

Propositions relatives à la diffusion d'informations relatives aux thèses de doctorat

Introduction

Près de 65.000 doctorants travaillent en France actuellement et plus de 10.000 thèses sont soutenues chaque année. Elles constituent un gisement d'information tout à fait remarquable : souvent rédigées de manière plus pédagogiques que les articles publiés dans les revues spécialisées, elles fournissent souvent un état des lieux très complet d'un sujet, accompagné d'une bibliographie exhaustive et d'explications complémentaires qu'il est parfois difficile de trouver ailleurs.

Ces dernières années, la généralisation de l'outil informatique rend possible la constitution d'un fond documentaire inestimable, à savoir la **mise en ligne intégrale** d'un catalogue, puis de toutes les thèses et de leurs résumés, ainsi que leur indexation par un moteur de recherche. Le réseau Internet fournit le canal de diffusion et d'accès idéal pour ce fond documentaire.

L'impact sur la Recherche d'un tel outil serait considérable : que de temps gagné pour un chercheur désireux d'utiliser plusieurs années après des travaux sur un sujet donné, que de temps gagné pour un jeune doctorant dans ses recherches bibliographiques... De plus, cet outil serait un beau moyen de valoriser la Recherche et la langue Française. Enfin, il y a fort à parier que, d'ici quelques années, d'autres pays se doteront d'un outil de ce type, ce qui leur donnera un avantage de plus dans la bataille de l'Intelligence.

Mais de qui relève cette mission ? Quelle organisation est la plus à même de mettre en place cet outil ? Nous pensons que cela est du ressort de l'État. En effet, d'une part les informations comme le titre, le résumé de la thèse, les mots-clefs, la composition du jury ainsi que la date de soutenance ont valeur officielle. Selon l'article 28 de l'arrêté du 30 mars 1992 dont le titre 3 régit les études doctorales :

« Art. 28. - Sur le diplôme de docteur délivré figure l'indication de l'établissement de soutenance. Y figure également une indication de spécialité ou de discipline, le titre de la thèse ou l'intitulé des principaux travaux, les noms et titres des membres du jury, la mention obtenue par le titulaire. »

Mais d'autre part et surtout, seul l'État pourrait garantir la quasi exhaustivité et la pérennité d'un tel service qui s'apparente à une composante d'une bibliothèque universitaire virtuelle nationale.

1 Etat des lieux

Divers outils existent déjà comme le service **Téléthèse** mais qui ne propose qu'un accès Minitel par 3615, un **telnet** payant ou un CDROM assez onéreux. Mentionnons aussi le fond de microfilms de l'atelier de reproduction des thèses. Clairement, ces outils appartiennent au passé et vivent leurs dernières années en tant qu'outil de consultation même si le microfilmage ou le stockage d'exemplaires papier sont évidemment indispensables pour l'archivage à long terme. A l'avenir, la consultation se fera de plus en plus via des réseaux à haut débits, les recherches se faisant en direct sur une base perpétuellement remise à jour.

A la lumière de ces principes, un état des lieux peut être dressé qui montre les forces et faiblesses des différents types de serveurs existants. Nos propositions visent à combiner les avantages des différentes approches pour permettre une meilleure diffusion de l'information relative aux thèses.

1.1 Les serveurs institutionnels

L'appel de la GDD a permis de recenser un certain nombre de sites internet qui tentent de pallier à ce déficit de diffusion de l'information. Deux sites institutionnels existent toutefois :

- l'ABES propose Téléthèses
- INIST pour le CNRS.

Le serveur de l'**INIST** (ARTICLES) donne accès à toutes les thèses soutenues mais avec un retard trop important et les informations fournies sont trop limitées pour être d'une réelle utilité. Les informations plus complètes sont pourtant disponible (base PASCAL) mais en accès payant.

L'Agence Bibliographique de l'Enseignement Supérieur (**ABES**) prépare un site web pour les thèses qui devrait être en accès gratuit mais ce sont les bibliothécaires qui catalogueront et indexeront les thèses dans le système universitaire de documentation quand l'information aura été transmise par les service des thèses de l'université. Cette action rassemble l'ART de Lille (Atelier de Reproduction des Thèses), la sous-direction des bibliothèques et de la documentation du MENRT, l'Agence Bibliographique de l'Enseignement Supérieur (ABES) et le CNUSC (Centre National Universitaire Sud de Calcul) à Montpellier.

