

Acronyme / Acronym	BDBfr
Titre du projet	Création d'une base de données bibliographique/bibliométrique des principaux contenus scientifiques de santé en français (journaux scientifiques, encyclopédie, livres) disponible gratuitement sur l'Internet
Proposal title	Creation of a bibliographic/bibliometric database including the main Health content in French (journals, encyclopedias, books) freely available on the Internet
Grands défis sociétaux 2014	Santé et bien-être

INTRODUCTION

Le web est une source majeure d'informations pour les professionnels de santé. Mais l'accès à ces informations et leur utilisation peuvent être plus difficiles pour les francophones à cause de la place prédominante et toujours croissante de la langue anglaise dans les sciences de la santé. Au sein des principales sources documentaires, les contenus comme les interfaces sont le plus souvent en anglais.

L'objectif de ce projet est de mettre, gratuitement sur l'Internet, à la disposition des professionnels de santé une base de données bibliographiques en français, BDBfr, concernant les principaux contenus de santé présents dans les journaux scientifiques, les encyclopédies et les ouvrages qu'ils soient en accès gratuit ou en accès payant, qu'ils proviennent de supports numériques ou de supports papier. BDBfr proposera des services à forte valeur ajoutée, notamment autour des technologies sémantiques et de la bibliométrie.

Pour réaliser cet objectif, nous avons réuni un consortium compact, avec deux équipes académiques spécialisées en e-santé (Rouen, Lille) qui se connaissent bien, qui ont des équipes pluridisciplinaires (sciences de l'information à Rouen & ergonomes à Lille) et trois industriels (un leader mondial de l'édition scientifique et deux TPE françaises ; l'une qui a développé des outils permettant d'évaluer la recherche de notre pays et l'autre qui est un leader français dans les dictionnaires électroniques, et qui proposera dans ce projet un moteur de recherche en texte intégral.

Ce projet a été déposé en mars 2013 au programme TecSan de l'ANR. Nous avons tenu compte des critiques, en ajoutant un partenaire industriel et en rééquilibrant le ratio académique/industriel.

1. OBJECTIFS SCIENTIFIQUES ET TECHNOLOGIQUES

La finalité de ce projet est clairement opérationnelle : fournir à la communauté des professionnels de santé un outil qui leur fait défaut. Nous nous assurerons tout au long du projet, dans une logique de conception centrée sur l'utilisateur, que cet outil est réellement adapté aux besoins de ses utilisateurs.

A notre connaissance, il n'existe pas de bases de données bibliographiques réunissant toutes les caractéristiques de BDBfr : spécialisée dans la santé, gratuite et en langue française. Les bases de données existantes sont soit payantes (comme Pascal issue du monde académique) soit anglophones (comme MEDLINE) ou encore trop généralistes (Google Scholar).

La cible principale de ce projet est clairement le professionnel de santé pour qui l'anglais est un frein, et qui souhaite restreindre sa recherche aux articles dans sa langue natale pour pouvoir s'appropriier l'information ou l'assimiler plus rapidement. Nous sommes conscients que les principales avancées de la science sont écrites en anglais, mais des revues de la littérature, des cas cliniques, des ouvrages continuent heureusement à exister dans notre langue. De plus, un travail récent a montré que la langue de publication avait peu d'influence sur la qualité des rapports d'essais cliniques randomisés dans une discipline donnée.

Une typologie d'utilisateurs est d'ors et déjà envisagée. Elle sera affinée par des études qualitatives.

- Médecins n'utilisant pas l'anglais dans leur pratique quotidienne, notamment les médecins libéraux, mais cherchant une information plus scientifique que celle disponible dans la presse généraliste.
- Médecins cherchant des réponses rapides à leurs questions. Il faut noter que les outils d'aide à la décision clinique au chevet du patient (Point-Of-Care Decision Tool) tel UPTODATE (URL : www.uptodate.com) n'existent pas en français. Il y a donc un intérêt à améliorer l'accès à une littérature de synthèse (dite secondaire) adaptée au contexte national et rapidement lisible.

