Ouvrir les données de la recherche pour la veille scientifique

Le cas des thèses électroniques

Bernard JACQUEMIN (*), Hélène PROST (*)(**), Joachim SCHÖPFEL (*), Marta SEVERO (*), Florence THIAULT (*)
bernard.jacquemin@univ-lille3.fr, helene.prost@inist.fr, joachim.schopfel@univ-lille3.fr, marta.severo@univ-lille3.fr, florence.thiebault@univ-lille3.fr

(*) GERiiCO, Université de Lille 3, BP 60149, 59653 Villeneuve d'Ascq Cedex (France)
(**) INIST-CNRS, 2, allée du parc de Brabois, CS 10310, 54519 Vandoeuvre-lès-Nancy (France)

Mots clefs :
Veille scientifique et technologique, information scientifique, données de la recherche, thèses électroniques, archives ouvertes, libre accès
Keywords:
Scientific and technical observation, scientific information, research data, electronic theses, open archives, open access
Palabras clave:
Escudriñar científico y tecnológico, información científica, datos de investigación, tesis electrónicas, archivos abiertos, libre acceso

Résumé

Avec le développement de l'*eScience*, l'accès aux données de la recherche devient un enjeu important pour les dispositifs et stratégies de la veille scientifique. Tandis que les projets d'infrastructure portent avant tout sur de grands réservoirs de données (*big data*), d'autres résultats de la recherche restent moins accessibles et sont peu ou pas exploitables par une veille scientifique. Il s'agit en particulier de petits ensembles de données (*small data*), produits et mis en ligne par des équipes ou chercheurs individuels, ou de données publiées avec ou dans des rapports, articles, *working papers*, communications, thèses, mémoires etc.
Dans la mesure où les thèses représentent un intérêt certain pour la veille scientifique, du fait de leur nombre, leur richesse et qualité mais aussi leur représentativité, nous nous sommes intéressés aux données de la recherche diffusées avec les thèses. Il s'agit d'annexes, d'enregistrements sonores, du matériel audio-visuel, de tableaux, bases de données, résultats bruts d'enquêtes etc. Hier déposé comme matériel complémentaire avec les thèses dans les bibliothèques universitaires, aujourd'hui, avec la mise en place des dispositifs de thèses électroniques, ce matériel est déposé et diffusé sur des serveurs et plateformes documentaires.
Source potentiellement riche d'information dans tous les domaines scientifiques, ce matériel paraît peu exploité à ce jour. Notre travail consiste à faire le point sur ce matériel et son intérêt pour la veille, et à dresser un premier état des questions et problèmes à soulever pour ouvrir ces *small data* aux dispositifs et stratégies de la veille scientifique.

1 Introduction

« La veille stratégique scientifique et technique est une discipline dont l'objectif principal est d'aider les chercheurs et les industriels à rester informés de ce qui se passe dans leurs domaines » [12]. Plus particulièrement, la veille scientifique assure un « suivi dans le domaine de la recherche, des productions scientifiques et de leur évolution, de sources diversifiées » [6], et recouvre « l'ensemble des actions coordonnées d'acquisition, d'analyse, de conservation et de diffusion de l'information de nature scientifique et technique en vue de son exploitation par les acteurs de l'organisation » [1]. Or, les nouvelles technologies de l'information et de la communication ont radicalement modifié la production scientifique. D'une part, de plus en plus d'information scientifique est librement accessible sur Internet, en particulier dans des archives ouvertes [35]. D'autre part, la nature même de l'information scientifique change, et aux publications traditionnelles s'ajoute ce que Peter Suber appelle simplement « research » – algorithmes, méthodes, données, hypothèses etc.

Avec le développement de l'*eScience*, l'accès aux données de la recherche devient un enjeu important pour les dispositifs et stratégies de la veille scientifique. Tandis que les projets d'infrastructure portent avant tout sur de grands réservoirs de données (*big data*), d'autres résultats de la recherche restent moins accessibles et sont peu ou pas exploitables par une veille scientifique. Il s'agit en particulier de petits ensembles de données identifiées sous le terme *small data*, ou parfois aussi *smart data* ; ces données sont produites et mises en ligne par des équipes ou chercheurs individuels, ou bien elles sont publiées conjointement ou directement dans des rapports, articles, *working papers*, communications, thèses, mémoires etc. Leur intérêt : permettre des comparaisons (et intégrations) avec d'autres données, l'identification de nouvelles tendances et de thèmes émergents, la détection de signaux faibles sans passer par l'interprétation des documents, publications et sources secondaires.

Dans la mesure où les thèses représentent un intérêt certain pour la veille scientifique, du fait de leur nombre, leur richesse et qualité mais aussi leur représentativité, nous nous sommes intéressés aux résultats de la recherche diffusées avec les thèses. Il s'agit d'annexes, d'enregistrements sonores, du matériel audio-visuel, de tableaux, bases de données, résultats d'enquêtes bruts etc. Hier déposé comme matériel complémentaire avec les thèses dans les bibliothèques universitaires, aujourd'hui, avec la mise en place des dispositifs de thèses électroniques, ce matériel est déposé et diffusé sur des serveurs et plates-formes documentaires. Mais ces serveurs documentaires ne sont pas nécessairement l'endroit idéal pour ces fichiers, d'autant plus que, souvent, ils ne sont pas ou peu « lisibles » par les machines [35].

