

● ● ● LES DONNÉES DE LA RECHERCHE DES THÈSES

Avec le développement de l'eScience, l'accès aux données de la recherche devient un enjeu pour la veille scientifique. Tandis que les projets d'infrastructure portent avant tout sur de grands réservoirs de données (*big data*), d'autres résultats (*small* ou *smart data*) restent moins accessibles. Certaines données sont déposées avec les thèses électroniques. Leur intérêt est double : d'une part, leur représentativité, richesse et qualité et, d'autre part, la possibilité de les réutiliser avec d'autres résultats. Il s'agit d'annexes, d'enregistrements sonores, de matériel audiovisuel, tableaux, bases de données, résultats d'enquêtes bruts, etc. Stockées auparavant comme matériel complémentaire dans les bibliothèques, ces données sont, avec la mise en place des dispositifs comme Star, désormais déposées en format numérique. Comment ouvrir ces données, les signaler dans l'environnement des archives institutionnelles, les rendre réutilisables et exploitables, notamment pour la veille scientifique ? On peut identifier trois verrous majeurs pour la mise en œuvre de cette accessibilité.

Barrières technologiques

Souvent, les formats sont peu adaptés pour une réutilisation et/ou les métadonnées sont absentes. Les métadonnées sont pourtant essentielles pour l'interprétation, la préservation, le partage et la réutilisation des données. Un problème majeur réside dans l'absence d'un identifiant unique et pérenne comme le DOI (*Digital Object Identifier*). Quant aux formats, la Commission européenne (CE) a proposé de mettre à jour la directive de 2003 sur la réutilisation des données publiques, afin de rendre obligatoire la mise à disposition de ces données dans des formats courants, lisibles par des machines. Quelles sont les options identifiées dans le domaine des thèses électroniques ?

Obstacles juridiques

Par rapport à la diffusion et à la réutilisation des résultats de la recherche, le cadre légal des thèses paraît inadapté. Étendre la protection du droit d'auteur de la thèse électronique aux résultats (données) est en conflit avec la politique du libre accès à l'information scientifique et des données ouvertes. La CE et le gouvernement français font la promotion d'une diffusion des données publiques avec une licence ouverte minimaliste, là où les auteurs et les établissements adoptent souvent une stratégie plus restrictive (pas de modification, pas d'exploitation à

but lucratif, accès en intranet, etc.) qui entrave ou interdit toute réutilisation.

Barrières organisationnelles

Avant toute exploitation, le traitement des données nécessite un *workflow* centré sur les besoins de l'auteur-producteur et adapté aux spécificités des données (acquisition, suivi, sécurité, mise à disposition, etc.). De même, il faudra de nouveaux services et fonctionnalités pour ajouter de la valeur aux *small data* et pour faciliter leur découverte et réutilisation. À ce jour, peu d'archives institutionnelles ont commencé à développer ce genre d'outils, à l'instar des archives de données.

Vers un programme de recherche

Les verrous évoqués ci-dessus sont souvent liés. Ensemble, ils représentent des obstacles à la réutilisation des données scientifiques déposées avec les thèses. Nous avons donc mis en place un projet de recherche européen (@pic-etd : *Academic Publishing in Change-Electronic Theses and Dissertations*) pour étudier ces questions et pour instruire à la fois de bonnes pratiques (recommandations) et un cahier des charges pour faciliter l'accès aux données en fonction des disciplines, des solutions techniques et procédures existantes. Cette recherche fait partie d'un programme plus large qui inclut deux autres projets, l'un sur la publication des thèses électroniques (avec l'université de Ljubljana, Slovénie), l'autre sur leurs restrictions d'accès (avec l'Institute for Science Networking de l'université d'Oldenburg, Allemagne).

Le projet @pic-etd, retenu sur la liste complémentaire du programme blanc 2013 de l'Agence nationale de la recherche (ANR), a été redéposé dans le cadre du programme 2014. Les résultats sont attendus pour début 2014. L'Abes sera associée à la réalisation de ce projet programmé dans le cadre d'un consortium composé de :

- laboratoires de recherche : Geriico (Lille), Elico (Lyon), Dicen-IDF (Paris) ;
- services, réseaux : Abes, CCSD, ANRT, British Library, TIB Hannover, DataCite, Dini, ProQuest ;
- experts : M. Foulonneau (H. Tudor, Luxembourg), C. Leduc (ISCID-CO), L. Maurel (BDIC, Paris).

JOACHIM SCHÖPFEL

Directeur de l'Atelier national de reproduction des thèses
joachim.schopfel@univ-lille3.fr