- Professionnels mettant à jour leurs connaissances. L'outil créé sera un support du développement professionnel continu (DPC) pour aller chercher ou recevoir de nouvelles connaissances.

La société de l'information générée par l'Internet a déplacé le besoin. Dans un monde où l'information est devenue pléthorique et où chacun est devenu producteur d'informations (patients compris), la question qui se pose est non plus l'accès à cette information mais l'identification et la qualification des sources et la discrimination –on ne peut tout lire- en fonction de critères fiables (comme la provenance d'un article) et personnalisés (comme le choix d'une langue).

Au total, nous positionnons BDBfr clairement comme un **outil de santé publique, de formation continue professionnelle et d'éducation thérapeutique**.

Nous souhaitons intégrer dans BDBfr plusieurs services à forte valeur ajoutée notamment autours de mesures bibliométriques qui seront évaluées. D'autres services seront également développés : pertinence des résultats (utilisations de terminologies enrichies, tri adapté), liens contextuels vers d'autres bases de données, services d'alertes, accès facilité au texte intégral, compatibilité avec les logiciels bibliographiques, représentation territorialisée des résultats, moteur de recherche sémantique pour le texte intégral.

Le principal verrou scientifique repose sur le développement d'un moteur de recherche sémantique de qualité permettant de requêter en français sur de nombreuses métadonnées, en priorité l'indexation à l'aide d'un vocabulaire contrôlé, mais aussi le titre, les auteurs, les journaux, les résumés. Le point méthodologique central à résoudre sera la création d'un système d'information capable de gérer les données issues de PubMed, d'Elsevier Masson et de SIGAPS. Ce moteur de recherche sémantique s'appuiera sur celui développé depuis 2000 dans l'équipe CISMef pour rechercher au sein d'un catalogue de ressources provenant du Web santé. Ce moteur est déjà multi-terminologique. Il doit devenir générique, pour s'adapter au volume et à la nature de nouvelles métadonnées en provenance de sources hétérogènes (PubMed, Elsevier Masson et SIGAPS). Sa structure doit s'appuyer sur le Web sémantique pour gérer lexiques, synonymies, terminologies, ontologies formelles, bilinguisme franco-anglais voire multilinguisme dans le cadre d'un futur projet européen. Cet outil de recherche doit aider l'utilisateur en interprétant intelligemment ses requêtes en les confrontant à sa base de connaissances. Il doit permettre d'élargir ou au contraire restreindre sa stratégie de recherche, en plus de dépasser les questions d'orthographe ou de phonémisation. Il doit permettre une navigation transversale entre documents de thèmes connexes en se fondant sur une expansion sémantique ou statistique (matrice de cooccurrence). SenseGates aura pour rôle de transmettre à CISMef le code source de son moteur de recherche et de concevoir, avec CISMef, depuis ce code, plusieurs moteurs de recherche en texte intégral, depuis une recherche brute sur du texte et des ontologies, au démarrage du projet, jusqu'à une indexation et une recherche mettant en jeu des enrichissements réciproques entre textes et ontologies à la fin du projet.

La version initiale du produit développée lors de la première année du projet sera l'accès à une base de données bibliographiques contenant les citations en langue française de PubMed et les résumés en langue française provenant de Elsevier Masson.

La version intermédiaire numéro 2 correspondra à l'intégration des services pour accéder au texte intégral des articles, des liens contextuels vers d'autres sources d'informations, des services d'alertes et des formats compatibles avec les logiciels de gestion bibliographiques.