Source potentiellement riche d'information dans tous les domaines scientifiques, ce matériel est peu exploité à ce jour. Notre travail consiste à faire le point sur ce matériel et son intérêt pour la veille, à dresser un premier état des questions et problèmes face à l'ouverture de ces *small data* aux dispositifs et stratégies de la veille scientifique. Il s'agit donc d'un travail sur le *sourcing*, ou plutôt en amont du *sourcing*, afin de rendre cette partie de la production scientifique accessible à la veille scientifique.

2 Contexte

Il y a peu de temps encore, la profondeur du *sourcing* d'un dispositif de veille s'arrêtait généralement au niveau du document. Même si Dousset et al. [15] avaient très tôt déjà souligné le potentiel d'Internet pour la veille, du fait d'un « accroissement gigantesque dans la production et l'accessibilité aux données », Chazelas et al. [8] considèrent la veille documentaire, c'est-à-dire le suivi des sources avec signalement des documents, toujours comme « premier degré » d'une veille dans le contexte de la recherche et à la base de toute veille scientifique (« analyse du contenu avec une sélection et une mise en perspective des documents »). Le concept de l'information brute, « matière première utilisée dans ce processus » [12], renvoie ici le plus souvent aux notices bibliographiques

des SGBD ou SIGB, aux éléments bien structurés et indexés des bases de données, ou simplement à du texte libre (hypertexte), sans établir le lien entre *datamining* et *datasets* au sens strict (données de la recherche, observations, mesures, statistiques, enquêtes, enregistrements etc.).

En fait, des milliers d'articles ont déjà été publiés sur l'acquisition et le traitement des données de la recherche, y compris dans une vingtaine de titres spécialisés (*Journal of Chemical and Engineering Data*, *Data Science Journal* etc.), en particulier en sciences de l'ingénieur, chimie et physique. Mais le rapprochement des données de la recherche et leur diffusion libre avec la veille est récent et à ce jour, relativement peu exploré. En plus, il s'agit en partie non pas de données scientifiques mais d'*open data* en général, pour créer des conditions favorables à leur réutilisation et exploitation y compris par des dispositifs de veille (cf. par exemple [22], [23], [25], [34]).

L'un des premiers à faire le lien entre les thèses, leurs données et la veille, sous l'aspect technologique mais aussi organisationnel et juridique, a été P. Murray-Rust [24] : « A PhD thesis represents 3 or more years' work and much of the text is actually detailed accounts of scientific experiments, facts, recipes, methodology. I shall refer to this as 'data' and argue that it is factual information which, when published, belongs to the scientific commons ». Pour ouvrir ces data à l'exploitation par des dispositifs automatisés, Murray-Rust demandait des thèses en XML, des métadonnées normalisées, l'acquisition et le dépôt des données à la source, et leur protection par des licences adaptées (*Creative Commons* etc.).

Par la suite, Sefton et al. [32] ont développé un système de gestion (*ICE-TheOREM*) pour créer, déposer et archiver des données de la recherche dans les archives ouvertes. « The *ICE* system manages both small data files and links to larger data sets. The result is research publications which are available not just as paper-ready PDF files but as fully interactive semantically aware web documents which can be disseminated via repository software such as ePrints, DSpace and Fedora as complete supported web-native and PDF publications ». Une autre solution (*GREET*) a été présentée par Por et al. [28], une *grid architecture*, qui supporte l'intégration des données de la recherche avec les thèses.

D'une manière plus générale, d'autres études soulignent l'intérêt d'une couche sémantique (*linked data*) pour l'exploitation automatique des données déposées dans une archive ouverte (cf. par exemple [17] ou [19]). Comment valider le dépôt d'un ensemble de données, comment gérer leur protection juridique – ce sont des questions récurrentes abordées par exemple par Downing et al. [14] dans le domaine de la chimie. Leur étude est un plaidoyer en faveur de formats standards, XML, métadonnées spécifiques, identifiants pérennes et une gestion des demandes d'embargo. Dans un autre domaine, la physique des particules, Praczyk et al. [29] arrivent à la même conclusion : l'importance des métadonnées et identifiants uniques pour faire le lien entre publication et données de la recherche, et pour la diffusion et la préservation de ces dernières. Un peu en marge de notre sujet mais intéressant pour la veille scientifique des *datasets*, Piwowar et Chapman [27] décrivent une méthode de traitement automatique de langage naturel pour identifier des données partagées mentionnées et/ou publiées dans des articles du domaine biomédical.

Cependant, ces études concernent surtout ce qu'on appelle les *big data*, les grands ensembles de données produits par le CERN, par les grands laboratoires en chimie ou biologie etc. Quid des résultats produits par les structures moins importantes, les équipes de recherche ou des chercheurs individuels ? Quid des données de la « *little science* » [4], de la « *dark science* » [19], de la « *small science* » [11], des « *small research collections* » [5] ? Borgman et ses collègues du *Center for Embedded Networked Sensing* (CENS) de l'université de Californie exposent, dans le domaine de l'écologie et de la recherche sur le terrain, la particularité de ces données : leur diversité, leur mode de production, de validation (contrôle), d'utilisation, ainsi que leur rôle dans le processus de la recherche. Ils soulignent l'intérêt d'une standardisation et d'une collaboration étroite avec l'informatique. Cragin et al. [11] prévoient pour cette raison – « the high level of variation and complexity in data forms » – mais aussi du fait de la grande diversité des pratiques de partage, une forte demande d'assistance et de traitement (*curation services*) pour l'acquisition et la diffusion de ces *small datasets* via les archives ouvertes, ceci aussi en fonction des particularités (formats, types de données, usage...) de chaque discipline.

