ANALYSES & VISUALISATIONS -

API HAL / RSTUDIO

Remaissa BENDIB - Hanjoon KO - Gracia YAN

M2 Information Communication, Données et Société
Université Paris Nanterre

Mars 2024

Sommaire

Le projet

Dans le cadre du cours de Gestion de projet dispensé en M2 Information Communication, nous travaillons avec la **PUDN** (Plateforme Universitaire des données numériques).

Ce projet, en partenariat avec la Fondation Deniker, porte sur la **maladie d’Alzheimer.**

En analysant les données présentes sur HAL, nous établissons un état des lieux des contenus, du référencement existant des publications et proposons une méthodologie à destination des étudiants.

Notre objectif est de montrer les possibilités de HAL mais aussi ses limites, en guidant les étapes d’extraction de données à partir d’une API jusqu’à la visualisation de ces données, tout en passant par le nettoyage, le traitement, l’analyse.

Les outils utilisés

Dans cette présentation, nous mobiliserons plusieurs outils :

- **L’API de HAL** (page 6)

Pour récolter toutes les données nécessaires à notre projet.

- **RStudio**

Pour traiter, analyser et visualiser ces données.

Vous pouvez retrouver toute la méthodologie de ce projet et avoir accès aux codes **ici**.

Présentation HAL

Créée en 2001 à l’inititative du CNRS (Centre National de la Recherche Scientifique), HAL est une **plateforme numérique pluridisciplinaire** qui vise à centraliser des dépôts scientifiques (articles, thèses, rapports...) réalisés par des chercheurs afin d’en faciliter la **diffusion**, et le **partage**.
La plateforme étant une **archive ouverte** commune, elle valorise la **collaboration** et le bien commun.

Les chercheurs peuvent publier eux-mêmes sur le site. Les contenus ne sont donc pas évalués en amont et sont archivés sur HAL que si une personne fait la **démarche** de publication.

C’est donc une immense **base de données**, alimentée par des chercheurs de toutes les disciplines, peu importe leur zone géographique, même si le portail n’est disponible qu’en français et en anglais.

Extraction des données

Qu’est-ce qu’une API ?

“Une API (application programming interface ou « interface de programmation d’application ») est une interface logicielle qui permet de « connecter » un logiciel ou un service à un autre logiciel ou service afin d’échanger des données et des fonctionnalités.” (CNIL)

Pour résumer, l’API de HAL nous permet, entre autre, de **connecter** la **base de données** à notre **interface de travail** (pour notre cas RStudio). L’avantage d’une API est la mise à jour **automatique** de l’ensemble de la base de données dès qu’on le souhaite, sans avoir à faire une extraction manuelle avec les nouveaux contenus régulièrement.

Extraction des données

Par exemple, notre base de données comprend l'ensemble des résultats de notre **requête** à partir du mot "Alzheimer". Cependant, de nouvelles publications sont déposées sur HAL régulièrement. Il serait assez fastidieux et chronophage d'extraire tous ces résultats, manuellement, plusieurs fois par mois pour avoir un corpus complet et à jour.

L'API nous permet donc de mettre à jour cette base de données et de récolter les **nouvelles publications récentes** directement lorsqu'on travaille sur RStudio.

Comment utiliser l'API de HAL ?

Il est fortement utile de lire la documentation propre à chaque API. Les étapes sont expliquées à partir de la ligne 53 du code.

Nettoyage

1. Étiquetage des variables

2. Étiquetage des modalités

3. Préparation des datasets

4. Exploration post-nettoyage

1. Étiquetage des variables

Les noms des variables du dataframe sont identiques aux champs sélectionnés lors de l'export depuis l'API. Pour une meilleure compréhension et lisibilité, nous ajoutons des étiquettes informatives à chaque variable.

“docid” remplace ‘Identifiant HAL du dépôt”.

2. Étiquetage des modalités

De la même manière, pour rendre les modalités plus explicites et compréhensibles, nous attribuons des libellés plus descriptifs pour les types de documents :
“Article dans une revue” remplace “ART”

Nous renommons les langues pour avoir le mot en entier :
“Allemand” remplace “de”

Pour les domaines, nous leur attribuons des étiquettes et nous les agrégeons pour créer des catégories plus larges :
“Informatique” remplace “info”

3. Préparation des datasets

a) Préparation d’un dataset général

Pour certaines de nos analyses, nous nous intéressons à l’ensemble des données afin de visualiser des tendances générales. Cependant, ces recherchent portent sur le domaine des Sciences Humaines et Sociales (SHS). Grâce à notre nettoyage précédent, nous pouvons chercher et sélectionner uniquement les sources catégorisées comme SHS ou “SSH” (en anglais) dans la variable de domaine.

