BILAN DU PROJET COLLEX DOPABAT

Introduction

Objectifs du projet

Ce projet original et expérimental a pour finalité de qualifier la place des thèses dans la production scientifique (leur audience et leur citation) dans les domaines de la physique, l'astrophysique et l'astronomie.
L'enjeu est également de pouvoir mesurer l'interdisciplinarité des publications et donc de mesurer à partir du classement thématique des revues scientifiques, la part des publications dans le domaine de la physique publiées dans d'autres revues, et la part des références issues d'autres disciplines.
Ce projet est né de la demande de chercheurs de nos établissements respectifs désireux de travailler sur une exploitation fine des publications de leur domaine scientifique.

Point de départ du projet

A l'origine, ce projet était envisagé comme un POC (proof of Concept). Cependant nous avons dépassé ce stade car l'application est désormais fonctionnelle et disponible sur le web. Elle est progressivement communiquée aux chercheurs susceptibles de l'utiliser.

Une application originale

L'application est une application web hébergée sur un serveur distant.

Elle est codée en langage R et déposée sur un serveur Shiny. Ce langage open source gratuit, utilisé par une large communauté, est très adapté à la gestion de données et la réalisation de statistiques. De plus, l'équipe de développement avait déjà une expérience approfondie de l'utilisation de ce langage.

Voici l'adresse de l'application :

https://uga-projet.shinyapps.io/applitodeploy/

Les principales fonctionnalités de DOPABAT

Les imports

L'utilisateur fournit un corpus de publications. Ce sont les métadonnées de ce corpus qui seront étudiées par l'application. Il est possible d'importer dans l'application des fichiers de différents formats.

- Les fichiers csv peuvent avoir des structures différentes et des formes d'encodages diverses. Une visionneuse permet à l'utilisateur de vérifier si son import fonctionne ou non.
- Les fichiers bibteX sont plus faciles à intégrer car ils ont une structure relativement simple et assez similaire d'un fichier à l'autre.

L'import dans l'application n'est pas limité à un seul fichier. L'utilisateur peut créer son propre corpus via plusieurs sources de données. Toutes ces données sont ensuite agrégées et envoyées vers des API pour être analysées.

Il est à noter que dans certains cas, le fichier peut ne pas être traité par l'application, par exemple quand le fichier présente une irrégularité dans sa structure.

PARTIE 1 METHODOLOGIE

1- Le choix des sources et la réalisation des API

Pour commencer, notre choix s'est porté sur les bases de données "sources" que nous avons choisies pour les informations qu'elles contiennent mais aussi pour leur spectre disciplinaire. L'objectif est d'obtenir des résultats sur des contenus de diverses origines.

Pubmed est une base de données bibliographique dans les domaines de la médecine, de la biologie et du biomedical, susceptible donc d'intégrer des articles relevant de la physique.

ADS est une base bibliographique développée par la NASA et recensant principalement des publications en physique et astrophysique.

LENS est une base de données à large spectre composée de plus d'une douzaine d'autres bases de données sur des sujets très hétérogènes.

Il est important de noter ici que l'accès à ces bases de données n'est pas fourni par l'application DOPABAT. Pour les APIs le demandant, c'est à l'utilisateur de fournir son accès (token).

Les API retenues pour l'application web DOPABAT sont donc celles d' ADS, Pubmed et LENS.

Chacune des API, sur lesquelles DOPABAT va travailler, a ses propres particularités qui ont été prises en compte dans le programme : différence des modes de requêtage, ou des formats des auteurs et des titres, ou encore de la limitation des caractères.

L'API de Pubmed a posé problème. Elle possède, en effet, un correcteur automatique de requêtes. Lorsque l'utilisateur fait une requête il peut enrichir les résultats obtenus grâce à une proposition de mots clés "similaires". Mais la correction peut avoir un impact sur les auteurs notamment et l'on risque, de ce fait, de n'obtenir aucun résultat.

