Cartographie de l'expertise des chercheurs de l'université de Toulon

David Reymond, Clara Galliano

Cartographie de l'expertise des chercheurs de l'université de Toulon

Etude préliminaire et prototype

Auteurs : David REYMOND (MCF-HDR), Clara GALLIANO (doctorante).
Université de Toulon, Université Aix-Marseille, Laboratoire IMSIC Toulon
Date : 6 novembre 2019

L'équipe présidentielle nous a sollicité (juillet 2019) pour étudier la mise en oeuvre d'un système de mise en lisibilité des compétences de la recherche pour notre établissement. Les objectifs établis pour ce dernier, sont de permettre à la fois : le pilotage stratégique de la recherche scientifique par la cartographie (interactive) des compétences présentes mais aussi la possibilité d'accroître la synergie des collaborations internes et/ou externes des chercheurs. Bien entendu comme tout outil, il s'agit aussi d'en favoriser l'acceptabilité malgré les risques inhérents de confusion ou de méprise quant aux objectifs du projet puisqu'il touche à la collecte d'informations de la production des membres de l'établissement. En phase avec nos travaux dans les domaines de la bibliométrie et de l'infométrie en général, au sein des sciences de l'information et de la communication, cette problématique relève de nos compétences et s'inscrit pleinement dans les travaux de Mme Galliano sur la science ouverte.

D'emblée, une proposition est faite sur l'hypothèse (réductrice) que les compétences des membres de l'établissement transparaissent dans les publications des chercheurs. Nous ciblons alors une archive ouverte (recommandée par l'état et les institutions de la politique scientifique) pour qu'au travers des publications identifiées nous extrayons les termes clés à rattacher aux individus. L'extraction de mots clés étant un problème difficile (cf. infra), il est impératif pour nous de permettre aux dits individus d'expertiser, valider et compléter les mots clés que le dispositif a extraits tout en leur présentant sur quels éléments informationnels utiles à leur propre activité (CV, collaborateurs). In fine, le dispositif convergera vers un système de cartographies dynamiques, auto alimenté par construction et collaboration avec les membres de l'institution présents ou à venir. A contrario des autres systèmes (commerciaux et très coûteux), nous contournons le problème de la consistance des données de départ en offrant aux collègues un outil de suivi de leur production sur les archives conformément aux réglementations en cours de mises en place. Collègues qui auront alors tout intérêt à renseigner le dispositif pour que leur expertise soit valorisée. Ainsi en ayant automatiquement associé l'ensemble des membres en mettant en miroir face à ce qui est accessible de leur production sur HAL, nous catalysons le phénomène "déposer en OA" et faisons converger la cohérence des cartographies et représentation des expertises. Pour le pilotage, les agrégations sont construites au niveau des laboratoires de l'établissement sur le même principe tout en offrant des indicateurs complémentaires. Réciproquement au problème difficile de la recommandation d'expert, nous l'abordons en construisant un dispositif d'émergence de cette dernière qui, appuyé au plan politique, permettra de s'absoudre des biais disciplinaires dont souffrent les techniques automatiques d'extraction d'information tout en convergeant vers un résultat optimal. De surcroît, le dispositif contribuera à la participation massive des membres de l'UTLN à la science ouverte.

Ce qui suit en présente les points d'appui scientifiques; fondements théoriques et justifications de choix stratégiques, techniques et opérationnels les plus critiques opérés pour la réalisation du prototype.

Introduction

L'identification des compétences d'expert est une problématique essentielle dans les milieux académiques (Bouillot et al., 2013; Zevio, Zargayouna, Santini, & Charnois, 2018) ou industriel. L'émergence de la "science ouverte" et la prise de conscience de la responsabilité sociale de la recherche pose la mise en avant au plan politique de la nécessité de pouvoir réaliser ces identifications au niveau d'un établissement. L'intérêt croise la facilitation des partenariats académiques et industriels mais encore l'aide à la construction d'équipes en réponse à un appel à candidature dans le cadre de la recherche de fonds. Au sein d'un établissement universitaire, la mise en oeuvre nécessite avant tout le recueil d'information sur les membres qui le composent, la saisie et la structuration en connaissances ainsi que la mise en place d'interfaces d'interrogation. Des outils commerciaux sont disponibles (Questel) mais ne dispensent pas de la réalisation de ces tâches préliminaires qui amènent de tels dispositifs à des coûts très importants. L'embryon d'un projet de cartographie des compétences chercheurs et laboratoires de l'université de Toulon est né avec pour objectif l'élaboration d'un prototype.

Dans ce qui suit, nous développons un état de l'art survolant les différentes recherches des nombreux domaines concernés et des outils afférents. Les imperfections des traitements automatiques proviennent d'une part de l'inconsistance des données initiales et d'autre part des limites de leur application à des domaines variés. Nous situons également le contexte politique/médiatique de la science ouverte en proposant de saisir l'opportunité de rallier les recommandations et directives européennes et étatiques à la reconstruction du "capital connaissance" de l'université sous l'angle des publications de ses membres sur l'archive ouverte HAL. Le prototype s'appuyant sur des technologies sélectionnées d'extraction d'information et relativisées dans ce qui suit sera présenté en dernière partie. Il se matérialise par un ensemble modulaire de chaînes de traitements pour la réalisation d'une cartographie évolutive et dynamique des compétences et expertises de l'établissement associé à un système d'interrogation. Nous proposons en parallèle de la mise en place de cette instrumentation automatique d'associer l'ensemble des chercheurs de l'établissement à la rectification ou complétion des résultats obtenus afin d'élaborer un mode d'accompagnement des chercheurs à son utilisation et son appropriation qui, de notre point de vue, sera la clé de son évolution en termes de fiabilité et de qualité de représentation toutes disciplines confondues. Globalement le dispositif proposé contourne les limites technologiques afférentes à l'extraction

d'information, l'inconsistance initiale des données par une co-construction collaborative (Levy, 2015) dont il s'agira de catalyser la dynamique et d'inscrire une instrumentation dans une logique d'intelligence collective d'usage (Bourdoncle, 2010) au service de la science.