Puis, nous agrégeons les types de publications pour rendre nos visualisations plus lisibles par la suite.

“ISSUE” et “ART” sont regroupés dans la catégorie “Article dans une revue”.

3. Préparation des datasets

b) Préparation d’un dataset spécialisé pour les revues

L’API nous donnant toutes les informations des colonnes sélectionnées, nous avons des contenus qui ne nous intéresse pas pour notre recherche. Afin de se concentrer seulement sur les informations relatives aux revues SHS, nous créons un sous dataset.

Grâce à notre agrégation précédente, nous avons toutes les données des articles de revues réunies.

3. Préparation des datasets

b) Préparation d’un dataset spécialisé pour les revues

Les publications sont répertoriées dans différentes langues selon la langue du texte détectée. Cependant, de nombreuses erreurs apparaissent. Pour pallier ces confusions et pouvoir établir une répartition plus juste des contenus, nous devons comparer si la détection de la langue correspond à ce que nous indique la colonne de la langue. Si non, nous nettoyons à la main.

Nous rencontrons un obstacle pour les publications incluant :

- un titre en 2 langues différentes (ex: titre français mais aussi écrit en agnalis à la suite)
- un titre qui contient des mots de langue étrangère au reste du titre (ex : le terme “care” utilisé dans un titre français)

3. Préparation des datasets

b) Préparation d’un dataset spécialisé pour les revues

Enfin, un nettoyage est nécessaire sur les éditeurs. En effet, l’éditeur “Elsevier” apparaît sous 2 formes : “Elsevier” et “Elsevier Masson”. Nous les regroupons sous un même terme.

4. Exploration post-nettoyage

Dans cette section dédiée à l'analyse des fréquences, nous allons explorer nos données nettoyées. L'objectif de cette étape est de mieux comprendre la distribution des valeurs dans certaines variables clés. En créant des tableaux de fréquence, nous pourrons identifier les tendances, les modalités les plus courantes, et détecter d'éventuelles anomalies.

Nous regardons la fréquence des types de publications, de la langue et des domaines.

