La Numérisation des Thèses en France: Pour une Meilleure Valorisation de la Recherche Scientifique Française

par CHRISTINE OKRET

Tels qu'ils existent actuellement, les procédés institutionnels de valorisation des thèses soutenues dans les universités et grands établissements français reposent sur deux circuits concomitants¹. En premier lieu les thèses soutenues sont déposées sous forme papier dans la bibliothèque de leur université de soutenance, où elles peuvent être consultées sur place, et prêtées par le biais du réseau de prêt entre bibliothèques. En second lieu, un exemplaire papier est transmis aux ateliers nationaux de reproduction des thèses² chargés du microfichage des textes et de la diffusion systématique de ces microfiches auprès de toutes les bibliothèques universitaires. Ces deux formes de diffusion non commerciale présentent en regard des exigences des chercheurs trois inconvénients majeurs: la diffusion des thèses est géographiquement restreinte, puisque limitée aux bibliothèques universitaires; le support microfiche est jugé peu convivial³; le prêt entre établissements est parfois trop lent ou onéreux⁴, ce qui peut décourager les utilisateurs qui désirent consulter sur papier les thèses non conservées dans leur bibliothèque universitaire.

Face à ces insuffisances, l'utilisation du réseau Internet offre la possibilité de multiplier les accès indépendamment des contingences physiques, de permettre une consultation rapide et souple grâce aux possibilités de navigation hypertexte. Une valorisation efficace des travaux scientifiques français doit donc s'appuyer sur les avantages du réseau, afin de leur assurer une visibilité accrue. Tel est le sens des diverses réflexions conduites autour de la mise en ligne de thèses.

Dans ce domaine, deux formes d'initiatives se complètent: l'expérimentation d'un serveur national de thèses pour instiller davantage de souplesse dans la diffusion institutionnelle des thèses, et dans de nombreuses universités des projets de numérisation de thèses motivés par un souci de valorisation de la production scientifique locale.

Parmi ses missions, le Ministère de l'éducation nationale, de la recherche et de la technologie se doit de veiller au signalement et à la diffusion des travaux de doctorat de façon exhaustive. Afin de tirer parti de l'utilisation des nouvelles technologies, il⁵ a lancé en partenariat avec l'Atelier national de reproduction des thèses (ANRT) de Lille, l'Agence bibliographique de l'enseignement supérieur (ABES) et le Centre informatique national de l'enseignement supérieur (CINES) un projet de serveur national de thèses numérisées. Sa viabilité est étudiée au cours d'une expérimentation menée durant l'année 1999, destinée à aborder tous les aspects, juridiques, techniques et économiques, liés à l'opération de mise en ligne et à la maintenance du service dans la perspective d'une éventuelle ouverture officielle⁶. Les thèses à numériser ont été choisies parmi les plus récemment soutenues en lettres, droit et sciences humaines, l'ANRT de Lille étant précisément chargé de microficher les thèses dans ces disciplines.

En préalable de toute intervention technique sur les textes, il convient de se prémunir sur le terrain juridique. En effet, la thèse possède un statut ambivalent. En tant que document administratif (diplôme d'Etat) elle est soumise au droit public, mais en tant qu'œuvre originale d'auteur, les règles de la propriété intellectuelle lui sont également applicables. Alors que l'arrêté du 25 septembre 1985 prévoit expressément le microfilmage des thèses⁷, rien n'est dit sur la diffusion par moyens électroniques de ces travaux. Il importe donc d'obtenir l'autorisation des auteurs pour pouvoir procéder à la numérisation et à la mise en ligne. A été proposée aux auteurs, soit la numérisation du texte intégral avec ou sans impression possible, soit la numérisation des parties significatives de la thèse (page de titre, table des matières, introduction, conclusion, avec ou sans bibliographie). Sur environ 2,500 demandes d'autorisation, 36% des auteurs ont renvoyé une réponse favorable, la majorité d'entre eux (les deux-tiers environ) se prononçant pour une mise en ligne du texte intégral. Ces résultats confirment la complémentarité de la numérisation avec les supports de diffusion traditionnels, et témoignent en outre d'un intérêt certain des chercheurs pour ce type de service.