L'étude la plus intéressante pour notre projet est sans doute celle de Collie et Witt [9] qui modélisent l'intégration des données de la recherche dans le flux de traitement des thèses électroniques au sein des archives institutionnelles (*data augmented approach*). Même s'ils ne parlent pas directement de la veille, il est évident que leur proposition d'une *data curation* systématique des thèses au sein des universités, par les écoles doctorales, bibliothèques universitaires et équipes informatiques, rejoint les besoins du *sourcing* de la veille scientifique.

3 Enjeux

Comment ouvrir les données déposées avec les thèses électroniques, comment les rendre exploitables pour la veille scientifique ? Notre travail consiste à identifier les problèmes afin de pouvoir proposer des solutions. Nous avons déjà souligné la différence avec d'autres ensembles de données, notamment avec les *big data* de l'*eScience*, puisqu'elles ne sont pas produites massivement ou dans des formats standards. Leur intérêt est lié à la richesse, la variété et l'originalité de ces ressources, y compris leur qualité dans la mesure où leur production a été supervisée, évaluée et validée par des experts scientifiques. Le plus souvent, une thèse est liée à un projet ou programme de recherche, et correspond à un axe scientifique de laboratoire. Un autre intérêt est la nature non commerciale et publique de la plupart de ces résultats.

D'après notre état de l'art, les problèmes d'accès, d'exploitation et de réutilisation relèvent de trois aspects:

Barrières technologiques : formats non adaptés et/ou absence de métadonnées. Les métadonnées sont essentielles pour l'interprétation, la préservation, le partage et la réutilisation des données [40]. Sans métadonnées adaptées, pas de portail web, d'outil de recherche efficace ni d'environnement innovant du type *discovery tools*. Un problème particulier est l'absence d'attribution d'un identifiant unique et pérenne aux données archivées. Le projet international *DataCite* propose l'attribution systématique d'un *Digital Object Identifier* (DOI). Cette solution vient d'être adoptée par l'*International Standard Randomised Controlled Trial Number Register* pour leur base de données *Current Controlled Trials* dans le but de simplifier le lien entre données et publications et de faciliter l'accès du public aux résultats de la recherche médicale. Quant aux formats, la Commission Européenne a récemment 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, pour faciliter leur réutilisation. Toutes ces initiatives et projets ont tendance à diminuer ces barrières technologiques. Quelles sont les options dans le domaine des thèses électroniques ?

Obstacles juridiques : licences non adaptées et/ou (sur)protection par le droit d'auteur. Trois parties sont considérées comme des détenteurs de droits d'une thèse :

- l'auteur qui détient tous les droits moraux et patrimoniaux et qui peut autoriser ou interdire la numérisation ou la diffusion,
- l'institution qui peut, au moins pour une période définie, limiter la diffusion de la thèse (confidentialité), interdire la reproduction ou demander des modifications,
- un tiers qui, dans certains cas spécifiques, peut faire valoir des droits d'auteur ou autres liés à la thèse.

Le format numérique permet la gestion de versions différentes, avec des droits différents. Par exemple, la version complète d'une thèse peut inclure des photos protégées par un tiers sans l'autorisation de reproduction, de diffusion, etc. Une autre version sans ces photos peut avoir une clause de confidentialité temporaire mais aucune autre restriction en ce qui concerne la reproduction, le téléchargement etc. Une version auteur non validée peut être disponible sur un serveur en libre accès (archive ouverte). Un projet d'édition conventionnelle peut inciter les auteurs à ne pas permettre la diffusion sur Internet, en particulier (mais pas seulement) dans les sciences humaines et sociales [31]. Tous ces problèmes concernent en premier lieu le document de la thèse en tant qu'œuvre de l'esprit et création intellectuelle. Inclure les résultats de la recherche (données) en respectant la protection du droit d'auteur de la thèse électronique est en conflit avec la politique du libre accès à l'information scientifique et des données ouvertes. Tandis que la Commission Européenne et le gouvernement

français font la promotion d'une diffusion des données publiques avec une licence ouverte minimaliste (open licence, équivalent à CC-BY), les auteurs et établissements adoptent souvent une stratégie de diffusion plus restrictive (pas de modification, pas d'exploitation à but lucratif, accès en Intranet etc.). Tout cela est trop restrictif pour réaliser le potentiel de réutilisation de ces données.