Analyses

1. Tendances annuelles de publications

2. Analyse de la répartition des publications SHS

3. Top des revues les plus actives

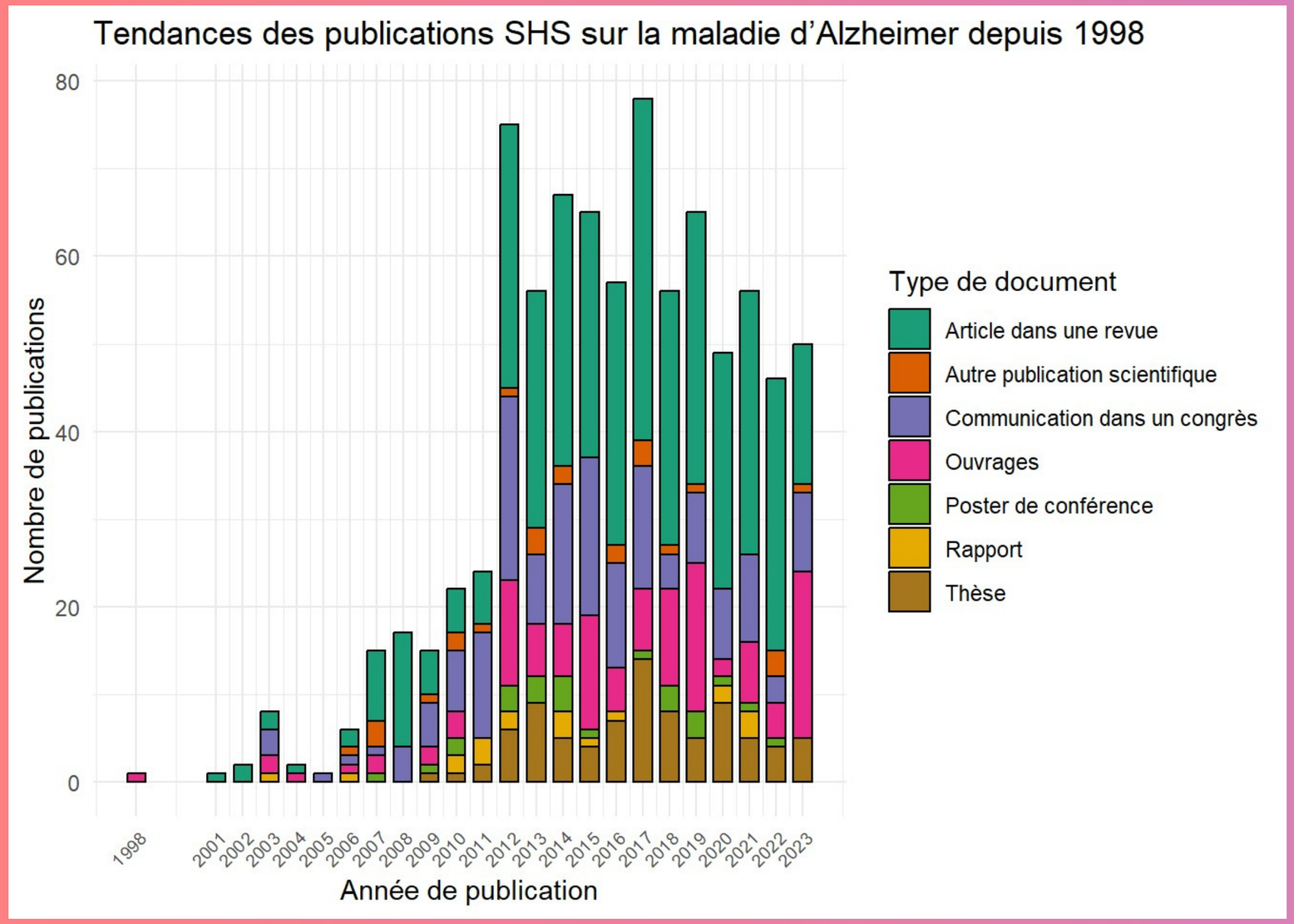
4. Auteurs actifs

5. Répartition des langues des publications des revues

6. Nuage de mots des titres des revues SHS

7. Les institutions les plus actives dans la publication des articles des revues SHS