En outre, la vérification de résultats se fait au prix d'un temps d'exécution plus long.

L'API de LENS contient les données de plusieurs bases, dont Pubmed. Si l'utilisateur possède un identifiant LENS (même gratuit), il pourra quand même avoir les données qui sont issues de Pubmed.

2 - D'autres pistes : ArXiv et INSPIRE

L'équipe a également envisagé d'exploiter l'archive ouverte Arxiv.

Son intérêt s'est avéré toutefois limité : nombre limité de requêtes simultanées , temps d'exécution trop long…

De plus, l'API ne fournit pas les références et les citations pourtant présentes sur le site web.

Enfin l'exploitation des données est rendue complexe par l'inexistence de nombreuses affiliations de laboratoires dans ArXiv.

Une autre option a été de passer par INSPIRE pour obtenir des références d'articles mais son site web a évolué et la récupération de données s'est finalement révélée impossible.

L'équipe a donc été contrainte d'abandonner ces options au profit de la base de données LENS.

PARTIE 2 LES FONCTIONS DE DOPABAT

1 - La fonction de contrôle de DOPABAT

Pour créer l'application DOPABAT, nous nous sommes basés sur plusieurs jeux de données de publications provenant principalement du domaine de la physique.

Or, les titres des articles peuvent contenir des symboles mathématiques, qui ne sont pas reconnus par les API de type url, soit la majorité des API du projet. Il peut également arriver que lors d'une requête portant sur plusieurs publications, l'API restitue une publication proche de celle demandée mais toutefois différente du titre voulu. Dans ce cas on parle de faux positif.

Pour éviter cela, nous avons mis en place un programme de vérification entre les titres demandés et la réponse apportée mais conservant les mêmes auteurs.

Si le titre obtenu s'éloigne à plus de 15% du titre original, la publication est rejetée. Si la publication s'éloigne entre 5% et 15% , elle est placée dans un jeu de données "doutes" qui devra être validé par l'utilisateur avant d'être analysé.

Si il y a moins de 5% de différence, on considère les données comme correctes.

Le format de réponse est le format json. Les éléments sont obtenus sous forme de grandes listes qu'il convient de transformer sous forme de tableau. Évidemment, chaque API ne retourne pas le même type d'informations mais nous avons identifié un noyau commun à chaque API qui nous permet de faire les analyses que l'on souhaite; on retiendra donc : identifiant, titre, auteur, dates, identifiant de citation, identifiant de références, journal. Pour

rappel, la citation est la publication qui cite une autre publication. La référence est une description bibliographique de la publication reprenant la source, le titre...

2 - La fonction analytique de DOPABAT

Dopabat permet la visualisation des données chargées dans l'application. mais également des données retournées par les API, telles que les références, les citations voire les erreurs.

Une fois l'analyse du corpus terminée, l'utilisateur peut récupérer les données sur lesquelles elle a porté.

Outre les données brutes, l'utilisateur a aussi accès à de nombreux graphiques. L'utilisateur peut disposer d'une analyse descriptive des mots clés et des domaines du corpus. Cela se fait par l'intermédiaire d'un nuage de mots-clés paramétrable (pour une analyse quantitative par années) et de plusieurs réseaux de mots appelés network (analyse en lien et en quantité par années).

Cela permet assez rapidement à l'utilisateur de se faire une idée des thèmes abordés par son corpus. Enfin l'utilisateur a aussi accès au graphique précisant l'interdisciplinarité du corpus, à partir des citations ou des références.

