressources informatiques
ERAD (Édition, reprographie, archivage et diffusion des documents) : service de publication électronique des documents scientifiques
RIME (Reprographie, Impression, MicroÉdition) : service de reprographie de l'Université Lyon 2

déroulent ces étapes. Mais le plus important, à cette étape, est de fournir le bordereau de métadonnées OAI, créé à partir des informations saisies pendant tout le circuit de la thèse, et de l'exporter au stade de la conversion (Cyberdocs) et ensuite au stade de la diffusion (SDX)ou dans tout autre système informatique comme, par exemple, le système intégré de gestion de bibliothèque (SIGB) du SCD, le catalogue du SUDOC ou prochainement STAR.

5 **Une dernière fonction** de OGET permet éventuellement d'indiquer si les différents documents électroniques liés à une thèse (documents sources, documents stylés, documents convertis) ont été sauvegardés sur d'autres supports (cédérom, DVD, serveur), suivant une procédure mise au point par l'institution de soutenance.

Une base de données et un outil de GED

Chaque thèse possède donc un espace unique dans OGET, où sont saisies toutes les informations la concernant. Cela évite la multisaisie dans différentes bases, en centralisant toutes les informations et en permettant des importations ou des exportations de données dans d'autres applications. OGET permet à chacun des acteurs d'intervenir à n'importe quelle étape du circuit pour travailler ou apporter des informations sur une thèse, et favorise ainsi le travail collaboratif. Il permet aussi de retrouver rapidement n'importe quelle thèse, ensemble de thèses ou information sur une thèse grâce à un moteur de recherche multicritères. Il permet encore le téléchargement des documents électroniques, c'est-à-dire l'accès direct à ces documents, mais aussi une sauvegarde supplémentaire sur le serveur. Il permet également de localiser physiquement le dossier de la thèse, qui contient les supports fournis par l'auteur, les imprimés légaux ou tout autre document papier lié à la thèse. Il permet enfin un suivi statistique précis du fonds des thèses électroniques et la mise en place des procédures permettant d'atteindre les objectifs fixés.

Grâce à OGET, les manipulations de documents« physiques » sont limitées et les risques d'erreur sont réduits par une saisie unique et par la possibilité de vérifier et modifier les données. La circulation des fichiers sur les différents postes de travail en est simplifiée, ainsi que l'organisation du classement et du suivi des thèses. Le partage d'information au sein d'une même équipe ou entre différents services est réelle. Les gains en terme d'espace de stockage, en temps de recherche et en temps de traitement sont importants. C'est un outil de travail pour l'ensemble du personnel intervenant dans le circuit de diffusion des thèses électroniques, et qui garantit un suivi précis de chacune.

OGET est une base de données et un outil de GED éditoriale et administrative adapté à la circulation de l'information au sein d'une institution universitaire. C'est un élément du système d'information de l'établissement, qui se caractérise par l'unicité documentaire, administrative et scientifique du traitement des thèses, garante de la qualité de l'information.

Magalie.Prudon@univ-lyon2.fr*

* Magalie Prudon est ingénieure vacataire à la division Erad Lyon2 qui a réalisé ce programme.

ÉCRAN DE PRÉSENTATION D'UNE THÈSE

Consultation d'une thèse

Revenir à la page de recherche | Modifier les informations de la thèse | Téléchargement des documents | Documents pour l'impression

Thèse n°3235

Titre : De la qualité de vie au diagnostic urbain, vers une nouvelle méthode d'évaluation
Sous-titre : Le cas de la ville de lyon

Faculté : Faculté de Géographie, Histoire, Histoire de l'Art et Tourisme
Laboratoire : Environnement, Ville, Societes (IRG)
Date de soutenance : 08/06/2005 14:30
Ecole doctorale : Sciences des Sociétés et du Droit

Auteur n°1

Nom : BARBARINO **Nom marital :** SAULNIER **Prénom :** Natalia
Adresses : 30 av. Garibaldi 69007 MEYZIEU **Num tel :** 07 51 58 58 59 **Emails :** n.saulnier@urbanismala.org
Num étu : 1515151 **Spé scientif :** géographie

Institutions - Modifier la liste

Nom : Université Lyon 2

Informations sur le dépot
Consulter - Editer

- Validation du dépôt électronique fait le **03/05/2005** par **Magalie Prudon** sur **PC**
- Support fournis par le doctorant de type : **Word 2000**
- Polices spéciales : **Oui, Mais Purement Esthétiques, Cf Cédérom**
- Documents à numériser : ***aucun***
- Dépôt bis : ***aucun***
- Dépôt ter : ***aucun***
- Remarques :
  Il ya beaucoup d'images, en format tiff : dépôt très lourd... Certaines n'apparraissent pas dans le fichier Word, mais il s'agit d'images uniquement esthétiques, donc pas la peine de les rajouter. Pensez à enlever les images flottantes. Attention, certaines images ont été insérées par lien.
- Clôture du dépôt : **19/07/2005**
- Prise en charge de l'impression : ***aucun***

Informations sur le stylage
Consulter - Editer

- Stylage commencé le **30/08/2005** par **Jérôme Serme** sur **PC** dans : ***aucun***