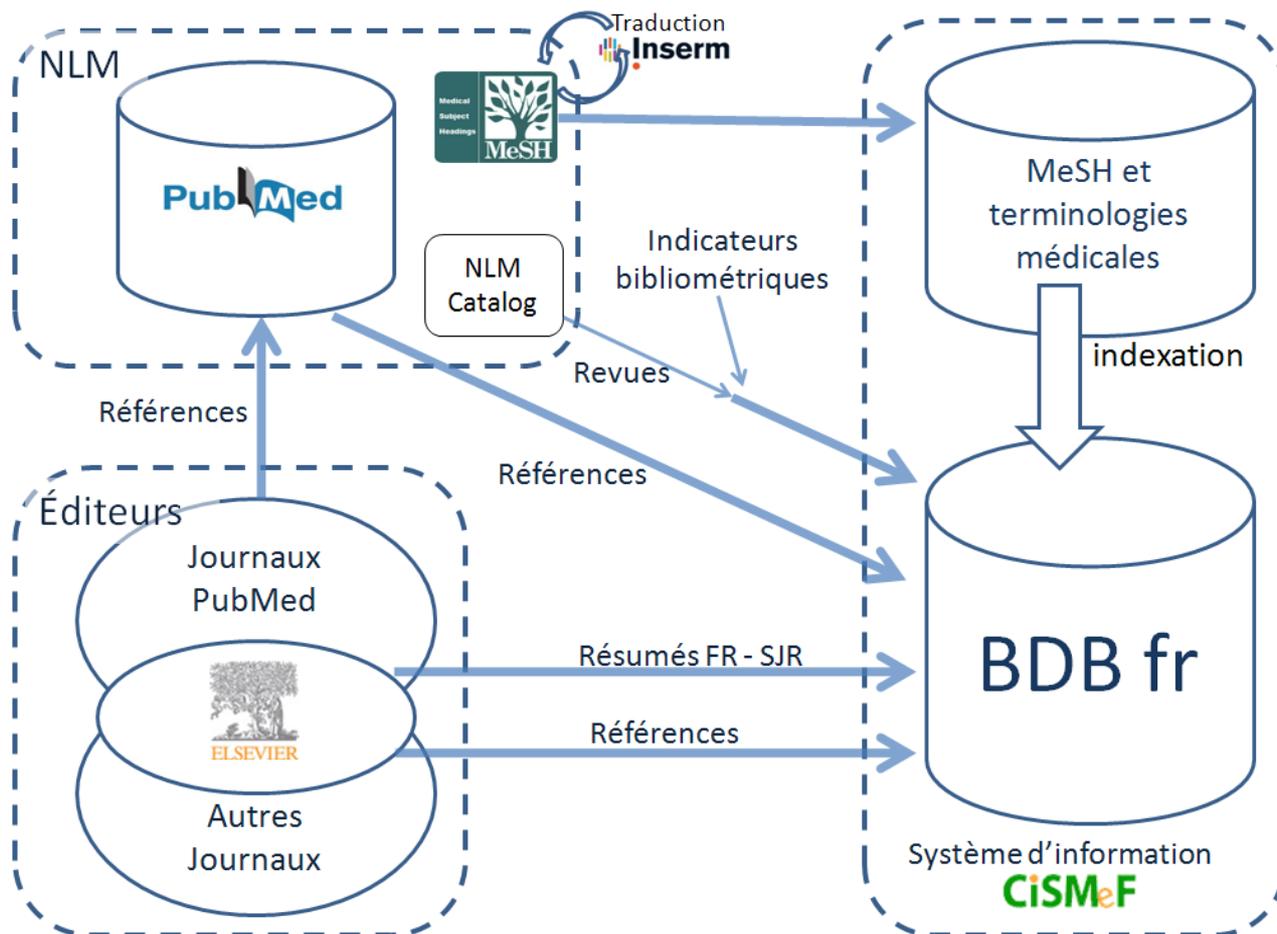
La version intermédiaire numéro 3 correspondra à l'ajout de critères bibliométriques, et en premier lieu le score SIGAPS, qui est devenu en moins de 10 ans un critère économique majeur dans les CHU français.

La quatrième version correspondra à l'extension de la base de données bibliographiques à des citations de contenus scientifiques (journaux, encyclopédies, ouvrages) non incluses dans MEDLINE, fournies dans le cadre de ce projet par l'éditeur Elsevier. Notre consortium est ouvert à toute source d'information provenant des principaux éditeurs nationaux et internationaux produisant des revues en français, mais aussi des revues indépendantes, qui sont partiellement présentes dans CISMef.