Barrières organisationnelles : *workflows* et/ou services non adaptés ou manquants. Avant toute exploitation, l'intégration et le traitement des données nécessitent une gestion et un *workflow* spécifique. Cet environnement concerne des aspects tels que le suivi, la sécurité, l'acquisition, la mise à disposition etc. [13]. L'absence d'aide et de soutien technique figurent parmi des raisons pour lesquelles les chercheurs ne déposent pas leurs données dans des archives ouverts [10]. Plus généralement, pour la diffusion des thèses électroniques en libre accès, si les établissements hébergeurs ou prestataires de services veulent accroître leur utilisation et impact, ils ont intérêt à adopter un processus de développement centré sur les besoins de l'utilisateur [16]. Dans son étude sur les archives ouvertes, Bester [2] distingue plusieurs options de navigation, de recherche, de personnalisation et de gestion de références bibliographiques. Dans une publication récente, nous avons identifié d'autres initiatives intéressantes qui permettent d'ajouter de la valeur aux thèses numériques dans des archives ouvertes : citons par exemples les outils sociaux, les *discovery tools* sophistiqués, les statistiques d'utilisation, les vidéos, l'impression à la demande, les licences *Creative Commons*, la conservation en plusieurs exemplaires etc. [30]. Certains de ces développements demandent un investissement conséquent, tandis que d'autres sont possibles sans déploiement de ressources importantes [16]. Sur un plan plus général, les services pour la diffusion des thèses électroniques doivent être souples, dotés d'une capacité d'adaptation rapide ; le logiciel devra être convivial et fiable, peut-être aussi ouvert à d'autres fournisseurs de services et/ou intégré dans un autre environnement de service, comme par exemple la formation à distance.

Ces aspects technologiques, juridiques et organisationnels sont souvent liés. Ensemble, ils représentent des obstacles évidents à l'exploitation et la réutilisation des données scientifiques déposées avec les thèses électroniques. Par la suite, nous allons décrire comment un programme scientifique pourrait répondre à ces questions. Ce faisant, nous nous gardons bien entendu de proposer une seule et unique solution. Il s'agit plutôt d'indiquer comment faciliter, dans des environnements spécifiques, l'accès aux données en fonction des disciplines, des solutions techniques et procédures existantes.

3.1 Distribution

Il n'existe à ce jour aucune information fiable sur la répartition des données complémentaires aux thèses parmi les différentes disciplines scientifiques, ni sur leur format, ni sur leur contenu ou évolution. Bien que l'état de l'art nous révèle des disparités de services ou de qualité de métadonnées parmi les référentiels de discipline différente, nous ne disposons pas actuellement d'éléments chiffrés sur la répartition disciplinaire de ces données.

Il faut donc, dans un premier temps, mener une enquête afin d'obtenir une idée réaliste sur ce matériel.

Une première analyse bibliométrique devra étudier la distribution du matériel complémentaire déposé avec les thèses dans un échantillon large et représentatif de thèses. Cet échantillon pourra être constitué de plusieurs grands catalogues et bases de données avec des thèses, tels que le *SUDOC* et le portail des thèses de l'ABES, l'archive de microfiches de thèses de l'ANRT, le catalogue des thèses numériques anglaises maintenu par la British Library (*EThOS*), la base de thèses de ProQuest (*ProQuest – Dissertation Publishing*) ou encore le catalogue de bibliothèque nationale technique et scientifique allemande à Hanovre (TIB). Il s'agira d'identifier tout d'abord les métadonnées ou champs de données qui renseignent sur l'existence de contenus supplémentaires, puis d'uniformiser l'intitulé des différents champs d'information afin de pouvoir analyser le contenu de chaque corpus en fonction du domaine scientifique, support, contenu ou année de publication. En fonction de la qualité des données, on pourra compléter une telle approche bibliométrique par une analyse de type *text mining* pour une étude plus fine de la distribution et de l'évolution dans le temps.

Dans un deuxième temps, il faudra conduire une enquête sur le traitement et la conservation de ces documents supplémentaires auprès des bibliothèques universitaires et écoles doctorales, en gardant les mêmes critères d'analyse que ceux choisis pour l'étude bibliométrique puis en ouvrant l'enquête à d'autres

domaines, en particulier sur les étapes de traitement, les bonnes pratiques ou la préservation des métadonnées (cf. la section suivante). Il s'agira de produire une sorte de *baseline* avec en même temps une image précise et fidèle de ce matériel, qui permettra à la veille scientifique d'évaluer l'intérêt de cette source en fonction de la demande et de définir le dispositif à mettre en œuvre.

3.2 Métadonnées

La veille scientifique a besoin d'une description précise, fiable et détaillée des ressources en question. Dans le cas des thèses électroniques, il s'agit d'assurer une description adaptée des données, fichiers et autre résultats déposés avec les thèses électroniques. À ce jour, cette description, si elle existe, reste souvent rudimentaire et liée à la thèse, sans permettre de faire le lien avec d'autres données.

Une solution technique capable d'assurer une interopérabilité viable entre des sources informationnelles hétérogènes nécessite de définir un schéma de métadonnées basé sur une approche sémantique cohérente à l'aide de nomenclatures, de vocabulaires contrôlés ou d'ontologies [7]. Ce travail s'effectuera en trois phases.