1. Les tendances annuelles des publications



Pour avoir une idée globale de la recherche et de la popularité de HAL, il est intéressant d'explorer l'évolution temporelle des publications sur le sujet de la maladie d'Alzheimer dans le domaine qui nous intéresse : les Sciences Humaines et Sociales.

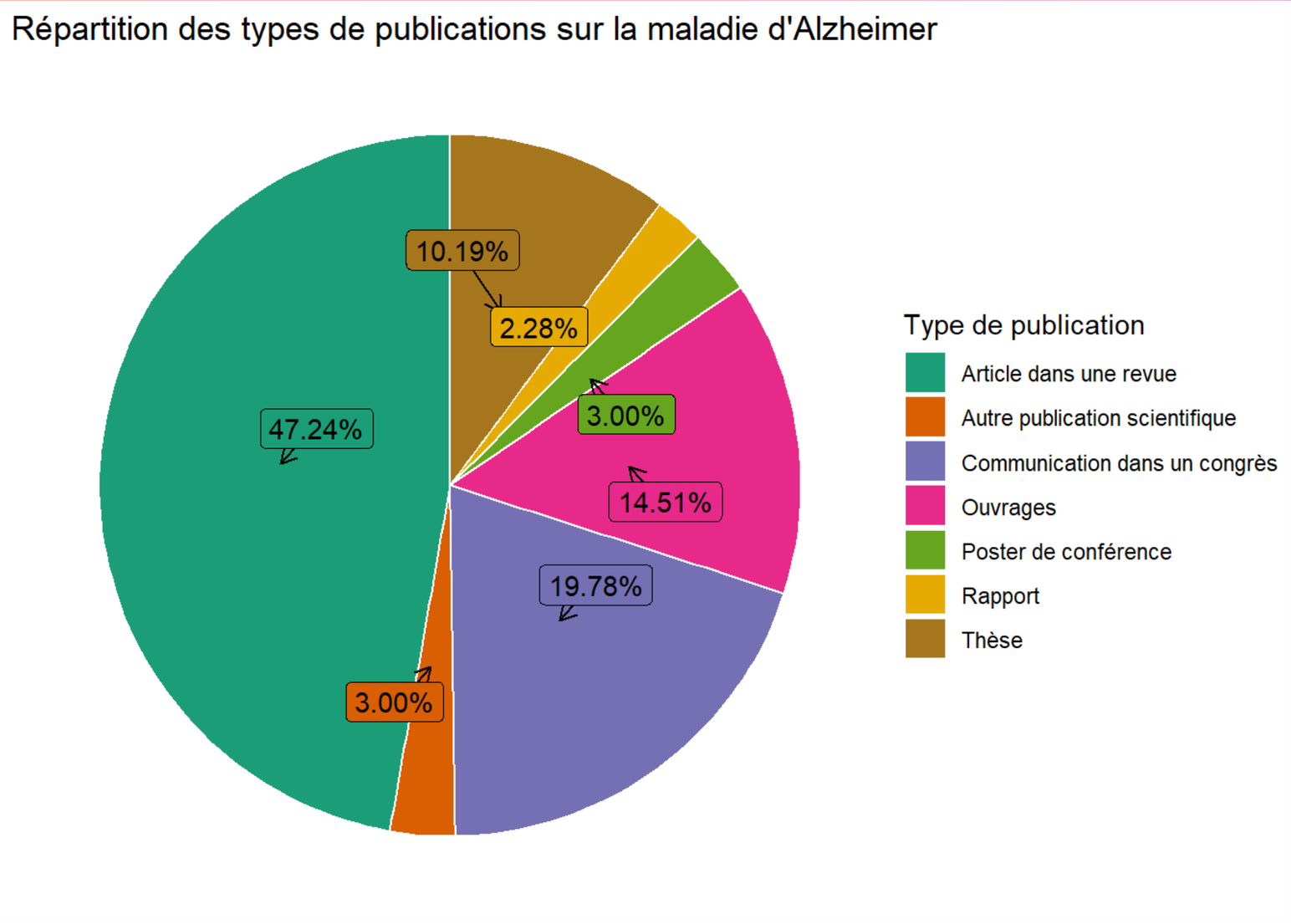
1. Les tendances annuelles des publications