Parallèlement à ces versions une version mobile de BDBfr, adaptée aux appareils de type tablette et smartphone sera développée. En effet le taux d'équipement des médecins français et leur utilisation d'Internet depuis ces appareils est d'ores et déjà très important. Cette version proposera le maximum de fonctionnalités possible en tenant compte de l'interface tactile et de la puissance de ces appareils.

Une tâche sera dévolue explicitement au recueil des besoins et à l'évaluation de chacun de ces produits. Le critère de réussite principal sera le trafic de cette base de données bibliographiques/bibliométriques (BDBfr) qui sera disponible gratuitement et en accès libre sur l'Internet. Un critère secondaire sera le trafic secondaire sur les sites d'Elsevier Masson et de SIGAPS, ayant comme origine BDBfr.

La figure ci-dessous permet de résumer l'ensemble des flux d'informations servant à constituer BDBfr :



2. PERTINENCE ET CARACTÈRE STRATÉGIQUE DU PROJET AU REGARD DES ORIENTATIONS DE L'APPEL

En analysant l'existant, nous pouvons affirmer, qu'à notre connaissance, BDBfr deviendrait la première base de données bibliographiques/bibliométriques disponible gratuitement pour l'information scientifique en santé en langue française. Toute proposition gardée, notre objectif est de devenir le PubMed à la française. A la différence de PubMed États-unien qui est un projet fédéral, BDBfr bénéficierait s'il est financé par le programme TecSan d'un fort partenariat public-privé. Nous espérons qu'à la fin du projet plus de 1.000 personnes par jour utiliseront BDBfr, et ce de façon gratuite. A titre de comparaison, le catalogue CISMef est utilisé par 10.000 personnes par jour et le Portail Terminologique de Santé par 500 personnes par jour.

Les données générées par l'usage de BDBfr, leur traitement et leur analyse constitueront à moyen terme un « fonds de commerce » : qualification des requêtes, provenance géographique, identification des visiteurs (médecins hospitaliers ou de ville, autres professionnels de santé, grand public), pages et arborescences consultées, temps passé, etc.

Enfin, BDBfr permettra naturellement une augmentation du chiffre d'affaires des éditeurs de la presse scientifique en français, dont notre partenaire Elsevier Masson qui en représente plus des deux-tiers. Nous espérons grâce à la position d'Elsevier Masson pouvoir intégrer les autres éditeurs de la presse scientifique française, ainsi que les revues indépendantes.

Un succès opérationnel de BDBfr ouvrirait la porte à de possibles déclinaisons en Europe à plus long terme. En effet, étant donné que notre outil de recherche est en passe de devenir multilingue, il deviendrait possible, en suivant la même démarche que pour la langue française, de créer un outil adapté aux principales langues européennes, dont l'allemand, l'espagnol, le portugais et le norvégien grâce à nos contacts avec d'autres équipes européennes. La faisabilité d'une telle démarche a déjà été étudiée et nous espérons qu'à plus long terme elle prenne sa place dans un projet européen.

Notre proposition se positionne au sein du défi « Santé et bien-être » dans l'axe 2 « Améliorer la Santé par la médecine personnalisée, le diagnostic, la prévention et la thérapie, les stratégies palliatives, en concevant le vivant dans son environnement ». Comme souligné dans le plan d'action 2014, les ponts avec le défi « Société de l'information et de la communication » sont nombreux. En effet, c'est en proposant une base de données à la pointe des sciences de l'information que notre projet s'inscrit dans l'amélioration des soins.

Les caractéristiques de notre projet apparaissent dans le sous-axe 8 « Systèmes informatiques et numériques et données de Santé, phénotypage » et dans le sous-axe 18 « Technologies pour la Santé ». Le projet BDBfr inclut la collecte, l'organisation, la structuration, le partage et l'exploitation de données et d'informations, en fonction des usages (en ce qui nous concerne bibliographiques et bibliométriques) et aide au diagnostic comme à la prise de décision. BDBfr participera notamment au développement des outils technologiques pour faciliter, accélérer et accompagner la recherche et la formation dans le domaine général de la santé.