La première phase consiste à analyser de manière approfondie les différents jeux de métadonnées utilisés pour décrire les thèses et mémoires électroniques : *Dublin Core* [26], TEI, *Metadata Core Set* (description des thèses et mémoires en Grande Bretagne), recommandation TEF (Thèse Électroniques Françaises). Il s'agit également d'analyser les choix effectués jusqu'à présent par les grands acteurs de ce domaine (British Library, ABES, CCSD, ProQuest), et en particulier d'étudier les options envisageables pour relier les documents décrits aux données de la recherche qui y sont attachées.

La seconde phase consiste à analyser les jeux de métadonnées destinés à la description des données de la recherche et des autres types de documents. Ce travail est effectué d'une part à partir des données collectées par le consortium EUDAT, qui se penche également sur les données de la recherche [39], et de l'autre sur l'expérience des fournisseurs institutionnels d'entrepôts de thèses, qui sont déjà confrontés aux données de la recherche. Nous nous intéressons bien entendu plus particulièrement aux métadonnées qui favorisent l'interconnexion entre les données hétérogènes et le texte intégral des thèses électroniques.

La troisième phase concerne l'attribution d'un identifiant unique non seulement aux données de la recherche, mais aussi aux thèses elles-mêmes, à leurs auteurs ainsi qu'aux institutions concernées. Ici, plusieurs initiatives, travaux de recherche et projets seront exploités, par exemple *DataCite*, le consortium DOI, le laboratoire DICEN-IDF, la Bibliothèque Scientifique Numérique et le Consortium interuniversitaire américain pour la recherche politique et sociale (ICPSR) qui travaille à Ann Arbor sur la mise en relation des données de la recherche avec les thèses recensées par ProQuest [37].

3.3 Dispositif et *workflow*

L'accessibilité des données pour la veille scientifique dépend également des dispositifs et procédures mis en place pour gérer les résultats de recherche, fichiers complémentaires et ensembles de données déposés avec les thèses électroniques dans les archives institutionnelles. Il est donc nécessaire d'étudier et d'évaluer les différents modes de collecte mis en place pour recueillir les thèses électroniques par les divers entrepôts institutionnels. L'objectif est de fournir soit des recommandations de bonnes pratiques sur les données de la recherche associées à des thèses, soit des critères permettant d'effectuer des choix concernant ces mêmes données, afin d'accroître leur disponibilité pour une exploitation et réutilisation dans le cadre d'une veille scientifique.

Une telle analyse s'appuiera nécessairement sur la modélisation de Collie et Witt [9], mais devra inclure les cinq principaux schémas de collecte élaborés par les grands acteurs du secteur : STAR pour le dépôt des thèses françaises auprès de l'ABES et du CINES, TEL pour les dépôts sur la plate-forme d'archives ouvertes HAL du CCSD, EThOS pour les thèses électroniques gérées par la British Library, les protocoles mis en place par la Technische Informations Bibliothek (TIB) de la Leibniz Universität Hannover, et le modèle exploité par ProQuest [37]. Cette analyse mettra en lumière le processus de collecte des

données associées aux thèses et produira pour chaque dispositif une représentation graphique du *workflow* sous la forme d'un diagramme de flux des données, de manière à faciliter la comparaison et l'évaluation.
La modélisation des différents schémas de *workflow* pour les thèses électroniques devra être complétée par l'étude d'autres dispositifs dans l'environnement émergent de l'*eScience*, sans lien direct avec les thèses, afin de rapprocher les bonnes pratiques des archives de données des archives institutionnelles et du traitement des thèses électroniques. Par exemple, il serait intéressant d'analyser et comparer les projets TARDis à l'Université de Southampton [33][18] et CAMERA (*Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis*) du California Institute for Telecommunications and Information Technology [36], par rapport à la collecte et l'intégration des données, l'automatisation de leur gestion, l'indexation au moment du dépôt etc.
Le cas du projet CAMERA est particulièrement intéressant. Sun et al. [36] ont développé une *gridsphere*, un portail avancé autour d'une base de données dans une infrastructure qui permet le dépôt, le partage, l'analyse, le téléchargement et la réutilisation des données afin de faciliter la collaboration dans le domaine du génome.
Sous l'aspect de la veille, ces études devront répondre en particulier à quatre questions :

- Le traitement de la thèse et des données associées doit-il être conjoint ou distinct ?
- Le dépôt de la thèse et des données doit-il se faire sur un même entrepôt ou sur des entrepôts séparés ?
- L'archivage à long terme des données doit-il être réalisé sur le même entrepôt que le dépôt ou sur un entrepôt distinct ?
- Comment relier la thèse et les différentes données de la recherche (indexation, identifiant etc.) ?

S'il n'y a pas forcément de solution idéale, mais plutôt des options pertinentes variées en fonction des situations techniques et/ou légales (archivage à long terme, matériels sous copyright etc.), il s'agira cependant de définir au moins des critères stables permettant d'opérer un choix optimal sur les données de recherche et les matériels associés.