La première publication accessible sur HAL portant sur notre sujet est un ouvrage qui a été publié en 1998. Cette date, antérieure à celle de la création de HAL, montre que la référence de publication indiquée peut aller au-delà des limites de la plateforme. HAL peut donc être un espace d'archivage de recherches scientifiques pour les ressources de plus de 25 ans.

Entre 2001 et 2006, il y a peu de publications. Mais il y a systématiquement un ou plusieurs articles publiés (à l'exception de 2004).
Il y a un pic de publications en 2012. Cette date marque la fin du Plan Alzheimer 2008-2012, projet politique qui tente d'améliorer la recherche française sur la maladie d'Alzheimer. On peut penser que de nombreuses recherches se sont intéressées, à cette période, à la maladie. Les résultats ont ensuite été publiés sur HAL.

2017 est une année où HAL enregistre le plus grand nombre de publications sur son site dans cette discipline.
Les articles de revues sont globalement le type de document qui y est le plus déposé.

2. La répartition des publications SHS



Pour visualiser toutes ces publications sous un autre angle, il est possible de regrouper ces données pour avoir une image résumant la répartition totale de ces publications en SHS sans l’aspect temporel.

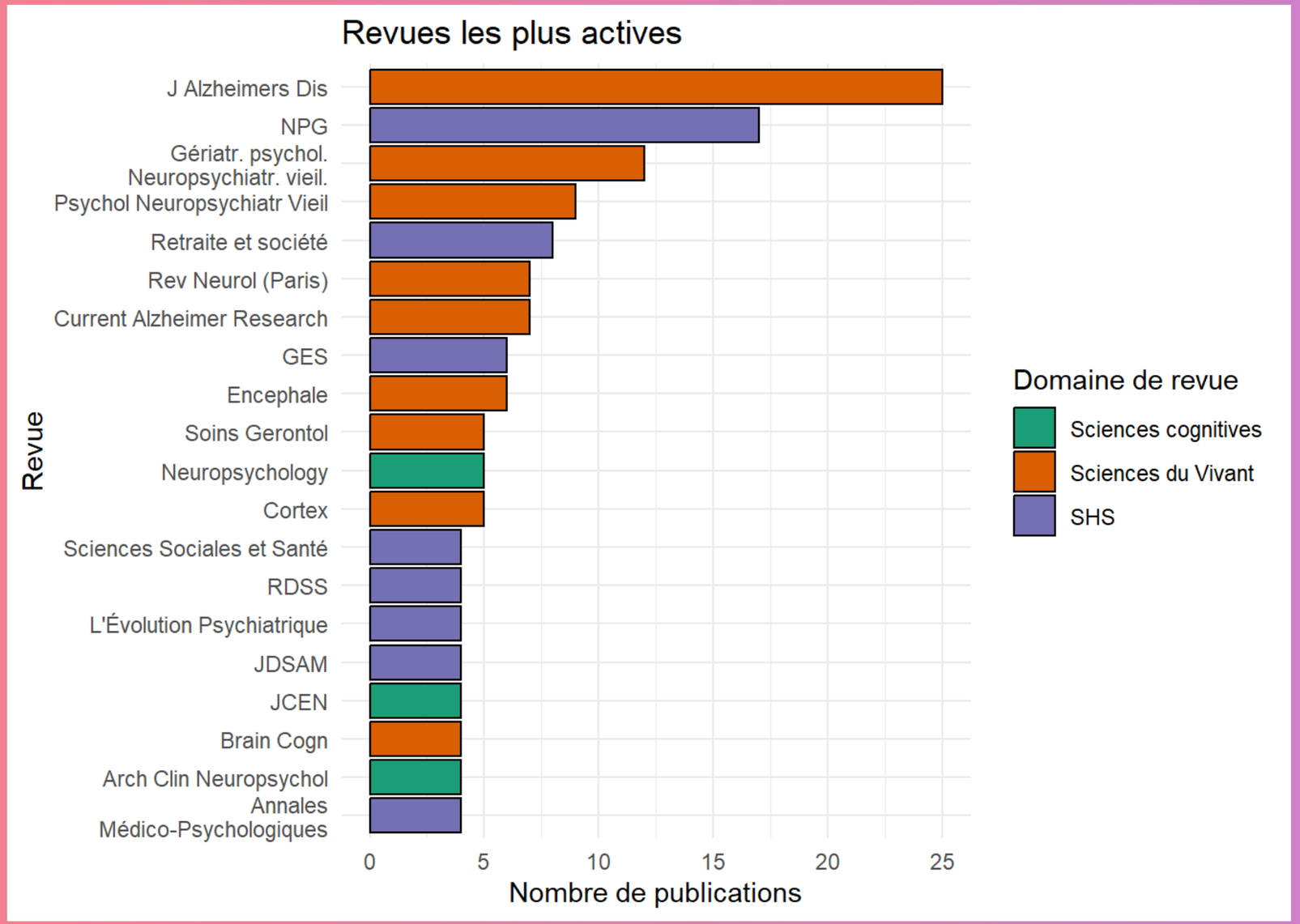
2. La répartition des publications SHS

Pour cela, l’analyse est représentée sous forme de camembert et gagne en lisibilité grâce aux couleurs et aux étiquettes de pourcentage.

Le grand nombre d’articles de revues est plus parlant dans ce graphique en comparaison au précédent. Ici, nous pouvons affirmer que presque 50% des publications sur l’Alzheimer, en SHS, disponibles sur HAL, sont des articles de revues.
Cependant, ce pourcentage est le résultat des dépôts sur HAL. Il n’est pas représentatif de la recherche globale dans ce domaine puisqu’il ne s’agit que d’une seule base de données pour la recherche scientifique. Il serait intéressant de compléter ces recherches et de les comparer avec des données sur d’autres sites comme Cairn.

En deuxième position, nous voyons que les communications dans des congrès sont davantage publiés que les ouvrages. On peut supposer que le dépôt sur HAL, une plateforme ouverte, n’incite pas les auteurs à soumettre leurs livres qu’il est possible d’acheter.

3. Top des revues les plus actives



3. Top des revues les plus actives

Notre recherche s’intéresse aux publications dans les revues, car elles représentent la moitié des documents disponibles sur le sujet.

Après avoir trié, dans l’ordre croissant, les 20 revues les plus actives sur le sujet de l’Alzheimer, il a été question d’associer un domaine principal à chacune de ces revues. Ces domaines ont été trouvés en cherchant via le moteur de recherche HAL, dans quels disciplines la revue publiait le plus.