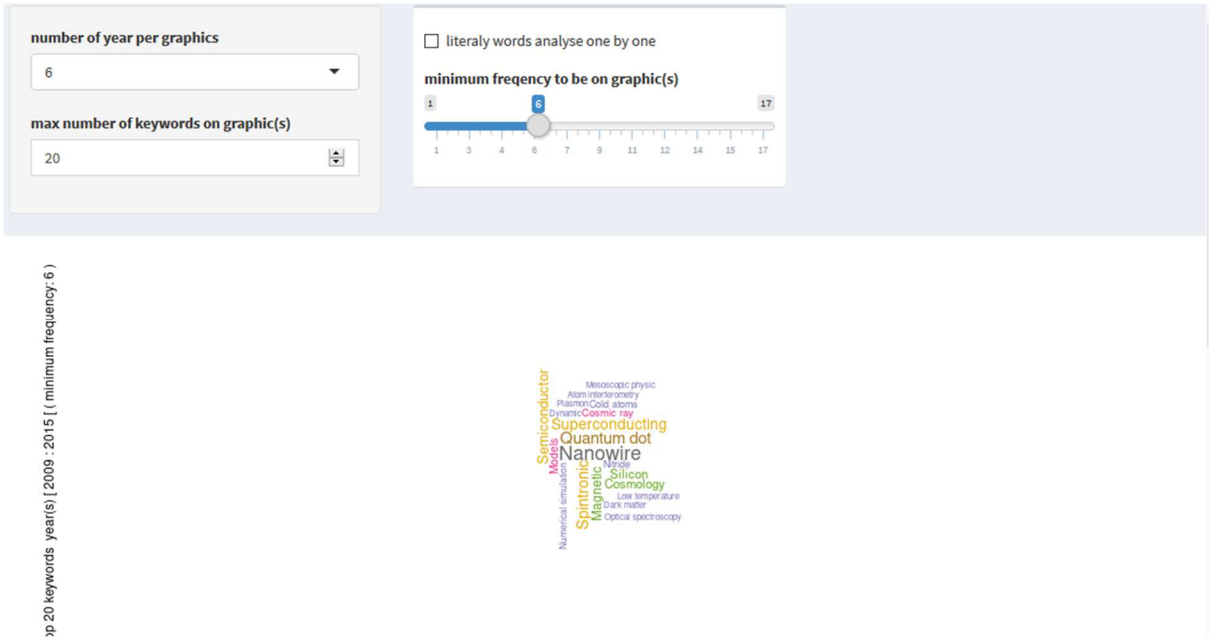
3 - La fonction Visualisation

Au fur et à mesure que l'utilisateur "valide" l'importation de ses fichiers de données, sont générés des graphiques évolutifs.

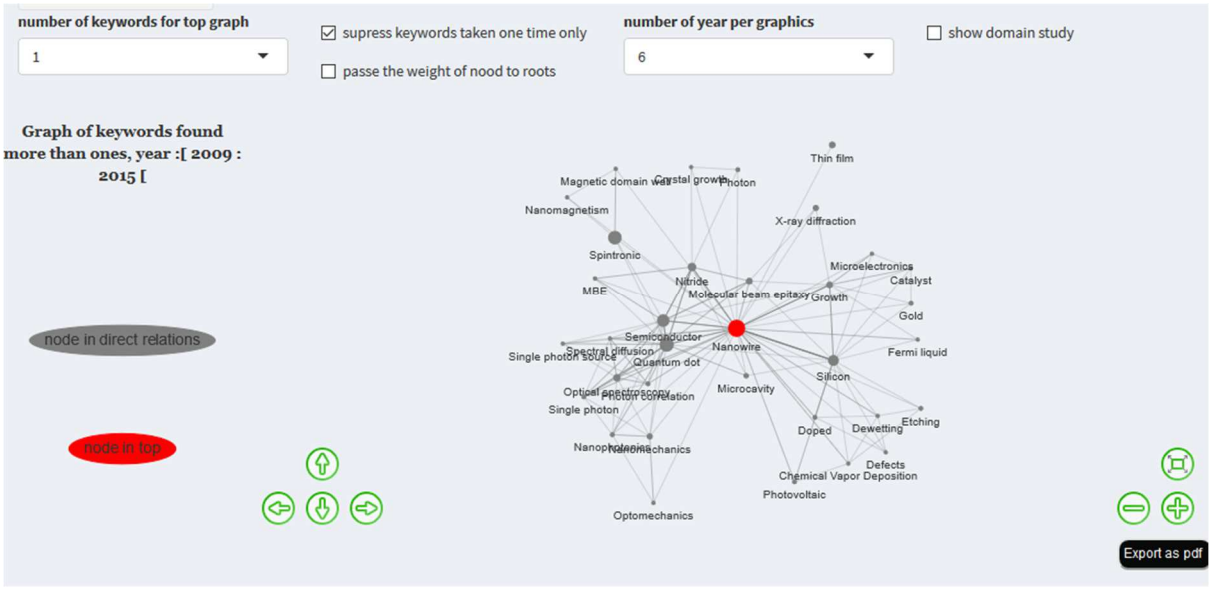
Sur l'application, chaque partie de l'analyse est divisée en onglets. Des messages en mode "pop up" apparaissent à l'utilisateur pour le guider dans les différentes visualisations proposées.