3.4 Services

Une partie de notre projet porte sur l'évaluation de services à valeur ajoutée proposés par les archives institutionnelles dans une perspective de diffusion et de réutilisation des données de la recherche publique. La publication de recommandations dans ce domaine permettra d'améliorer les outils existants et de contribuer au développement de nouveaux produits ou services.
Par rapport à la veille scientifique, il s'agira en particulier d'établir l'offre de service adaptée aux besoins et fonctionnalités des dispositifs de veille, comme l'accès non protégé aux données, des flux RSS, une indexation fine etc.
Il existe plusieurs initiatives et réalisations qui, même s'ils ne concernent pas directement les thèses, pourront servir de modèle ou de référence. Ainsi, Yang et al. [40] décrivent les composantes fonctionnelles pour l'interface utilisateurs d'un système géospatial, comme par exemple le traitement de données multidimensionnel, le Web sémantique et le partage des connaissances, etc. Par rapport au *workflow* (cf. la section précédente), ils mentionnent des outils pour la collecte et l'intégration de données hétérogènes, la préservation des données et l'accessibilité.
Si les données sont stockées sous des formes réutilisables, elles peuvent être partagées sur des réseaux distribués et stimuler la recherche et d'autres formes d'exploitation. Pour cette raison et toujours dans le domaine de la géolocalisation, Wang et Liu [38] demandent la transformation des systèmes de gestion de données centralisées en systèmes de collaboration décentralisées (réseaux sociaux, technologies Web 2.0), afin de réduire les coûts et de proposer des interfaces personnalisables.
Lors de la première conférence EUDAT, Laure et Livenson [21] ont présenté un projet de service pour les *small data* appelé « *Simple Store* », basé sur le logiciel Invenio du CERN et conçu pour la longue traîne des *small data*. Avec *Simple Store*, le chercheur ou amateur citoyen est capable de créer et/ou

manipuler des données de la recherche avec des fonctionnalités du genre YouTube et DropBox, il peut les inclure dans une présentation ou les ouvrir au débat et aux commentaires.
La première étape du plan de travail est une revue des études publiées sur les services à valeur ajoutée de répertoires ouverts [2], et en particulier les dépôts institutionnels. Cet examen permettra de produire une première liste de services et de fonctionnalités nécessaires pour la recherche, la diffusion et la préservation du contenu déposé. L'étude sur les référentiels sélectionnés permettra d'enrichir l'examen des informations détaillées sur les avantages, les spécificités et les inconvénients de ces fonctionnalités. Dans cette partie du travail, les conseils des fournisseurs de services dans le domaine des thèses – ABES, le CCSD, ProQuest et d'autres – seront essentiels. La troisième étape sera essentiellement empirique et s'appuiera sur une enquête sur les dépôts et les initiatives de données sélectionnés, tels que le projet de EUDAT « *Simple Store* » déjà évoqué [21] et l'analyse des services basés sur des ontologies biomédicales [3]. L'objectif sera d'identifier les services et les fonctionnalités appropriés et mis en œuvre pour la récupération et surtout la réutilisation des données scientifiques dans des répertoires ouverts. Les deux listes de services et de fonctionnalités accompagnés des caractéristiques importantes, d'exemples, avantages et limites identifiés seront validées par rapport aux critères de qualité tels que les recommandations DINI. Enfin, nous proposerons également des spécifications pour les services et les fonctionnalités qui permettent de relier des documents et des données, afin de faciliter l'utilisation et la réutilisation autonome de chaque élément, y compris la récolte par d'autres prestataires de services. Les données déposées sous forme de fichiers supplémentaires doivent être réutilisables indépendamment de la thèse connexe. Une attention particulière devra être accordée aux métadonnées des deux documents et des données, dans le contexte de l'OAI-PMH et du TEI.
Parmi les autres pistes d'études, nous prendrons en compte les services basés sur les principes de médias sociaux, le partage et la collaboration, tels que l'évaluation sociale et le *tagging*.

3.5 Aspects juridiques

Le régime juridique des données et d'autres fichiers déposés avec les thèses n'est pas le même que pour des thèses elles-mêmes. Les deux doivent être considérés indépendamment, même s'ils sont liés. Par exemple, alors que d'une manière générale, dans le contexte des *open data*, les données scientifiques pourront être diffusées avec peu de restrictions, le plus souvent ces données sont diffusées au mieux de la même façon que les thèses ; au pire, elles ne sont pas accessibles du tout.
Dans la perspective d'ouvrir ces données aux dispositifs de la veille scientifique, il convient donc d'évaluer leur nature et statut juridique, en faisant le lien avec l'analyse des différents types et formats (cf. plus haut). L'évaluation s'appuiera d'abord sur des catégories telles que suggérées par Murray-Rust [24] pour les données ouvertes :

- données scientifiques appartenant aux biens communs (par exemple, le génome humain),
- données d'infrastructure essentielles à l'activité scientifique (par exemple dans les systèmes d'information géographique),
- données publiées dans des articles scientifiques qui sont factuelles et sans protection par le droit d'auteur,
- données par opposition aux logiciels et donc pas couverts par des licences Open Source et potentiellement susceptibles d'être détournés,
- cartes et autres objets nécessaires à l'infrastructure communautaire.

Ensuite, il faudra évaluer le dépôt, la conservation et la diffusion des données, comme un cas général de fichiers supplémentaires déposés avec le texte intégral des thèses, ceci sous l'aspect de leur accessibilité et réutilisation. Cet examen doit tenir compte de plusieurs aspects, tels que la protection sui generis des bases de données, la question des données personnelles, la confidentialité des informations sensibles ou stratégiques (y compris le secret professionnel), la restriction de réutilisation, de l'éthique (par exemple d'éventuels conflits d'intérêt) ou de plagiat.