En 1999, le serveur sera encore dans une phase expérimentale qui devrait s'achever en 2000.

Source : Lettre du CNUSC - Num 67 (mais pas de serveur Web)

Mais il existe d'autres initiatives très intéressantes initiés par des structures légères (associations le plus souvent) qui démontrent la faisabilité d'index des thèses. Ils sont tous limités soit à une discipline (serveurs thématiques), soit à un organisme (serveurs locaux).

1.2 Les serveurs de thèses thématiques

- ATM (Annuaire des thèses de Mathématiques),
- Association du DEA de Biophysique de Paris: ADBPhi,
- Thèses en Génie informatique, automatique et traitement du signal sur le serveur de l'ISIS,
- Thèses en Neurosciences.

L'intérêt de ces serveurs est indéniable car il permet à une communauté (scientifique) de trouver l'information nécessaire aux avancées notamment en mettant en contact des équipes travaillant sur des sujets très proches. Le problème est que ces sites –souvent non soutenus par l'Etat– ont des difficultés à obtenir l'information nécessaire à leur fonctionnement.

1.3 Les serveurs de thèses locaux (propres à un organisme, une université)

Diverses organismes de recherche ou instituts proposent des serveurs de thèses. D'autres sont en cours d'élaboration mais nous mentionnons deux exemples opérationnels :

- Le <http://www.inria.fr/RRRT/publications-fra.html> (accès aux documents en format PDF, PS...),
- Les <http://iep.univ-lyon2.fr/IEP/Recherche/theses.html> de l'Université Lumière Lyon2 (accès restreint aux documents).

L'intérêt de ce type de site est la visibilité de la recherche de l'établissement, et les informations transitant nécessairement par les services compétents de l'organisme, l'information peut y être exhaustive. Par contre, ils sont limités à une discipline ou à un organisme bien délimité.

Un dernier projet (REDOC) nous semble intéressant et sera décrit dans la dernière partie.

2 Les problèmes posés

Tout d'abord, la typologie des différents serveurs que nous avons présentés permet de voir les avantages et surtout les inconvénients des diverses solutions. Ainsi :

- Les serveurs thématiques n'ont pas accès à l'ensemble de l'information, mais recourent une notion de communauté scientifique.
- Les serveurs locaux sont souvent développés de façon indépendante et non compatible entre eux, ce qui pose des problèmes pour une mise en commun de leur base de données. En revanche ils ont souvent une bonne réactivité.

La multiplication des sites de thèses (soit thématiques, soit locaux) montre de façon évidente le manque d'informations sur les thèses et rend urgente l'amélioration des services nationaux chargés de la diffusion de ces informations. Il est à noter que la situation est la même, si ce n'est pire, dans les autres pays européens (cf le No 11 d'*ARABESQUE* - article sur les serveurs de thèses et la situation en Allemagne).

- Concernant un tel service national, la méthode envisagée par l'ABES nécessite une gestion centralisée et ne permettra sans doute pas de réduire de façon suffisante le délai de mise à disposition des informations sur les thèses. La lourdeur du dispositif nous semble un peu trop importante. Mais l'avantage est celui d'un fond global.

Le défi principal consiste donc à concilier rapidité de saisie des données et exhaustivité.

Concernant le contenu, il importe de distinguer la question de la mise en ligne d'un catalogue¹ et celle des thèses elles-mêmes. En effet, le problème de la multiplicité des formats de documents électroniques, et la question de leur pérennité sur de longues périodes ne doit pas être prise à la légère. Une attitude pragmatique consiste donc à mettre rapidement en place un catalogue le plus complet possible des thèses. Il pourrait mentionner en plus l'existence ou non d'un document électronique, et dans l'affirmative son format². Le document électronique, lorsqu'il existe pourrait être pointé par un lien hypertexte. Dans un second temps, la mise en ligne des thèses à divers formats et le développement d'outils de conversion pourrait alors être abordée.

1. Comprenant les données institutionnelles relatives à la thèse (nom et prénom du doctorant, université et discipline de rattachement, composition du jury et date de soutenance mais aussi les résumés en français et en anglais ainsi que les listes de mots clefs.