Les graphiques des mots-clés, identifiés par l'utilisateur, font naître un nuage de mots. L'utilisateur a à sa disposition plusieurs paramètres pour faire varier la taille, le nombre ou encore le type d'analyse des mots-clés en fonction du temps (de la colonne date).

L'utilisateur peut faire varier le nombre de mots sur le graphique, la fréquence nécessaire pour l'apparition d'un mot et il peut définir un nombre d'années spécifique pour un graphique



Autre forme de graphique proposé : des réseaux qui mettent en évidence les mots-clés utilisés dans la ou les mêmes publications. En fonction de la taille des nœuds (représentant la fréquence du mot utilisé comme mot-clé dans le corpus) et de la fréquence du lien, l'utilisateur peut avoir une idée des thèmes abordés dans son corpus. Il est également possible de faire varier le nombre de graphiques afin de mettre en valeur une période donnée.



Une relation entre deux mots se crée lorsque ces deux mots-clés sont présents dans la même publication. Les mots-clefs les plus présents dans ces relations constituent des nœuds.

Plus le nœud est important, plus le mot-clé revient souvent. Plus le lien entre les mots est foncé, plus la relation est fréquente.

A tout moment il est possible d'exporter le ou les graphiques en format pdf

4 - Analyse des références et citations

Nous nous sommes attachés à identifier des citations ou des références des articles pour pouvoir visualiser l'interdisciplinarité d'un corpus de publications.

Pour un corpus, l'application analyse :

- les références des articles pour identifier les domaines auxquels elles appartiennent
- les citations des articles pour identifier à quel domaine appartiennent les revues citantes

L'équipe a construit un fichier de 20 000 titres de revue qu'elle a rattaché au domaine tel que défini par l'OST.

5 - La mesure de l'interdisciplinarité

Pour procéder à l'analyse de l'interdisciplinarité, l'application va interroger via les API, les bases ADS, LENS ou Pubmed.

L'application va utiliser l'un des identifiants de l'article (titre/auteur ou DOI) pour interroger via une ou plusieurs API une des 3 bases qui renverra les informations, si cette publication est bien présente dans la base. Puis l'API va ensuite récupérer, selon la demande de l'utilisateur, les références et les citations relatives à la publication en question.

Chaque revue étant rattachée à une ou plusieurs catégories OST, on récupère les domaines des journaux concernant les publications citantes ou les références grâce au fichier des 20000 titres de revues. Cela nous donne une représentation de la distribution des disciplines au sein de la publication (dans la limite du pourcentage de journaux retrouvés dans le fichier).