Ce travail ne fait pas suite directement à des précédents projets de recherche financés par l'ANR programme TecSan. Néanmoins, ce projet BDBfr s'appuie *indirectement* sur de nombreux projets précédents, dont l'équipe rouennaise a fait partie :

- InterSTIS, par la création d'un portail terminologique inter-lingue HeTOP ; projet dont Dominique Dutoit, dirigeant de SenseGates était partenaire (comme ingénieur à Mémodata).
- RAVEL, par le développement de nouvelles fonctionnalités à notre moteur de recherche sémantique, originellement dévolu aux ressources Web, et, dans ce projet, avec la finalité de rechercher des informations dans le dossier du patient informatisé.

D'autre part, ce projet s'appuie sur le score SIGAPS, qui a maintenant une véritable équipe opérationnelle portée par l'équipe lilloise et Alicante, et dirigée par la direction générale de l'offre de soins.

3. PRÉSENTATION DU CONSORTIUM

Ce projet fait suite à une longue collaboration entre les équipes d'informatique médicale (e-santé) de Rouen et de Lille ; depuis 12 ans, nous avons participé à de nombreux projets de recherche ensemble, qu'ils soient financés par les Ministère de la Recherche ou de l'Enseignement Français (projets COQUAS et UMVF) ou au niveau Européen dans le cadre du 7ème programme cadre (projet PSIP 2007-2011). L'équipe CISMef a participé avec EVALAB CIC-IT de Lille aux projets ANR TecSan L3IM qui vient de se terminer et sa suite SIFADO qui a démarré en janvier 2012. Nous souhaitons collaborer avec les trois industriels pour intégrer d'autres critères bibliométriques dans BDBfr. L'équipe Lilloise et Alicante collaborent depuis plus de 10 ans autour du projet SIGAPS.

Partenaire 1. Equipe CISMef, CHU de Rouen

L'équipe CISMef du CHU de Rouen est membre de l'équipe TIBS du laboratoire LITIS EA 4108 de l'Université de Rouen. Cette équipe est multidisciplinaire, composée de 16 ETP (Equivalent Temps-Plein), avec 1,5 médecins informaticiens, 3,0 documentalistes spécialisées en sciences de l'information, 4 ingénieurs de recherche, un assistant régional de recherche, un maître de conférences en informatique, un post-doc, 4 doctorants en informatique et un mi-temps d'assistante.

Depuis son origine en février 1995, CISMef a développé le catalogue éponyme pour Catalogue et Index des Sites Médicaux de langue Française (URL : www.cismef.org). Nous avons utilisé pour sa création à deux standards principaux : le Dublin Core pour la description des métadonnées et le thésaurus MeSH pour l'indexation des ressources Web. Depuis 2005, nous sommes à notre connaissance la première équipe à avoir développé un catalogue de santé qui reposait sur un univers multi-terminologique tant pour l'indexation que pour la recherche d'information. Nous avons commercialisé une licence du droit d'usage du catalogue CISMef auprès de l'industrie pharmaceutique (GSK, Lilly, Vidal). Depuis 2007, notre équipe a investi plus de 14 années-hommes pour développer PTS le portail terminologique de santé (voir plus haut), accessible en français et en anglais. Enfin, intégré au PTS (service à la demande), nous avons également développé InfoRoute, un « infobutton » à la française (URL: <http://inforoute.chu-rouen.fr/>)

Référence : Darmoni SJ, Soualmia LF, Griffon N, Grosjean J, Kerdelhué G, Kergourlay I, et al. Multi-lingual Search Engine to Access PubMed Monolingual Subsets: A Feasibility Study. Stud Health Technol Inform. 2013;192:966.