Dans un second temps, l'opération de numérisation est effectuée à partir de la version papier⁸ détenue par l'ANRT. La thèse est numérisée en format PDF image, et découpée en fichiers suivant la construction du texte, selon un maximum de deux degrés de détail (partie – chapitre). Les fichiers sont mis sur le serveur localisé au CINES. L'accès aux thèses se fait soit à partir d'un catalogue⁹, soit par moteur de recherche (Altavista). Un choix de consultation donne accès à une page présentant la liste des fichiers de la thèse. A partir de chacun des fichiers, le texte est accessible, et la navigation d'un fichier à l'autre est libre sans revenir à la liste initiale. Cette possibilité de navigation à l'intérieur du texte de la thèse offre une grande souplesse de consultation. L'affichage des numéros de page à l'intérieur des chapitres permet le

feuilletage de l'ouvrage et l'accès rapide à une référence précise pour des recherches pointues¹⁰. La question de l'archivage suscite encore de nombreuses interrogations, parmi lesquelles la possibilité d'utiliser le format normalisé XML, plus léger et plus souple que SGML, pour structurer les thèses et, soit les diffuser telles quelles, soit à partir de ce „format-matrice“ envisager une diffusion en différents formats existants, après réalisation de programmes de conversion.

En dehors de considérations techniques et économiques, les perspectives d'évolution de ce service seront déterminées par la réaction des chercheurs et des professionnels de la documentation, dont l'opinion sera sollicitée au cours d'une phase de test qui aura lieu au second semestre 1999 dans quelques bibliothèques de sciences humaines.

A côté de ce projet d'envergure nationale existent des projets locaux complémentaires de valorisation de la production scientifique des universités, répondant à des objectifs similaires selon des modalités diverses. Parmi ces projets, deux exemples témoignent d'approches novatrices voisines avec des choix techniques divergents¹¹, démontrant ainsi la richesse des possibilités techniques utilisables dont la fiabilité doit être testée par le biais d'expérimentations multiples: l'Institut national des sciences appliquées (INSA) de Lyon et l'université Lyon 2.

Ces deux projets, actuellement en phase expérimentale, partent du même postulat: la remise d'une version électronique des thèses soutenues par les étudiants, ce qui nécessite une formation spécifique des doctorants aux feuilles de styles pour pouvoir mener l'opération de numérisation sur des thèses présentées selon des règles précises. Le projet de Lyon 2 propose également une formation des étudiants à la conversion de leurs documents afin d'assurer une certaine homogénéité des versions de traitement de texte utilisées. Les contours techniques de ces deux projets diffèrent cependant.

L'INSA de Lyon¹² a pour ses travaux en sciences mis au point une procédure semi-automatique de conversion de fichier pour passer des formats natifs (tels que Word, LaTeX) à PDF texte, format de diffusion actuellement utilisé. L'accès aux textes intégraux des thèses se fait par catalogue, par liste d'auteurs classés par année ou par texte intégral grâce à un moteur de recherche Altavista local indexant les textes entiers des thèses. Le découpage en fichiers est fin, et permet une consultation efficace augmentée par la possibilité de visualiser les pages par vignettes (ce qui permet de voir les nombreux schémas présents dans ces types de thèses). Ce projet offre plusieurs pistes d'évolution. Sur un plan strictement local et propre à l'INSA, la fonction valorisatrice de ce site doit être développée par la mise en place d'un comité scientifique de sélection des thèses chargé de donner son accord à la mise en ligne des travaux soutenus. Par ailleurs, l'INSA inscrit son

expérimentation dans une dynamique régionale forte en matière de documentation électronique en cherchant à développer avec d'autres organismes documentaires rhône-alpins une collaboration qui peut prendre des formes diverses: concertation pour définir un guide de présentation des thèses, réflexion sur les méthodes d'archivage, coopération avec les centres de documentation et le service commun de documentation de Lyon 1 situés sur le même campus (la Doua) pour accroître la portée du serveur de thèses scientifiques, plus généralement échange régulier d'informations.