2. Format mais aussi version de logiciel et de système d'exploitation l'ayant produit.

Toutefois, les travaux de thèse doivent être, nous l'avons dit, diffusés rapidement; ceci est d'ailleurs confirmé par les études de l'ABES qui montrent que les thèses sont consultées essentiellement dans les trois années qui suivent la soutenance (cf. ARABESQUE - No 11). Dans un premier temps, et avant le développement ou l'adaptation³ d'outils de conversion interformats, le choix du format dans lequel doit être stocké le document doit être guidé principalement par les formats les plus utilisés au moment de la soutenance (actuellement les formats postscript et PDF semblent bien adaptés).

L'archivage à long terme des manuscrits doit être conservé (sous forme papier et microfiche) même si les thèses restent rarement des documents de référence (à moins qu'elles ne conduisent à la publication d'un livre).

3 Les solutions proposées

Le plus important est donc de permettre de retrouver très rapidement les informations principales sur la thèse (celles qui sont contenues dans le formulaire d'enregistrement de thèse soutenue). Il faudrait, pour réduire les délais de mise à disposition de l'information, **permettre aux doctorants de remplir le formulaire d'enregistrement de thèse soutenue directement par voie électronique**. Ceci permettrait en outre d'alléger la charge de travail des organismes chargés de cette diffusion.

Ces données devraient ensuite être validées par le service compétent de l'établissement délivrant la thèse et indexées ensuite dans la base de données. Cette méthode présente deux avantages :

- Elle élimine les erreurs dues à la double saisie.
Actuellement, les secrétariats doivent retaper les informations fournies par le jeune docteur. Ils doivent les dactylographier sur le formulaire d'enregistrement (ce qui nécessite une machine à écrire ; objet qui se fait rare dans certains laboratoires) alors que ces données sont le plus souvent déjà dans le fichier utilisé pour écrire la thèse.
- Les délais de mise à disposition de l'information seraient notablement réduits.

Nous pensons donc que la solution réside dans la mise en place d'un serveur à consultation centralisée mais à gestion décentralisée. Une telle initiative existe. Il s'agit de celle de REDOC pour les universités de Grenoble, qui devrait permettre au jeune docteur d'inscrire lui-même sa thèse en ligne et s'il le souhaite de diffuser le document complet de sa thèse. Cette possibilité devrait être disponible sur un serveur institutionnel des thèses. Concrètement cela passerait soit par l'extention de l'initiative REDOC à l'échelle nationale, soit par une collaboration forte entre REDOC et l'ABES. En clair, il importe de choisir une entité qui sera maître d'œuvre dans la réalisation du projet. Elle aurait une large liberté d'action et aurait pour mission de faire aboutir le projet, au moins dans une version minimale, en un temps raisonnable (de l'ordre de l'année).

L'interface avec les universités nécessite une consultation de celles-ci (par le biais de la Conférence des Présidents d'Université) comme c'est déjà le cas dans le cadre du projet REDOC afin de préciser les moyens permettant aux doctorants de déclarer leur thèse et aux universités de valider ces données. Cette consultation devrait déboucher en quelques mois sur un cahier des charges simple précisant quelles sont les informations nécessaires au catalogue et précisant la procédure de validation électronique (à travers Internet) des données par les services administratifs des universités. L'implémentation pratique serait à charge de l'entité chargée de la réalisation du projet.

3. De nombreux logiciels libres permettent déjà d'effectuer un grand nombre de conversions plus ou moins facilement.

Enfin, l'indexation (classification) sera effectuée à l'échelle nationale mais les docteurs pourront le cas échéant proposer des modifications lorsque cette classification ne reflète pas la spécialité dans laquelle ils ont effectué leur doctorat.

Rapidement, l'ensemble des bases de données de thèses existantes pourraient être remplacées par ce système, qui permet une saisie délocalisée mais assurerait la disponibilité globale des informations. Un premier pas pour constituer le fond de départ serait **la diffusion gratuite de l'intégralité des données de la base PASCAL**. En clair, nous demandons donc aux services concernés (l'ABES pour le Ministère de l'Education Nationale de la Recherche et de la Technologie, et l'INIST pour le CNRS) de **rendre publiques et gratuites ces informations qui existent déjà sous forme électronique payante** (minitel, CDROM...).