Etat de l'art

Expertise et connaissance

Les systèmes qui guident les usagers par la recommandation d'experts sur un sujet donné sont un domaine récent en extension des systèmes de recommandation. On en distingue deux grandes catégories : les systèmes qui aident à identifier les personnes les plus reconnues dans un domaine spécifique (*expert finding system*) et, ceux qui déterminent dans quel domaine une personne est experte (*expert profiling syste*m). Balog et al. ont identifié les différentes méthodes et applications de ces systèmes experts (Balog, Fang, de Rijke, Serdyukov, & Si, 2012). En général, cette détection se fonde à la fois sur l'extraction d'information des activités des individus et sur le contenu des documents qui les concernent (Nikzad–Khasmakhi, Balafar, & Feizi–Derakhshi, 2019). De surcroît, ces systèmes construisent à partir d'une requête usage (thème ou mot-clé) une liste d'individus triés par score de pertinence en réponse à cette requête.

Ces systèmes peuvent être modélisés par des approches de recherche d'information, avec des moteurs de recherche ou une combinaison de systèmes de traitement du langage naturel en regard des problématiques de représentation des connaissances et d'extraction d'information. Ces techniques sont fondées à la suite de l'utilisation du texte informatisé comme source informative (DeRose, Durand, Mylonas, & Renear, 1990) et le traitement statistique du texte (Lebart & Salem, 1994). La distinction de différents niveaux textuels (composition et structure d'un document et métadonnées) et les travaux fondateurs ont permit alors de convoquer linguistique et informatique pour développer le traitement automatique du langage (TAL) et rendre opérationnels des procédés de nettoyage et de filtrage nécessaires à l'extraction de connaissances et d'entités nommées (Chakrabarti, 2018). Les développements actuels visent l'utilisation de bases de connaissances ontologiques (Uren et al., 2006) pour affiner les résultats et développer une couche sémantique plutôt que statistique (Cifariello, Ferragina, & Ponza, 2019).

Dans ce contexte de travaux et développement en cours, différentes étapes et leviers technologiques ont été franchis à différent degrés d'efficacité. Trois éléments composant de tels systèmes sont retenus pour notre proposition : l'indexation documentaire, les technologies de fouille de texte et de traitement automatique du langage et, source informative, les bases d'archive ouverte.

Indexation documentaire

L'indexation des documents, automatique (Salton, 1971) ou supervisée (Boyce & Lockard, 1975; Amar, 2000 ;Cleveland & Cleveland, 2013) ne sont pas récentes. Ce sont traditionnellement deux étapes qui sont appliquées : dans un premier temps l'identification de termes clés, puis la sélection des meilleurs candidats. L'étude des cooccurrences de termes et l'utilisation de connaissances du domaine permet depuis longtemps d'obtenir des résultats exploitables (Toussaint et al., 1998). TAL et recherche d'information (RI) se rejoignent dans la mise en application de la représentation des textes, des mesures de similarité (Claveau, 2012) qui constituent les fondements mathématiques de l'indexation. Des expériences d'indexation contrôlée automatique et manuelle sur un corpus en français ont démontré l'intérêt de combiner différentes approches pour améliorer les résultats (Savoy, 2005). Des approches plus récentes en matière d'indexation automatiques prennent en compte la cooccurrence des termes associées à la structure des documents (Paroubek, Zweigenbaum, Forest, & Grouin, 2012).

Fouille de texte et traitement automatique du langage

L'analyse instrumentée de documents est un thème qui a particulièrement évolué depuis l'avènement technologique et la capacité de traitement que l'on a conféré aux machines. Le traitement automatique des langues (TAL) développe des procédés informatique mettant en jeu des règles linguistiques (Jacquemin & Zweigenbaum, 2000) pour produire des briques opérationnelles à plusieurs niveaux d'application. Ces applications sont accessibles après un prétraitement commun à la plupart des techniques :

1. nettoyage (signes de ponctuation, quelquefois les chiffres et caractères spéciaux) et identification des phrases et des termes du texte
2. extraction des mots vides
3. Lemmatisation et éventuelle racinisation des mots

Une fois les pré-traitement établis, fouille de texte et linguistique s'adjoignent pour répondre à la problématique issue de la massification de la donnée textuelle : modéliser puis mettre en œuvre des méthodologies appliquées aux données textuelles afin d'en déterminer le sens et/ou découvrir des connaissances nouvelles (Torres-Moreno, 2007). Dans ce processus, le descripteur linguistique constitue un élément pivot (Roche, 2011) qui permet de dépasser la dimension statistique du "sac de mots" en intégrant la position relative des termes et leur fonction linguistique dans les modèles .

C'est ainsi que de nombreuses applications atteignent aujourd'hui des niveaux opérationnels remarquables. Nous ne retiendrons que les applications les plus

pertinentes à nos objectifs : l'analyse et annotation automatique de texte pour l'identification des catégories lexicales (noms, verbe, ...), des motifs (Charnois, 2011) et des entités nommées qui constituent les briques élémentaires de l'extraction de phrases clés et de mots clés en objectif principal de traitement des données textuelles.

Notons que si ces outils et techniques ont atteint des niveaux de fiabilité en moyenne très honorables, il n'en demeure pas moins qu'ils sont perfectibles et sujets à des difficultés encore en vigueur dès lors qu'il s'agit de les appliquer à certains domaines de spécialité (Bougouin, Boudin, & Daille, 2014), notamment en SHS.

Référentiels documentaires et bases de données ouvertes

La problématique de départ, visant à extraire des informations concernant les chercheurs d'une communauté scientifique, est grande et bien réelle. Les informations relatives aux enseignants-chercheurs, de type "information scientifique" (autrement dit, toute la production scientifique : articles, communications, ouvrages, chapitres d'ouvrage…) sont largement accessibles aujourd'hui grâce à la numérisation, l'archivage et l'ouverture des données. Si le dépôt en archive ouverte n'est pas encore rendu obligatoire, l'obligation de la diffusion des résultats de la recherche/des connaissances est largement envisagée par les politiques (cf. Plan National pour la Science ouverte[1]).