Partenaire 2. Equipe d'Accueil 2694 : "Santé Publique : épidémiologie et qualité des soins"

L'équipe d'accueil EA2694 : "Santé Publique : épidémiologie et qualité des soins" est une équipe pluridisciplinaire regroupant des méthodologistes et des ergonomes dans les domaines des biostatistiques, de l'informatique médicale, de l'économie de la santé et de l'ergonomie. Cette équipe de recherche de l'Université Lille 2 et du Pôle de Recherche et d'Enseignement Supérieur Université Lille Nord de France est rattachée à l'Institut de Médecine Prédictive et de Recherche Thérapeutique (IFR 114) de la Faculté de Médecine. Le CIC-IT / EVALAB est notamment dédié à l'analyse des contextes d'usage et l'évaluation des technologies de Santé pour une optimisation des environnements de santé.

Référence :Beuscart-Zéphir MC, Borycki E, Carayon P, Jaspers MWM, Pelayo S. Evolution of human factors research and studies of health information technologies: the role of patient safety. Yearb Med Inform. 2013;8(1):67-77.

Partenaire 3. Elsevier Masson

Elsevier Masson est issu du rapprochement, fin 2005, des Éditions Elsevier France, filiale de la division Health Science du leader mondial de l'édition médicale et scientifique Elsevier, et des Éditions Masson. De cette association est né l'éditeur francophone médical et paramédical de référence : plus de 10 000 auteurs et rédacteurs, 130 revues publiées chaque année (environ 70% de la presse médicale scientifique), 47 traités EMC, un catalogue riche de 1 500 livres... Intégré à Elsevier International, Elsevier Masson d'une diffusion internationale large notamment grâce à ScienceDirect® Elsevier Masson a mis en place un workflow de contenus structurés et indexés. Ceci permet de disposer d'une base de données exhaustive en contribuant de manière décisive à la qualité des produits et services. Ce renversement de modèle économique est au cœur des développements de la société qui adapte, en conséquence, ses structures et ses modes de travail avec ses partenaires. Cet engagement numérique permet à Elsevier Masson d'être le leader de l'édition en ligne dans le domaine médical et paramédical en France, à travers plusieurs des plates-formes interactives. Au fil des années, Elsevier Masson a su tisser des relations privilégiées avec les sociétés savantes, les collèges des spécialités médicales et les associations en assurant une large diffusion de leurs ouvrages, vecteur de visibilité, et en leur fournissant des services dédiés (actes, référentiels...).

Partenaire 4. Alicante

La société ALICANTE est l'actuelle concessionnaire du développement de SIGAPS pour la période 2013-2015. Elle réalise les développements de SIGAPS et de SIGREC (outil de gestion de la recherche clinique au niveau d'un établissement), en collaboration avec l'équipe opérationnelle SIGAPS/SIGREC, financée par le Ministère, et située au CHRU de Lille. La société ALICANTE, au-delà de SIGAPS et de SIGREC, a une activité de services autour des systèmes d'information de la recherche, en particulier de la recherche clinique, pour des clients hospitaliers et universitaires, centres de lutte contre le cancer, éditeurs de logiciels. La société ALICANTE s'apparente à une SSSL (société de services en logiciels libres).

Partenaire 5. SenseGates

Sensegates est une jeune entreprise innovante depuis 3 ans dont le dirigeant, Dominique Dutoit, est un spécialiste -Docteur, HDR-, en Traitement Automatique des Langues. Sensegates édite le site web sensagent.com qui est un dictionnaire électronique multilingue connu sur le Web (entre 200 000 et 500 000 visiteurs par jour).

Pour maintenir ou accroître sa position, Sensegates a eu besoin de développer un moteur de recherche 1) simple à installer, 2) aisé à utiliser, 3) peu coûteux à administrer et 4) et dont les parties sont aussi bien des unités atomiques, comme des mots, que des unités complexes, comme dans le cas des mots composés, de sous-documents, de grammaires, d'arbres ou de graphes. Si les critères 1 à 3 sont essentiels pour l'entreprise qui ne peut se permettre les coûts élevés récurrents d'un moteur de recherche courant, le critère 4 est pour sa part fondamental. En effet, si la maîtrise de la recherche d'un tout depuis des parties facilite et rend utile l'application des dictionnaires, sous-documents, grammaires, ontologies et documents sur eux-mêmes, l'expérience montre que sa non-maîtrise la rend infiniment coûteuse en coût de travail et en durée de calcul.