L’évaluation devrait aboutir à des recommandations sur les conditions juridiques du dépôt, de la préservation et notamment de la diffusion de ces fichiers, en s'appuyant entre autre sur la politique française en matière d’*open data* et le choix d'une licence ouverte minimale (CC-BY) pour la mise à disposition des données numériques produites par les services et organismes publics. Contrairement aux thèses électroniques, les données ne devront pas seulement être ouvertes et « gratis » (au sens du mouvement du libre accès), mais aussi « libres », c'est-à-dire réutilisables.

4 Conclusion

Dousset et al. [15] ont décrit la veille sur Internet en huit étapes, à partir de la délimitation du sujet et des objectifs de l’analyse jusqu'à l'interprétation et la restitution des résultats. La récupération des données brutes, c'est-à-dire le choix des sources, les stratégies d'interrogation et la préparation des données, est d'une importance cruciale pour le processus du *datamining* « car de ses résultats dépendent la véracité et la validité de la généralisation des connaissances extraites ». Or, la qualité du pré-traitement des données brutes, de leur nettoyage, intégration et reformatage, n'est pas uniquement tributaire du dispositif de veille mis en œuvre, mais aussi des sources elles-mêmes. Notre contribution poursuit un double objectif : décrire une source peu ou pas exploitée, et proposer une approche afin de rendre cette source accessible et exploitable pour la veille scientifique.

La source en question se trouve à l'interface entre les thèses numériques, les données de la recherche et les archives institutionnelles. Autrement dit, notre proposition concerne une source numérique en libre accès sur Internet, faisant partie de la cyberinfrastructure scientifique (ou *eScience*). L'intérêt de la source est sa représentativité (tous les domaines scientifiques), sa qualité (validation par jury et établissement) et sa richesse thématique. Son problème est une méconnaissance, une accessibilité limitée et une grande variété de formats et sites (*small data*).

Notre proposition est de mieux connaître ces sources à travers une double enquête représentative et d'augmenter leur visibilité et accessibilité par plusieurs moyens, par rapport aux métadonnées, conditions de diffusion, services et procédures de gestion (*workflow*). Ainsi, pour faciliter la veille il faudrait probablement adopter une licence de diffusion avec un minimum de contraintes pour l'exploitation (*licence open data*, CC0 ou CC-BY) et pas seulement permettre mais exiger un dépôt des données et autres matériels avec les thèses mais dans d'autres conditions, avec un autre dispositif et d'autres procédures. Autrement dit, comme pour les articles de revues il faudrait rapprocher les archives institutionnelles actuelles des *data repositories*, ou bien créer cette fonctionnalité comme option au sein des dispositifs existants.

La qualité de veille dépend (aussi) de la qualité des sources. Voici, dans l'environnement du *open access* et de l'*eScience*, une source intéressante mais peu exploitée qui fait partie de la nébuleuse des données de la recherche publique. L'enjeu aujourd'hui est de trouver des moyens pour rendre ces données réellement publiques, librement, sans restriction d'accès. Notre contribution essaie d'en faire profiter la veille scientifique, en amont du *sourcing*.

5 Bibliographie

[1] BASSET H. (2013). *`De la veille à l'intelligence scientifique : définitions et concepts de base'*. In H. Basset (ed.), Maîtriser la veille pour l'intelligence scientifique, pp. 10-13. Techniques de l'Ingénieur, Paris.

[2] BESTER E. (2010). *`Les services pour les archives ouverte : de la référence à l'expertise'*. Documentaliste - Sciences de l'Information 47(4):4-15.

[3] BLAKE J. A. et BULT C. J. (2006). *`Beyond the data deluge: Data integration and bio-ontologies'*. Journal of Biomedical Informatics 39:314-320.

[4] BORGMAN C., et al. (2007). *`Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries'*. International Journal on Digital Libraries 7(1-2):17-30.

[5] BORGMAN C., et al. (2012). `*Who's Got the Data? Interdependencies in Science and Technology Collaborations'*. Computer Supported Cooperative Work (CSCW) 21(6):485-523.

[6] BRACHET-DUCOS C. (2007). `*Quel apport des professionnels de l'Information Scientifique et Technique dans le dispositif de veille d'un organisme de recherche*?'. Mémoire INTD CNAM, Paris.

[7] BULT C. J. (2002). `Data *integration standards in model organisms: from genotype tp phenotype in the laboratory mouse'*. TARGETS 1(5):163-168.

[8] CHAZELAS M., et al. (2006). `*Les rencontres 2006 des professionnels de l'IST. Les archives ouvertes et la veille scientifique, deux axes de réflexion'*. Documentaliste-Sciences de l'Information 43(3):232-241.

[9] COLLIE W. A. et WITT M. (2011). `*A Practice and Value Proposal for Doctoral Dissertation Data Curation'*. International Journal of Digital Curation 6(2):165-175.

[10] COSTELLO M. J. (2009). `*Motivating Online Publication of Data'*. BioScience 59(5):418-427.

[11] CRAGIN M. H., et al. (2010). `*Data sharing, small science and institutional repositories'*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368(1926):4023-4038.