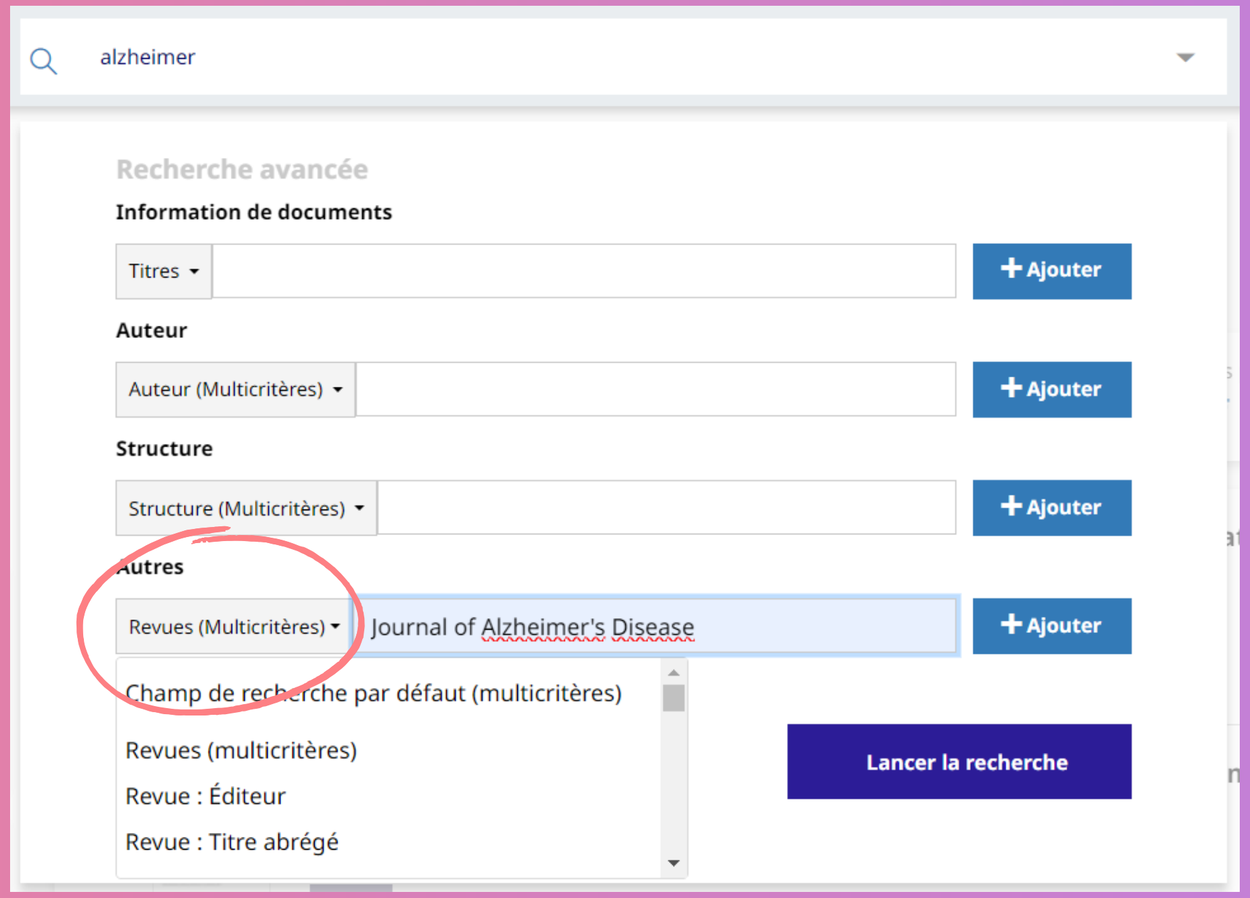
3. Top des revues les plus actives

Voici la méthode :

Après avoir tapé le mot “alzheimer” dans la barre de recherche principal, il est possible d’effectuer une recherche avancée.

Dans le paramètre “Autres”, on sélectionne “Revues [Multicritères]. Puis, on écrit le nom de la revue qui nous intéresse et on répète cette opération pour chaque revue.