Le même souci de mise en valeur locale caractérise le projet de Lyon 2¹³, qui se place cependant dans une perspective particulière en participant à un projet de diffusion des travaux scientifiques francophones avec l'université de Montréal. Lyon 2 a constitué une base de données locale des thèses soutenues (en sciences humaines) depuis 1902, aux notices de laquelle il est envisagé d'ajouter une URL afin d'accéder directement au texte de la thèse. L'expérimentation actuellement menée concerne principalement les formats de diffusion et d'archivage des textes, centrée essentiellement sur le format SGML et ceux qui en découlent, HTML et XML, afin de réaliser une structuration des documents permettant un traitement efficace des contenus et l'exploitation des métadonnées pour leur indexation. L'objectif est de parvenir à une automatisation poussée de la procédure de numérisation des textes.

Ces quelques exemples de projets en voie de réalisation ne doivent cependant pas masquer l'existence d'une dizaine d'autres projets de même type, en cours d'élaboration selon des choix techniques très divers (mise en ligne ou sur cédérom, diffusion en mode texte ou image). En effet, deux conditions essentielles convergent pour créer un contexte favorable à leur développement: l'intérêt des chercheurs pour une production riche mais insuffisamment visible, et un progrès technologique permettant la mise à disposition rapide, conviviale et relativement peu onéreuse de ces travaux scientifiques.

ANNOTATIONS

- 1 L'arrêté du 25 septembre 1985 relatif aux modalités de dépôt, signalement et reproduction des thèses ou travaux présentés en soutenance d'un doctorat décrit précisément ces circuits.
- 2 Il existe deux ateliers de reproduction des thèses: à Lille pour les thèses de lettres, sciences humaines et droit, et à Grenoble pour les thèses de sciences.

La Numérisation des Thèses en France: Pour une Meilleure Valorisation de la Recherche Scientifique Française

- 3 Voir les conclusions de *l'Enquête sur les pratiques des utilisateurs du signalement des thèses et des utilisateurs de thèses* (1997): <<http://www.sup.adc.education.fr/bib/acti/fnThes.htm>>
- 4 *Ibid.*
- 5 Ce projet est piloté par la Direction de l'enseignement supérieur – Sous-direction des bibliothèques et de la documentation.
- 6 Cette expérimentation s'insère dans le cadre plus vaste de la réorganisation du circuit de signalement des thèses en cours; mais les seules modalités de diffusion seront ici décrites.
- 7 Article 6.
- 8 La version papier est celle qui a été validée par le jury.
- 9 Ce projet est lié à la réalisation du nouveau catalogue collectif des bibliothèques universitaires françaises (Système universitaire de documentation). Il est prévu que les notices de thèses actuellement regroupées au sein d'un catalogue spécifique (Téléthèses) soient intégrées dans la base de notices communes à toutes les bibliothèques universitaires. Il est alors envisagé d'ajouter une URL au bas de chaque notice lorsque la thèse numérisée sera disponible.
- 10 *L'Enquête sur les pratiques des utilisateurs du signalement des thèses et des utilisateurs de thèses* citée ci-dessus fait état de pratiques de consultation des thèses caractérisées par une lecture partielle ou ponctuelle des thèses.
- 11 Les préalables juridiques sont également abordés, dans ces deux cas.
- 12 Contact Monique Joly: Monique.Joly@insa-lyon.fr, et le site du projet appelé CITHER <<http://csidoc.insa-lyon.fr/these/index.html>>
- 13 Ce projet, dont l'objectif est de valoriser les contenus de l'édition savante de la francophonie, est mené en coopération avec l'université de Montréal. Contact Jean-Paul Ducasse: ducasse@univ-lyon2.fr

Christine Okret
Ministère de l'éducation nationale, de la recherche et de la technologie
Direction de l'enseignement supérieur
Sous-direction des bibliothèques et de la documentation
christine.okret@education.gouv.fr