[12] DKAKI T., et al. (1997). `*Recherche de l'information stratégique dans les bases de données : veille scientifique et technique'*. In INFORSID 97 : informatique des organisations et systèmes d'information et de décision, Toulouse, 10-13 juin 1997.

[13] DOOLEY R., et al. (2006). `*From Proposal to Production: Lessons Learned Developing the Computational Chemistry Grid Cyberinfrastructure'*. Journal of Grid Computing 4(2):195-208.

[14] DOWNING J., et al. (2008). `*SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories'*. J. Chem. Inf. Model. 48(8):1571-1581.

[15] DOUSSET B., et al. (1997). `*Veille scientifique et technique sur internet*'. In 6ème Conférence sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, Ile Rousse, Corse, France., 12-19 mai 1997.

[16] HALBERT M. (2007). `*Integrating ETD Services into Campus Institutional Repository Infrastructures Using Fedora'*. In ETD 2007 10th International Symposium on Electronic Theses and Dissertations, June 13-16, 2007, Uppsala, Sweden.

[17] HASSANZADEH O., et al. (2009). `*A framework for semantic link discovery over relational data'*. In Proceedings of the 18th ACM conference on Information and knowledge management, CIKM '09, pp. 1027-1036, New York, NY, USA. ACM.

[18] HEY T. et HEY J. (2006). `*e-Science and its implications for the library community'*. Library Hi Tech 24(4):515-528.

[19] HEIDORN P.B. (2008). *'Shedding Light on the Dark Data in the Long Tail of Science'*. Library Trends 57(2): 280-299.

[20] LATIF A. et TOCHTERMANN K. (2012). `*Webbing Semantified Scholarly Communication Datasets for Improved Resource Discovery*'. Journal of Digital Information Management 10(4):245+.

[21] LAURE E. et LIVENSON I. (2012). `*Simple Store. An Overview of a Potential New EUDAT Service'*. In EUDAT 1st Conference, October 22-24, 2012, Barcelona, Spain.

[22] LUZI D., et al. (2012). `*Enhancing diffusion of scientific contents: Open data in repositories'*. The Grey Journal 8(2):71-82.

[23] MOLLOY J. C. (2011). `*The Open Knowledge Foundation: Open Data Means Better Science'*. PLoS Biol 9(12):e1001195+.

[24] MURRAY-RUST P. (2007). `*The Power of The Electronic Scientific Thesis'*. In ETD 2007 10th International Symposium on Electronic Theses and Dissertations, June 13-16, 2007, Uppsala, Sweden.

[25] OLENDORF R. et KOCH S. (2012). `*Beyond the low hanging fruit: Archving complex data and data services at University of New Mexico'*. Journal of Digital Information 13(1).

[26] PARK E. G. et RICHARD M. (2011). `*Metadata assessment in e-theses and dissertations of Canadian institutional repositories'*. The Electronic Library 29(3):394-407.

[27] PIWOWAR H. A. et CHAPMAN W. W. (2008). `*Identifying data sharing in biomedical literature*.'. AMIA... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium pp. 596-600.

[28] POR L. Y., et al. (2012). `*A grid enabled E-theses and dissertations repository system'*. International Arab Journal of Information Technology 9(4):392-401.

[29] PRACZYK P., et al. (2012). `*Integrating scholarly publications and research data - Preparing for open science, a case study from high-energy physics with special emphasis on (meta)data models'*. Communications in Computer and Information Science 343 CCIS:146-157.

[30] SCHÖPFEL J. (2013). `*Adding Value to Electronic Theses and Dissertations in Institutional Repositories'*. D-Lib Magazine 19(3/4).

[31] SCHÖPFEL J. et LIPINSKI T. A. (2012). `*Legal Aspects of Grey Literature'*. The Grey Journal 8(3):137-153.

[32] SEFTON P., et al. (2010). `*ICE-theorem - end to end semantically aware eResearch infrastructure for theses'*. Journal of Digital Information 11(1):1-19.

[33] SIMPSON P. et HEY J. (2006). `*Repositories for research: Southampton's evolving role in the knowledge cycle'*. Program: electronic library and information systems 40(3):224-231.

[34] STEBE J. (2012). `*Responsibility for research data quality in open access: a Slovenian case'*. In Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries, JCDL '12, pp. 401-402, New York, NY, USA. ACM.

[35] SUBER P. (2012). *Open access*. MIT Press, Cambridge Mass.

[36] SUN S., et al. (2011). `*Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource'*. Nucleic Acids Research 39(suppl 1):D546-D551.

[37] WALKER E. P. (2011). `*What We Can Learn from ETDs: Using ProQuest Dissertations et Theses as a Dataset'*. In USETDA 2011: The Magic of ETDs…Where Creative Minds Meet. May 18-20, Orlando, Florida.

[38] WANG S. et LIU Y. (2009). `*TeraGrid GIScience Gateway: Bridging cyberinfrastructure and GIScience'*. International Journal of Geographical Information Science 23(5):631-656.

[39] DE WITT S. (2012). `*Metadata and EUDAT'*. In EUDAT 1st Conference, October 22-24, 2012, Barcelona, Spain.

[40] YANG C., et al. (2010). `*Geospatial Cyberinfrastructure: Past, present and future'*. Computers, Environment and Urban Systems 34(4):264-277.