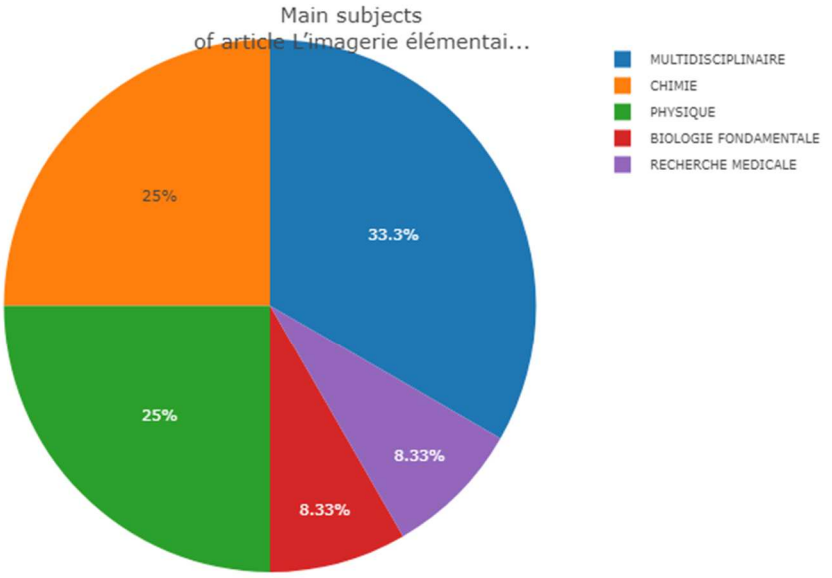
Si l'on travaille sur un corpus entier, le processus peut être évidemment plus long.

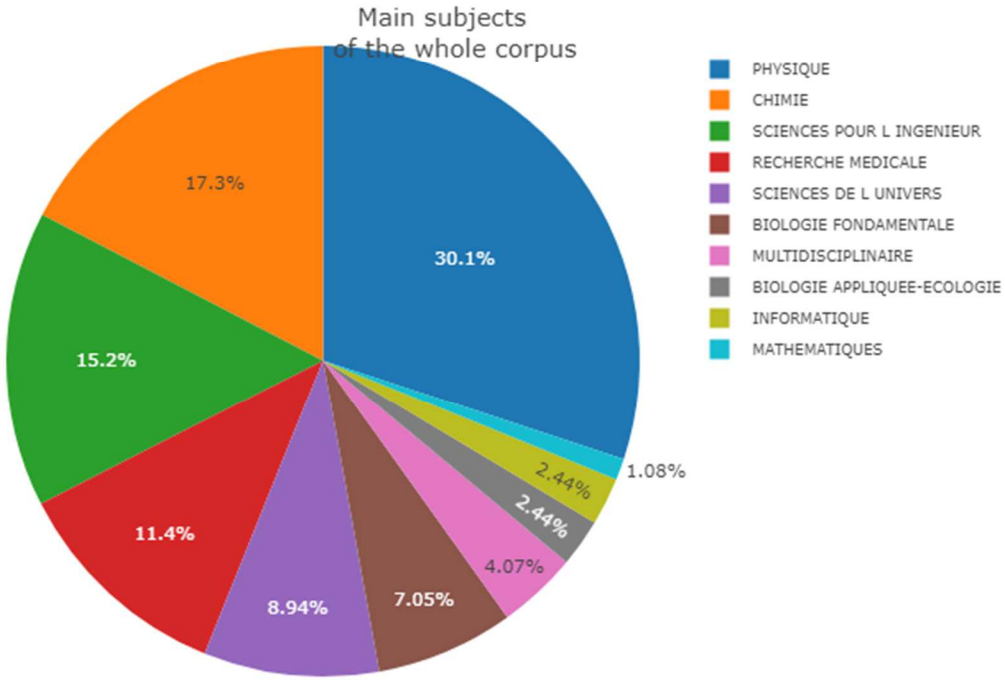
Dans l'exemple ci-dessous, on peut observer sur le graphique les différents domaines d'un article à gauche (choisi en amont par l'utilisateur) et, à droite, les principaux domaines du corpus entier. Ainsi l'utilisateur peut comparer les deux.

Au-dessus de chaque graphique, l'indicateur statistique nous montre le pourcentage de journaux retrouvés. Plus cet indicateur est élevé, plus le graphique est représentatif. Ici la représentation est relativement bonne.