La diffusion des résultats de la recherche passe aujourd'hui par deux vecteurs/circuits :

- la publication dans les revues à comité de lecture chez un éditeur (plus longue et plus coûteuse)
- les transmissions électroniques : revues en libre accès (PLos Biology, Biomed Central) qui restent payantes pour les institutions, et les archives ouvertes (ArXiv, HAL), totalement gratuites.

Les archives ouvertes jouent un rôle essentiel dans la diffusion des résultats de la recherche (ouverture au public, accès rapide, pas de limitation financière). Cependant, elles sont toujours impactées par l'évaluation/la validation des articles par les éditeurs, ce qui peut demeurer un frein sur le long terme. Le système des archives ouvertes s'ajoute également aux archives institutionnelles, déjà présentes dans les organismes, les universités en fonction des politiques et des choix stratégiques de celles-ci.

[1] https://www.enseignementsup-recherche.gouv.fr/cid132531/plan-national-pour-la-science-ouverte-discours-de-frederique-vidal.html

Cette nouvelle approche peut également être faussée par les hiérarchies des moteurs de recherche en fonction de la pertinence des méta-données : référencement, indexation (du contenu), liens... visant à une monopolisation de la diffusion des connaissances sur Internet par certain moteur de recherche.
Le *Directory of Open Access Repositories*[2] compte près de 4 447 référentiels/réservoirs en Open Access dans le monde (dont 4,8% en Afrique, 28,8% en Amérique, 19,7% en Asie, 44,2% en Europe et 2,27% en Océanie). L'Open Access en France se traduit par la présence de plateformes d'archives ouvertes nationales et thématiques : HAL, Pleiadi, OpenAire, ArXiv, Pubmed, REPEC et SSOAR. Le plus utilisé en France est l'archive ouverte pluridisciplinaire HAL[3] (Hyper Articles en Ligne) destinée au dépôt de la production scientifique (déjà publiée ou non) émanant des établissements d'enseignement et de recherche, des laboratoires publics et privés. HAL recense aujourd'hui plus de 1 951 927 ressources bibliographiques, dont 624 384 en full text (48,9% d'articles scientifiques). Depuis sa création par le CCSD (Centre pour la Communication Scientifique Directe) en 2001, l'archive ouverte est devenue le référentiel de base, **soutenu par le CNRS et la politique française concernant le dépôt ou la diffusion des connaissances**. Les fonctionnalités de HAL se sont également ouvertes et ne sont plus limitées au simple dépôt d'article scientifique : création de CV, API d'interface, traitements sémantiques, désambiguïsation de l'homonymie. De nombreux outils dérivés de la plateforme ont été créés par la suite, au profit de la communauté scientifique et des universités/institutions/laboratoires de recherche.
En juillet 2006, la signature d'un protocole d'accord entre le CNRS, l'INRIA, l'INRA, l'INSERM, le CEMAGREF, l'IRD, l'institut Pasteur, la CPU, la CGE et plus récemment par le CEA, l'IFREMER, l'INERIS, l'INRETS positionne HAL au coeur de l'archive ouverte de la recherche française.

Le référencement des publications est également un problème (sensible) qui ne favorise pas la diffusion des connaissances. Aujourd'hui, aucun outil bibliométrique ne permet d'avoir l'extraction exacte de la production d'un laboratoire. En France, certains chercheurs sont affiliés à plusieurs laboratoires, ou des laboratoires en co-tutelles : informations souvent négligées, oubliées ou volontairement effacées par les politiques éditoriales des journaux (dont le but est financier, scientifique et non bibliométrique). On peut également rajouter à cela le nombre de contributeurs à un publication et leur positionnement relatif qui peut entraîner des difficultés dans l'évaluation. L'Observatoire des sciences et des techniques, ainsi que le CNRS, se penchent actuellement sur cette question et tentent de réfléchir à une normalisation de ces aspects[4].

[2] https://v2.sherpa.ac.uk/view/repository_by_country/countries=5Fby=5Fregion.html
[3] https://hal.archives-ouvertes.fr/
[4] http://www.cnrs.fr/fr/presentation/ethique/comets/index.htm

Des initiatives en faveur de la science ouverte se développent de plus en plus, au niveau national et à l'échelle européenne. Si le mouvement de l'Open Access tire son origine de la création du premier journal scientifique "*Philosophical Transactions of the Royal Society of London*" en 1665, de la première plateforme de dépôt électronique ArXiv en 1991 ou encore de la création de la première revue en libre accès Revues.org en 1999 : plusieurs faits historiques ont fait état des lieux des différentes politiques mondiales en faveur du libre accès. Au niveau national et européen, les textes fondateurs du mouvement de l'Open Access sont connus sous le nom de "BBB" (Suber, 2016) en références à : l'*Initiative de Budapest* pour l'accès ouvert (2002), dont Suber est un des initiateurs et signataires ; la *Déclaration de Bethesda* sur la publication en accès ouvert (2003) ; et la *Déclaration de Berlin* sur l'accès ouvert aux connaissances dans les sciences et dans les humanités (2003).
A ces déclarations internationales, Frédérique Vidal, Ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation , lors de son discours sur le *Plan National pour la Science Ouverte* en Juillet 2018, a affirmé la position et l'engagement de la France dans le but de suivre les propositions de la *Déclaration de San Francisco* (2013) et d'appliquer les principes du *Manifeste de Leiden* (2015) pour la mesure de la recherche. Ces initiatives donneront naissance en France à la *Loi pour une République Numérique* votée en 2016, dont notamment les articles 30 qui reposent sur le droit du libre accès à l'auteur et non plus l'éditeur, sous forme d'embargo de durée variable suivant les disciplines et les éditeurs, et 6 sur l'obligation des universités et des organismes concernant les données de la recherche. Ces mesures politiques seront appuyées quelques années plus tard par l'*Appel d'Amsterdam* (2016) repris l'année suivante par l'*Appel de Jussieu* (2017) pour la science ouverte et la bibliodiversité.
En Septembre 2018, une coalition d'agences de financement de la recherche lance le *Plan S* en complément des objectifs fixés par le programme H2020.