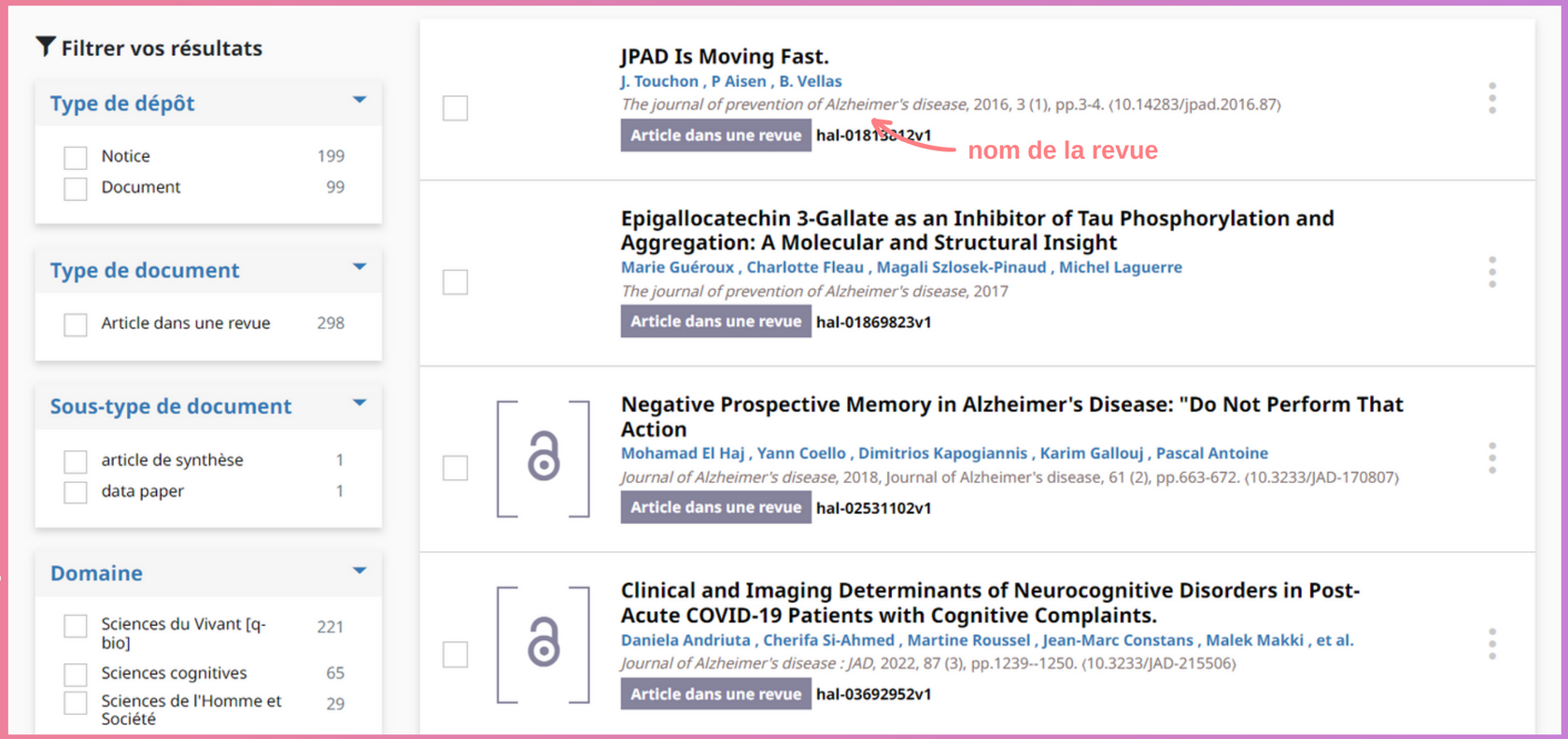
On lance la recherche.



3. Top des revues les plus actives

Voici les résultats lorsqu'on cherche "Journal of Alzheimer's Disease", la revue la plus active. Dans la colonne à gauche, on regarde la partie "Domaine". On voit que près de 4 fois plus de publications sont présentes sur HAL dans cette revue pour le domaine des Sciences du Vivant en comparaison aux Sciences cognitives. Ainsi, cette revue a été catégorisée dans les Sciences du Vivant.

les nombre de publications dans cette revue d'après leur domaine



3. Top des revues les plus actives

Cette méthode nous a permis de segmenter les domaines mais le manque d'information fournie au préalable sur la spécialité de ces revues pose un problème.
En effet, cette catégorisation par nos soins ne représente pas la réalité de ces revues.
Par exemple, le *Journal de droit de la santé et de l'assurance maladie* est très présent sur HAL dans le domaine SHS alors qu'il s'agit d'une revue qui porte davantage sur le droit et la médecine. Pour pallier ces erreurs, les auteurs doivent publier davantage sur HAL pour essayer d'atteindre une représentativité plus juste.