Les graphiques sont complètement interactifs; en cliquant sur une partie, l'utilisateur aura accès aux données liées, il peut activer ou désactiver certaines disciplines pour voir en détail les catégories les moins représentées du graphique.

Cela permet à un chercheur d'identifier, via les références, les disciplines qui ont été mobilisées et la diversité des apports pour produire l'article étudié. Il peut également, grâce aux citations, mesurer l'audience d'une publication et son rayonnement dans différents domaines scientifiques. Ce processus permet donc de caractériser à deux niveaux l'interdisciplinarité d'une publication (ou d'un corpus).





PARTIE 3 LE PROJET AU FIL DE L'EAU

1 - Le blog

Le blog de suivi du projet a permis à l'équipe de tenir un véritable carnet de bord consignant les étapes, les avancées, les difficultés rencontrées durant toute la période. Au fil des articles, les chercheurs peuvent ainsi découvrir le développement du projet DOPABAT.

Par ailleurs, une documentation en ligne et un guide de l'application, en cours de rédaction, seront proposés aux chercheurs.

2 - Les difficultés rencontrées

Tout au long de la première année de ce projet, l'équipe a également pris le temps d'analyser les difficultés rencontrées. Celles-ci ont notamment concerné les conditions d'interrogation des bases et l'exploitation des données.

- L'exploitation de l'API du Web of Science nécessitait de prendre un abonnement supplémentaire payant.
- Il était impossible d'importer des résultats en format csv depuis le Web of Science notamment.
- La granularité des différentes bases était très diverse, certaines couvrant des domaines scientifiques trop généraux. L'incohérence des résultats obtenus par certaines des requêtes a nécessité d'apporter des correctifs.

L'équipe DOPABAT a eu conscience, durant cette première année, que le projet expérimental dans lequel elle s'était lancée, allait également se heurter à des écueils.

Cela ne l'a pas empêché à chaque fois d'imaginer des solutions pour contourner ces problèmes. Par exemple, l'équipe a préféré s'appuyer sur des outils solides (ex. la base ADS) sans perdre du temps sur ceux qui s'avéraient trop complexes (ArXiv).

Cette expérience a donné à l'équipe l'occasion de se poser des questions sur le contenu des bases (par exemple, comment les bases bibliographiques intègrent-elles les thèses ?) et, au-delà, sur la production scientifique elle-même. Ne pas retrouver une thèse dans une base signifie-t-il qu'une thèse n'est pas systématiquement citée par son titre ? Existe-t-il d'autres pratiques ?...

Enfin, les membres de l'équipe ont eu à cœur de chercher à construire les bonnes requêtes d'interrogation des bases, d'élaborer des méthodologies pas à pas, de ne pas prendre pour acquis les premiers résultats obtenus mais devoir parfois les retravailler…

L’exécution du projet, notamment dans la première phase, avant les tests d’imports, a demandé à l’équipe un contrôle des métadonnées, un regard comparatif et critique des outils disponibles, notamment pour le choix des API.

Au milieu du projet, le langage source R a passé une version majeure. Cela a eu des incidences sur les fonctions de base du langage, nous obligeant à mettre tous nos programme à jour, à repenser certaines fonctions et à vérifier la quasi totalité du code déjà construit.

3 - Retour des chercheurs

L’application a été présentée à quelques chercheurs pendant la phase de développement. Elle a recueilli leur intérêt, sur la bibliométrie des thèses, l’évolution des mots-clés et sur l’interdisciplinarité.

Un ancien directeur d’école doctorale de physique a demandé à l’équipe, une fois que seront finalisés les développements, de produire des éléments précis concernant la citation des thèses (sur quelle temporalité, selon quels thèmes, etc.). Ces indications peuvent en effet nourrir la réflexion des directeurs d’écoles doctorales et des directeurs de thèse.

Au-delà de l’équipe projet, qui s’est consacrée aux premiers tests, des chercheurs vont être sollicités ces prochains mois pour tester l’application afin d’identifier les derniers écueils à résoudre ou évolutions à prévoir.