Si les enjeux de l'ouverture de la science reposent avant tout sur des facteurs économiques et politiques, il est important de souligner les enjeux éthiques qui ont également une place très importante dans la diffusion de la science. Evaluation de la recherche, attitude élitiste de restriction, conflits d'intérêt, monopole des langues, décision de publication : toute action produite/faite par les chercheurs ou par les maisons d'édition présente un frein, à un moment donné, non négligeable, dans ces dimensions éthiques, pour la publication scientifique et l'ouverture la science dans son ensemble. Il est clair de rappeler l'une des missions du chercheur, trop souvent oubliée au profit d'intérêts stratégiques, qui est de diffuser ses connaissances et ses résultats à la communauté scientifique et à la société (cf. Responsabilité sociale du chercheur). De nombreuses procédures sont mises en place par les universités, leurs bibliothèques et les agences de financement de la recherche pour encourager, appuyer, guider et assurer (voire même déléguer) les pratiques du chercheur afin de

privilégier l'accès ouvert. Comme l'affirme Peter Suber : "*mais ce sont ces derniers qui, ultimement, font ou non le choix de ce mode de diffusion. Quelques résistances sont observées chez les chercheurs, inquiets de ce qu'implique ce changement de paradigme sur la diffusion et la reconnaissance de leurs travaux. Ce sont des obstacles culturels qu'il faut surmonter à ce moment-ci*".
Marin Dacos, conseiller pour la Science Ouverte DGRI, affirme lors d'un entretien[5] "*une grosse amélioration ergonomique a déjà été apportée sur HAL en 2017, qui a consisté à demander beaucoup moins de métadonnées aux chercheurs. Nous travaillons également sur un autre projet de simplification de la vie du chercheur, en mettant en place un moissonnage de publications déjà déposées ailleurs en un seul clic"*.

Récemment, le baromètre français de la science ouverte (Jeangirard, 209) a été dévoilé. L'étude menée sur les publications accessibles en 2017 a été mesurée en 2018 à partir de la ressource ouverte Unpaywall et via les données du Ministère de l'Enseignement, de la Recherche et de l'Innovation (MESRI).
Selon le baromètre, la part des publications en accès ouvert représente 41% (dont 24% chez les éditeurs, et 16% sur les archives ouvertes).
L'évolution du taux d'accès ouvert ainsi que la répartition par discipline ont également été étudiées[6].
Comme le souligne Marin Dacos (2019) lors de son intervention pour la science ouverte : "aussi ouvert que possible … aussi fermé que nécessaire".

Le prototype

Se décompose en deux étapes modulaires et complémentaires : un script de collecte et un script de traitement et visualisation des données.

Le script de collecte

Données d'entrée

Le script de collecte schématisé sur la figure suivante prend en entrée l'organisation des membres de l'établissement sur l'organisation hiérarchique de décomposition par laboratoire eux-même composés de membres. A ce stade, le laboratoire est

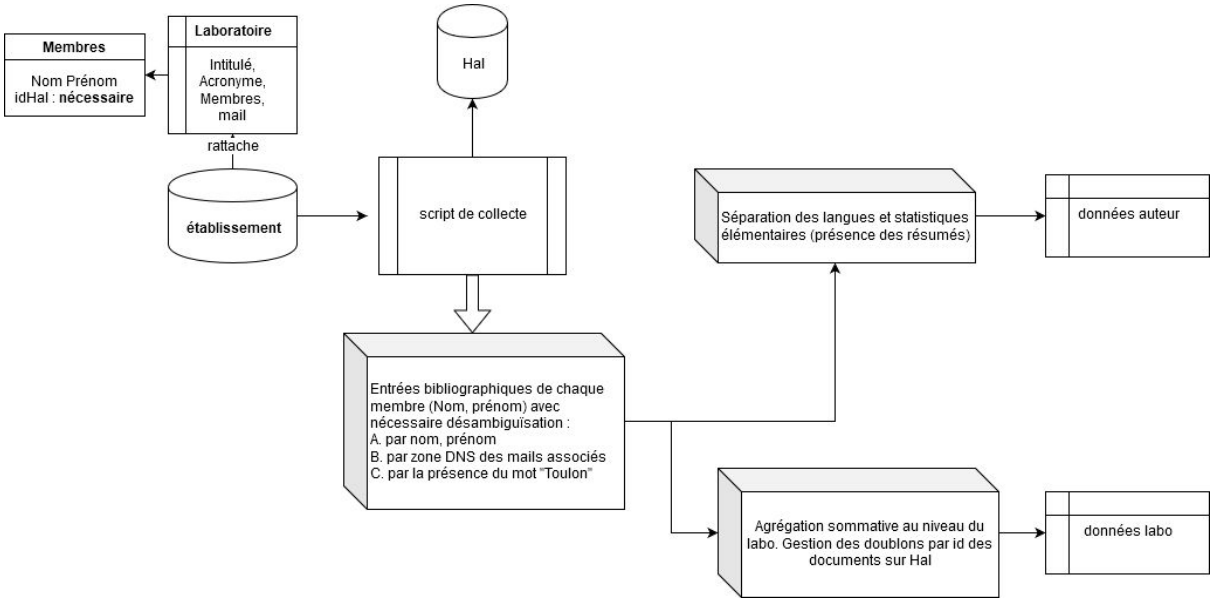
5 https://education.newstank.fr/fr/tour/news/135985/science-ouverte-dacos-detaille-strategie-entrainer-tous-acteurs-esr.html

6 https://data.enseignementsup-recherche.gouv.fr/explore/dataset/open-access-monitor-france/information/?disjunctive.oa_host_type&disjunctive.year

décrit par son intitulé, son acronyme, les membres le composant et une liste d'e-mails. Chaque élément est une liste ce qui permet d'associer des dénominations multiples et suivre de fait les évolutions historiques.

Les membres sont eux-mêmes décrits par leur nom et prénom. Notons que le prototype a permis de montrer la nécessité de la création d'un idHAL seule forme de désambiguïsation fiable pour les objectifs d'élimination de l'homonymie des auteurs.

Collecte et traitements



Sur les noms et prénoms des auteurs, le script récupère toutes les publications associées sur HAL et désambiguïse les résultats sur le critère "Nom Prénom". Mais, souvent les formes de signature sont très variées (plusieurs prénoms au lieu d'un, seule la première lettre, etc.). Ainsi, une procédure complémentaire est mise en place sur le critère de mail des auteurs comme relevant de la zone dns du laboratoire (ou de l'université) puis encore, la présence du mot Toulon dans l'adresse associée à la publication. Nous notons que ces éléments ne sont pas suffisant pour garantir la fiabilité de la collecte. L'homonymie est un problème difficile de la bibliométrie (Bertin, Desclés, & Krushkov, 2006) et plutôt que de développer un affinage automatique dont l'efficacité ne sera que temporaire, nous préconisons de déléguer aux auteurs eux-même cette tâche. Ceci peut s'effectuer au travers de la création d'un identifiant HAL qui associe à l'individu l'ensemble des écritures de son nom et prénoms trouvées dans les publications en proposant de les rattacher interactivement avec l'auteur. Cette méthode assure une fiabilité et une maintenance de ce point d'écueil du système.
Le script génère alors pour chaque auteur et, par agrégation pour le laboratoire, des corpus d'individu par langue. La langue est celle décrite par les métadonnées OU à

défaut identifiée automatiquement sur les résumés. Actuellement, français, anglais, espagnol et portugais sont automatiquement identifiées. Les productions dans d'autres langues sont abandonnées : les traitement linguistiques appliqués par la suite sont dépendant de la langue et imposent cette identification préalable pour atteindre des degrés suffisants de pertinence en regard des objectifs.

Limites et extensions

La limite décrite est celle de l'ambiguïté liée à l'homonymie. L'idHAL permettra de contourner et de résoudre à la limite ce problème difficile.
Sur le même principe, des extensions du collecteur sont possibles (prototype v2) en s'appuyant sur deux autres bases officielles et d'intérêt pour cette cartographie :

- la base des thèses française
- la base européenne des brevets

Le script de traitement et de création d'interfaces

Une fois la collecte réalisée, le traitement opère sur les données auteurs et laboratoire précédemment collectées. Les opérations sont réalisées à deux niveaux : le niveau bibliographique général permet la création d'indicateurs élémentaires et le niveau des contenu l'extraction des mots clés que l'on associe aux auteurs et aux laboratoires. Le schéma suivant récapitule les différents traitements effectués.

Eléments bibliographiques

A ce niveau, sont traitées pour un individu ou pour le laboratoire:

- la liste des publications
- la détermination des entrées avec résumé et par langue
- la création des réseaux de collaborations reconstruits comme la liste des co-auteurs (considérés par leur nom seulement) de chaque publication. Le réseau des co-auteurs sépare par couleur les co-auteurs internes au laboratoire et identifie le degré de collaboration par le nombre d'apparition du nom dans l'ensemble global des publications recensées).

Extensions possibles (prototype v.2) : historisation des publications, différentiation de agrégations sommatives d'une collecte directe "laboratoire" et des collections des laboratoires.

Traitement des contenus et extractions d'informations

Traitement génériques

Pour chaque résumé identifié (et pour chaque langue), les traitements suivant sont opérés afin d'extraire les phrases clés, les mots-clés et les entités nommées. L'indexation (i.e associer à un expert des mots dans une base permettant

d'interroger par des termes) n'est pas encore réalisée et est présentée en pointillées dans le schéma.
Certaines entrées bibliographiques présentent des mots-clés dénommés "mots-clé auteur" par la suite. Cette information est utilisée actuellement dans la représentation des résultats. Notons que ces termes choisis par les auteurs sont en règle générale posés à destination d'experts des domaines et sont en ce sens insuffisants pour répondre à une logique d'interrogation de recherche d'expert. D'autant plus que toutes les entrées bibliographiques ne présentent pas ces éléments. Notons que l'on pourrait aussi caractériser les productions en associant à chaque publication la thématique des revues dans lesquelles elles ont été publiées ainsi que les mots-clés afférents (Bouillot et al., 2013).
Pour compléter ces mots-clés auteurs ou combler leur défaut de présence, nous procédons au traitement des résumés par langue. En nous positionnant sur le résumé et le titre des entrées bibliographiques nous maximisons la quantité informationnelle représentant la production, et en ce sens, opèrerons avec les méthodes dites "extractives" de sélections des phrases et mots clés dans les résumés. Deuxièmement, les résumés étant grossièrement de taille comparable nous éviterons les effets de différence de taille que l'on peut identifier dans la plupart des techniques de traitement documentaire. Nous nous appuyons sur les techniques d'extraction suivantes :

- des mots clés selon l'algorithme *TextRank* (Mihalcea & Tarau, 2004) d'efficacité prouvée sur l'extraction de mots clés à partir de textes courts ;
- des N-grammes retenus significatifs par leur occurrence (Callon, Courtial, & Turner, 1991; Callon, Courtial, Turner, & Bauin, 1983) ;
- des phrases-clés selon l'algorithme *PositionRank* (Florescu & Caragea, 2017).

Les prétraitements (identification de langue, découpage en phrase, lemmatisation, et identification d'entités nommées) sont réalisés à l'aide de spaCy (Explosion, 2017).
Notons que l'aspect modulaire des développements nous permettra d'intégrer l'état de l'art en la matière de façon récurrente. La version 2 mettra en oeuvre des techniques dépassant au plan des résultats sur des corpus contrôlés les technologies précédentes (Boudin, 2016; Campos et al., 2020).

Sélections d'informations pertinentes

Un poid prépondérant est donné par l'ordre suivant :