Il est enfin prévu une campagne d’information autour de Dopabat dans la communauté, afin de faire connaître ses fonctionnalités et d’accompagner son usage par les chercheurs et les équipes de pilotage des laboratoires.

CONCLUSION

Quelle suite donner au projet ?

Le projet Dopabat se poursuit avec de nouveaux tests sur des jeux de publications. L'ergonomie et l'habillage de l'application sont également travaillés par l'équipe.

Cette première phase nous a permis de pousser notre raisonnement un peu plus loin, et de nous lancer dans une suite à ce projet : Alt Dopabat.

Alt Dopabat vise à étudier la vie d'une publication de son premier jet (preprint) à sa version éditeur lorsqu'elle existe. L'intérêt est ici d'examiner non plus seulement la vie d'une publication dans l'univers académique mais de l'étudier au sein des réseaux sociaux « grands publics ».
Cette étude permettra d'étudier les impacts médiatiques, académiques et sociaux d'une publication en fonction de son statut - preprint, postprint et version éditeur – au cours du temps.

Une riche expérience pour l'équipe

En conclusion, le projet au long cours mené jusqu'à présent, a, d'une certaine manière, permis de valider le choix de l'équipe de travailler sur les corpus de thèses, type de document moins exploité et moins bien appréhendé que les autres publications.

L'équipe a pu obtenir satisfaction au bout d'une année de projet, en dépit des écueils cités plus haut.

Ainsi l'expérimentation a fait ressortir de nombreux points forts du projet DOPABAT.

Le défi de travailler sur plusieurs bases à la fois a pu être relevé, sur un périmètre relativement large.

L'atout majeur de l'application est d'offrir une bonne visualisation de l'interdisciplinarité, des citations. Les outils de visualisation sont un véritable plus dans le projet.

L'objectif d'obtenir un outil interactif et en accès libre a été atteint.

Annexes

Calendrier

Janvier 2019 : ouverture d'un blog de projet, mis en production sur WordPress, afin de faire connaître l'avancée de DOPABAT : **https://dopabat.inist.fr/**

Février 2019 : extraction d'un corpus de 1085 thèses soutenues entre 2009 et 2018, sur périmètre de l'UGA et de l'Observatoire, et issues de TEL.

Chargement du corpus traité dans l'outil Lodex par l'INIST
Objectif : publier des jeux de données et fournir un rapport web dynamique (tableau de bord et graphiques).

Mars-mai 2019 : création de requêtes manuelles pour l'exploitation du corpus sur les bases ADS (Astrophysics Data System), le Web of Science (WoS) et ArXiv et tests pour vérifier la citation des thèses dans ces bases
Objectif : vérifier la présence de thèses dans les bases bibliographiques et leurs citations par d'autres publications.

Eté 2019 : début de développement d'un programme informatique en langage R, servant à interroger les bases de données et récupérer des données statistiques qui pourront révéler des informations sur l'interdisciplinarité. Cet outil est libre, évolutif et pleinement maîtrisé par le contractuel recruté pour le projet.

Automne 2019 : travail théorique (méthodologie) sur les citations et l'interdisciplinarité d'un corpus.

Début d'intégration d'une interrogation des bases par API afin de mesurer les citations des publications du corpus après une analyse de ces données. Analyse de la cohérence des résultats obtenus.
Objectif : identifier les liens entre les publications et les thématiques communes.

Printemps 2020:

Poursuite du développement du programme qui interroge les différentes bases de données, notamment Arxiv er ADS.

Mise à disposition de l'application shiny sur un serveur R.

Eté 2020

Tests sur l'application en ligne, corrections de bugs et évolution de l'interface et des paramétrages.

Automne 2020:

Mise à jour de l'interface et du code source.

Ajout du format Bibtex et spécialisation de l'import des fichiers WoS

Ajout de Lens comme base de données interrogeable pour l'interdisciplinarité.