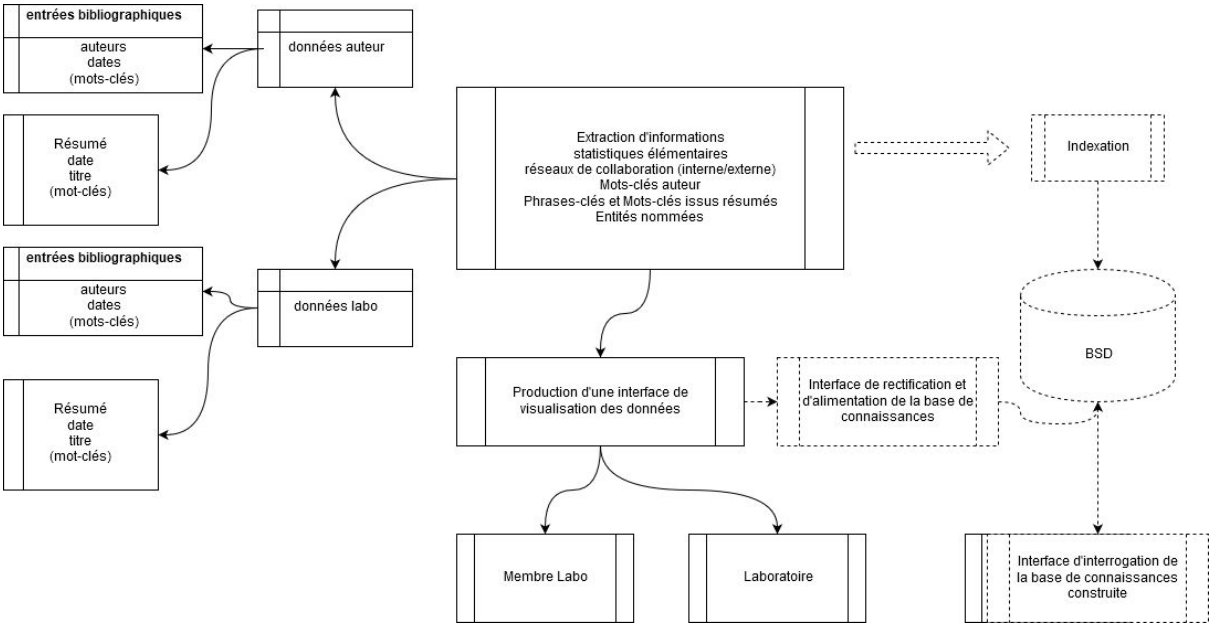
1. Mots-clés auteur
2. N-Grammes autrement appellé co-occurrence terminologique Entités nommées (Lieux, organisations, symboles chimiques)
3. Mots-clés extraits appartenant aux phrases clés.

Ainsi, la sélection des termes clés s'effectue selon les critères suivant :

- si le mot clé n'est pas un mot auteur (sinon ce dernier est prépondérant)

- si le mot-clé n'est pas une entité nommée
- si le mot-clé appartient à une phrase clé
- si le mot-clé ne compose pas un bigramme
- si son occurrence est significative
- alors ce dernier est retenu en tant que mot extrait

L'occurrence des termes extraits dans le corpus sert alors de pondération dans sa représentation.



Nous pensons pouvoir faire évoluer à terme cet algorithme de traitements en incluant une hiérarchisation de la terminologie clé qui utilise l'information structurelle issue de la collecte d'information (laboratoire, auteur, publications) qui devrait permettre d'imbriquer la "spécialisation terminologique" à la terminologie générique issue des mots-clés laboratoires.

Interfaces

Construite dans l'état de l'art des technologies du Web (HTML5, CSS3, Flex) pour la composition des interfaces, les visualisations s'appuient sur les technologies "Data Driven Document" pour la représentation des résultats (Bostock, Ogievetsky, & Heer, 2011). Aussi modulaires que les traitements précédemment décrits, les interfaces sont construites et alimentées par le script de traitement des données.

Pour chaque membre une interface présente la synthèse des données collectées :

- données bibliographiques et référence à la moyenne du laboratoire quant au nombre de publications individuelles présentes et au degré de consistance en regard des résumés
- liste (partielle) de publications identifiées présentées selon les trois éléments majeurs des critères AERES (ACL, OS, COMS)

- Réseau des collaborateurs
- Nuage des mots-clés identifiés par couleur sur les trois éléments informationnels suivant :
    - mots clés auteur
    - mots-clés extraits
    - entités nommés

Pour le laboratoire, la même interface présente les mêmes éléments par agrégation de l'ensemble de ses membres.
Les fonctions de rectification et d'alimentation de la base de connaissance ne sont pas encore implémentées (en pointillé sur le schéma). De même que l'interface d'interrogation de ladite base. Les interfaces sont accessibles là : http://tests.vlab4u.info/Auteurs/IN2MP
Les fichiers html avec la convention de nommage suivante :
- *PrénomNoms.html* pour les interfaces "auteur"
- *SigleLaboratoire.html* pour l'interface laboratoire

Limites et "to do list"

Le prototypage est suivi d'une phase de test/ de vérification afin de connaître les limites (techniques et opérationnelles) de notre projet.
Rapidement, plusieurs problèmes (connus de nos autres projets) ont refait surface :

- problème d'homonyme : d'où l'importance de créer l'identifiant unique IdHAL
- problème de norme de saisie : par exemple, pour un enseignant-chercheur de l'IM2NP, nous avons détecté 11 façons différentes de citer l'auteur sur une publication. Tout le travail de la désambiguïsation lexicale (Cuxac, Collignon, Gregorio & Parmentier, 2019).

En définitive, le projet de cartographie des compétences ne reflète que les éléments qui sont présents sur l'archive ouverte HAL. Par conséquent, il est d'autant plus souhaitable que les enseignants-chercheurs déposent davantage et qu'ils soient plus vigilants quant à la saisie des informations relevant de leur production scientifique. C'est pourquoi une phase de "validation" sera nécessaire pour chaque profil établi, par l'enseignant-chercheur en amont.

Les différentes limites ont été notées dans ce qui précède. Nous en rappelons la teneur en précisant la quantité de travail nécessaire à leur dépassement :
La limite du collecteur se situe essentiellement sur le point d'entrée unique HAL. Cependant, en incitant à conférer à ce dépôt un caractère de référentiel unique il s'agit aussi à terme de palier au levier technologique de l'homonymie dans les données bibliométriques. Le collecteur pourra à terme intégrer :

- des éléments issues de la la base des thèses (mots-clés, résumés, collaborateurs, domaines etc.).
- les différents brevets des laboratoires

Les traitements d'extractions de données doivent être stabilisés en étant confrontés aux membres de l'établissement dans une phase "béta" de validation du dispositif. Cette phase servira l'initialisation de l'accompagnement en établissant le degré d'intérêt des chercheurs en regard des fonctions offertes et potentiels en vigueur.
Au plan des interfaces, il s'agit de stabiliser une première version en phase avec les informations collectées et traitées par la sélection de représentations plus adéquates (par ex. l'historisation des publications serait plus pertinent que la carte proportionnelle imbriquée).
Enfin, l'indexation des termes extraits est à relier aux différents experts et laboratoire en mettant en oeuvre une base de connaissance et l'interface d'interrogation associée.
Nous suggérons d'inviter des experts nationaux des différents domaines convoqués pour assurer un regard critique et performatif des solutions proposées et mise en oeuvre tout en conduisant une phase de communication sur le projet.
En ce qui concerne les perspectives, nous pensons sur le long terme développer le projet à d'autres sources d'informations. En effet, pour le prototype, nous nous réservons uniquement la production scientifique "académique". Cependant, nous pourrons également, dans le cadre "des compétences chercheurs" ajouter des informations concernant les dépôts de brevets, la vulgarisation (journaux, radios, presse), les logiciels, les données de la recherche...
Il existe de nombreuses possibilités d'extension par rapport à notre projet. Nous gardons à l'esprit toutes ces éventuelles améliorations à apporter.

Propositions de continuité, planification et coûts

Structuration annuelle

Sur trois ans nous proposons de développer de la sorte le projet :
2020 : prototype, tests et méthodologie d'accompagnement (quelques laboratoires)
2021 : intégration ENT, consolidation des extractions de données et extension (tout l'établissement)
2022 : Validation, bilan et mesures d'efficacité. Portage externe.

Tâches principales (1ere année)

Le tableau suivant présente les grandes tâches de développement du projet :

Tâches (livraison)	**Responsable**	**Durée**
Collecteur (janvier 2020) - consolidation	DR DR	3 mois - 1 mois

- intégration SI - extensions des sources	DSIUN DR	- 2 mois - 2 mois
Traitements (janvier 2020) - consolidation - vérifications et phase “béta”	 DR CG	3 mois - 10j - 3 mois
Interfaces (janvier 2020) - consolidation - représentations optimisées - extensions fonctionnelles - intégration outils HAL	DR DR LC DR, CG CG	stagiaire DASI 1 mois
Indexation (juin 2020 - pré-étude) - choix technologiques - Mise en oeuvre - tests et validation	 DR, DSIUN (service externe ?) consortium esup	6 mois
Accompagnement (juin 2020) - plan de formation - supports - tests	DR, CG	5 mois stagiaire DASI

Planification prévisionnelle

Globalement les tâches de consolidation et de représentations optimisées (collecteur, traitement, interfaces) peuvent être menées d’ici décembre 2019. La phase d’intégration au SI (opérationnalisation et faisabilité quant aux interfaces) pourra démarrer en janvier 2020. En parallèle la phase “bêta version” de vérification et d’adéquation en regard des retours des experts pourra démarrer à ce moment là avec pour objectif un échantillonnage représentatif des différentes disciplines, l’identification d’acteurs dans différents laboratoires pour entretiens et présentation du système.

Après validation par la DIREP et les instances politiques nous pourrions ouvrir une concertation auprès d’experts nationaux (à identifier) ayant oeuvré à des travaux connexes ou du domaine (extractions de connaissance, recommandation, indexation, etc.) en mai ou juin 2020.

Un bilan établi en septembre 2020 pour établir l’avancement et les premier résultats des travaux.

Conclusions

L'état de l'art dans les domaines connexes ou composant un système de cartographie des compétences des chercheurs a montré la dynamique des développements actuels en soulignant que la charnière d'un outil automatisant le système se situe à la croisée de travaux actuels variés. En saisissant la vague commanditée par les instances étatiques et européenne de favoriser le dépôt en archive ouverte par les chercheurs nous proposons de construire un système de cartographie qui, d'une part, s'alimente et se valide directement à partir des chercheurs et, d'autre part, contribue à accompagner, faciliter et favoriser les dépôts des publications.

Le système construit dans l'état de l'art des technologies tant d'interface que d'extraction des données est modulaire. De fait, sa maintenance est facilitée, ouverte et le code sera mis à disposition de la communauté. Nous pensons qu'il est vital que cet outil soit intégré aux ENT des établissements, et le consortium e-sup avec qui nous avons déjà travaillé pourrait être adjoint à cette intégration.

La continuité proposée est de délivrer une version alpha dès le début 2020 pour réaliser les tests et la mise en conformité du système sur le SI. Des experts nationaux pourraient s'adjoindre au projet (cette problématique est générale dans les établissements) pour conforter choix techniques et stratégiques tout en diffusant son existence. Nous pouvons développer un **plan de formation** pour accompagner les agents (laboratoire, secrétariats, services de documentation) pour conduire une progression très rapide du renseignement des données sur HAL et conforter la stratégie adoptée : l'utilisation d'une gamme d'outils qui permettent de retrouver les publications existantes de chaque chercheur, de les transposer dans un format compatible afin de les intégrer à l'archive ouverte. Le plan de formation est autant d'intérêt que le dispositif construit : l'objet est de catalyser le démarrage de l'appropriation en facilitant la tâche du chercheur par son accompagnement et la mise en oeuvre de procédures automatisées qui participeront de l'appropriation générale du dispositif et contribueront à son succès.

Références

Amar, M. (2000). *Les fondements théoriques de l'indexation. Une approche linguistique.* ADBS Éditions.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise retrieval. In *Expertise retrieval*. Consulté à l'adresse https://ieeexplore.ieee.org/document/8187432

Bertin, M., Desclés, J.-P., & Krushkov, Y. (2006). Critique de la bibliométrie comme outil d'évaluation, vers une approche qualitative. *Digital humanities 2006*. Présenté à Paris, France. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-02125151

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3 data-driven documents. *IEEE Transactions on Visualization And Computer Graphics*, *17*(12), 2301‑2309. https://doi.org/10.1109/TVCG.2011.185

Boudin, F. (2016). pke : An open source python-based keyphrase extraction toolkit. *Proceedings of COLING 2016, the 26th international conference on computational linguistics: system demonstrations*, 69‑73. Consulté à l'adresse http://aclweb.org/anthology/C16-2015

Bougouin, A., Boudin, F., & Daille, B. (2014). Influence des domaines de spécialité dans l'extraction de termes-clés. *Traitement Automatique des Langues Naturelles (TALN)*, 13‑24. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-01021452

Bouillot, F., Gout, O., Magnier, P., Pénin, C., Poncelet, P., & Roche, M. (2013). *Vers un outil de cartographie : Qui est l'expert ?* Consulté à l'adresse https://hal-lirmm.ccsd.cnrs.fr/lirmm-00798073

Bourdoncle, F. (2010). L'intelligence collective d'usage. In J.-M. Noyer & B. Juanals (Éd.), *Technologies de l'information et intelligences collectives*. Hermès Science Publications-Lavoisier.

Boyce, B., & Lockard, M. (1975). Automatic and manual indexing performance in a small file of medical literature. *Bulletin of the Medical Library Association*, *63*(4), 378.

Callon, M., Courtial, J.-P., & Turner, W. (1991). La méthode leximappe : Un outil pour l'analyse stratégique du développement scientifique et technique. In D. Vinck (Éd.), *Gestion de La Recherche. Nouveaux Problèmes, Nouveaux Outils* (p. 207‑277). Bruxelles: De Boeck.

Callon, M., Courtial, J.-P., Turner, W. A., & Bauin, S. (1983). *From translations to problematic networks : An introduction to co-word analysis*. *22*(2), 191–235.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257‑289. https://doi.org/10.1016/j.ins.2019.09.013

Chakrabarti, S. (2018). Knowledge extraction and inference from text : Shallow, deep, and everything in between. *The 41st international ACM SIGIR conference on research & development in information retrieval*, 1399‑1402.

Charnois, T. (2011). *Accès à l'information : Vers une hybridation fouille de données et traitement automatique des langues* (Habilitation à diriger des recherches, Université de Caen). Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00657919

Cifariello, P., Ferragina, P., & Ponza, M. (2019). Wiser : A semantic approach for expert finding in academia based on entity linking. *Information Systems*, *82*, 1‑16.

Claveau, V. (2012). Vectorisation, Okapi et calcul de similarité pour le TAL : pour oublier enfin le

TF-IDF. *TALN - Traitement Automatique des Langues Naturelles*. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-00760158

Cleveland, A. D., & Cleveland, D. B. (2013). *Introduction to indexing and abstracting : Fourth edition*. Consulté à l'adresse https://books.google.fr/books?id=JfPXAQAAQBAJ

Cuxac, P., Collignon, A., Gregorio, S., & Parmentier, S. (2019). Des bases de données massives au Web de données : désambiguïsation et alignement d'entités géographiques dans les textes scientifiques. *12ème Colloque international d'ISKO-France : Données et mégadonnées ouvertes en SHS : de nouveaux enjeux pour l'état et l'organisation des connaissances ?*. Octobre 2019, Montpellier

Dacos, M. (2019). Pourquoi la Science ouverte ? Un point de vue français. *Colloque international « Science ouverte au Sud »*. 23 au 25 octobre 2019, Dakar (Sénégal)

DeRose, S. J., Durand, D. G., Mylonas, E., & Renear, A. H. (1990). What is text, really? *Journal of computing in higher education*, *1*(2), 3‑26.

Explosion, A. (2017). SpaCy-Industrial-strength natural language processing in python. *URL: https://spacy. io*.

Florescu, C., & Caragea, C. (2017). PositionRank : An unsupervised approach to keyphrase extraction from scholarly documents. *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, 1105‑1115. https://doi.org/10.18653/v1/P17-1102

Jacquemin, C., & Zweigenbaum, P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. In *Le document en sciences du traitement de l'information* (Vol. 4, p. 71‑109). Toulouse: Cépadués Éditions.

Jeangirard, E. (2019). Monitoring Open Access at a national level: French case study. *LPUB 2019 23rd edition of the International Conference on Electronic Publishing*, Jun 2019, Marseille, France. ⟨10.4000/proceedings.elpub.2019.20⟩.

Lebart, L., & Salem, A. (1994). *Statistique textuelle* (Dunod). Consulté à l'adresse http://lexicometrica.univ-paris3.fr/livre/st94/st94-tdm.html

Levy, P. (2015). *The emergence of reflexive collective intelligence*. Consulté à l'adresse http://pierrelevyblog.com/tag/ieml/

Mihalcea, R., & Tarau, P. (2004). Textrank : Bringing order into text. *Proceedings of the 2004 conference on empirical methods in natural language processing*, 404‑411.

Nikzad–Khasmakhi, N., Balafar, M., & Feizi–Derakhshi, M. R. (2019). The state-of-the-art in expert recommendation systems. *Engineering Applications of Artificial Intelligence*, *82*, 126‑147.

Roche, M. (2011). *Fouille de Textes : De l'extraction des descripteurs linguistiques à leur induction* (Habilitation à diriger des recherches, Université Montpellier II - Sciences et Techniques du Languedoc). Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00816263

Salton, G. (1971). *The smart retrieval system : Experiments in automatic document processing*. Upper Saddle River, NJ, USA: Prentice-Hall.

Suber, P. (2016) *Qu'est-ce que l'accès ouvert ?* Nouvelle édition [en ligne]. Marseille : OpenEdition Press. Disponible sur Internet : . ISBN : 9782821869806. DOI : 10.4000/books.oep.1600.

Torres-Moreno, J.-M. (2007). *From text to digital : Automatic analysis and classification* (Habilitation à diriger des recherches, Université d'Avignon). Consulté à l'adresse https://tel.archives-ouvertes.fr/tel-00390068

Toussaint, Y., Namer, F., Daille, B., Jacquemin, C., Royauté, J., & Hathout, N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. *Proceedings, TALN*, *98*, 182‑191.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., & Ciravegna, F. (2006). Semantic annotation for knowledge management : Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, *4*(1), 14‑28.

Zevio, S., Zargayouna, H., Santini, G., & Charnois, T. (2018). Vers une cartographie automatique des thématiques et profils d'experts associés à une conférence scientifique : 9 ans d'ateliers Recherche d'Information SEmantique (RISE). *10ème Atelier Recherche d'Information SEmantique (RISE)*. Présenté à Rennes, France. Consulté à l'adresse https://hal.archives-ouvertes.fr/hal-02004675