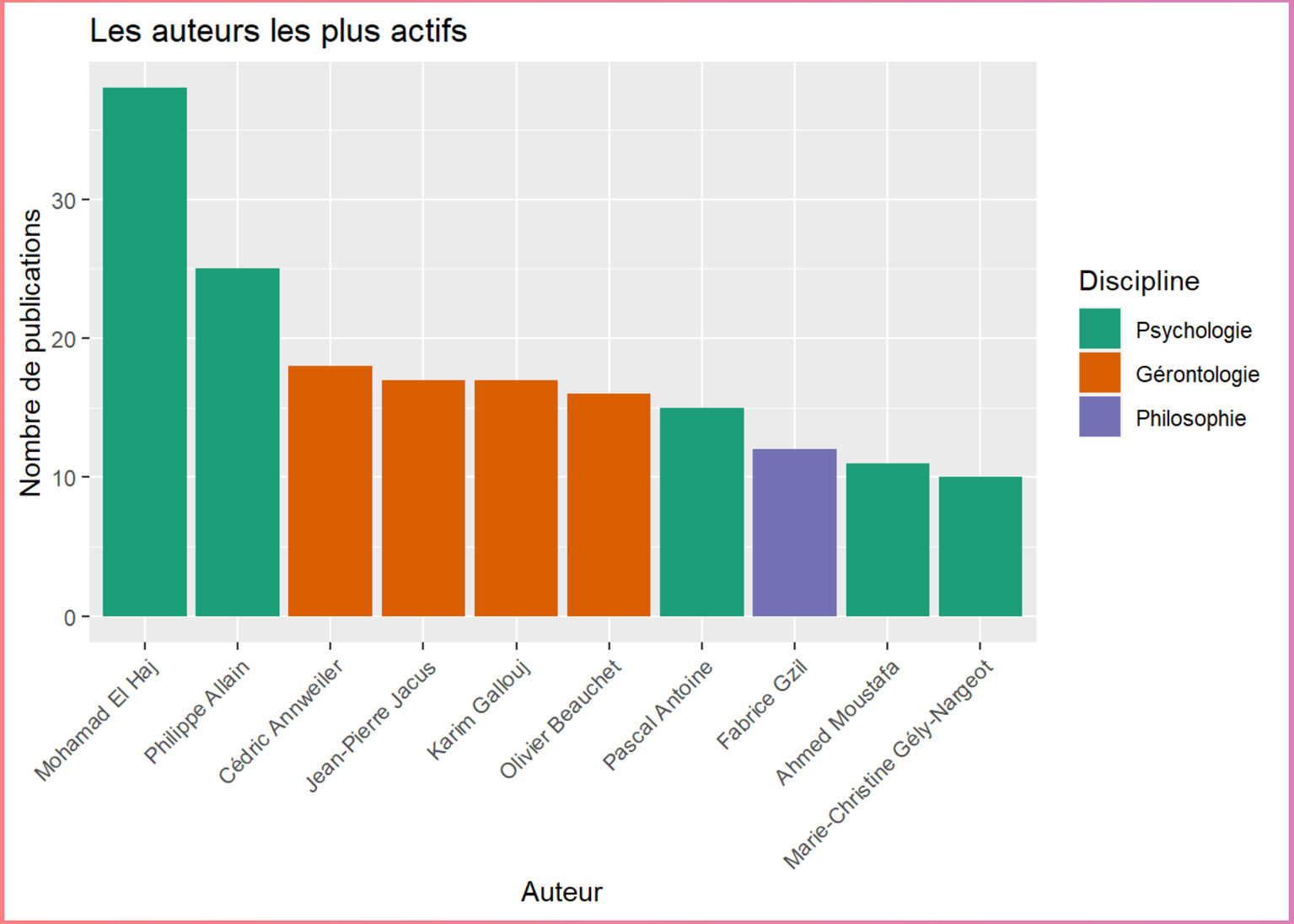
Puis, pour une meilleure lisibilité et compréhension de notre graphique, nous avons renommé les noms des revues par des noms plus courts.

3. Top des revues les plus actives

D’après les 20 revues les plus actives, les trois domaines majeurs qui traitent du sujet de l’Alzheimer sur HAL sont les sciences cognitives, les sciences du vivant et les sciences humaines et sociales.

En s’intéressant aux noms de ces revues, on remarque une récurrence de ce qui relève du domaine de la psychologie.

4. Les auteurs les plus actifs



Il était intéressant de voir si des auteurs contribuaient beaucoup en publiant sur HAL.

Nous avons établi le classement des 10 des auteurs les plus actifs en leur ajoutant une discipline majeure d'après leurs informations présentes sur d'autres sites Internet.

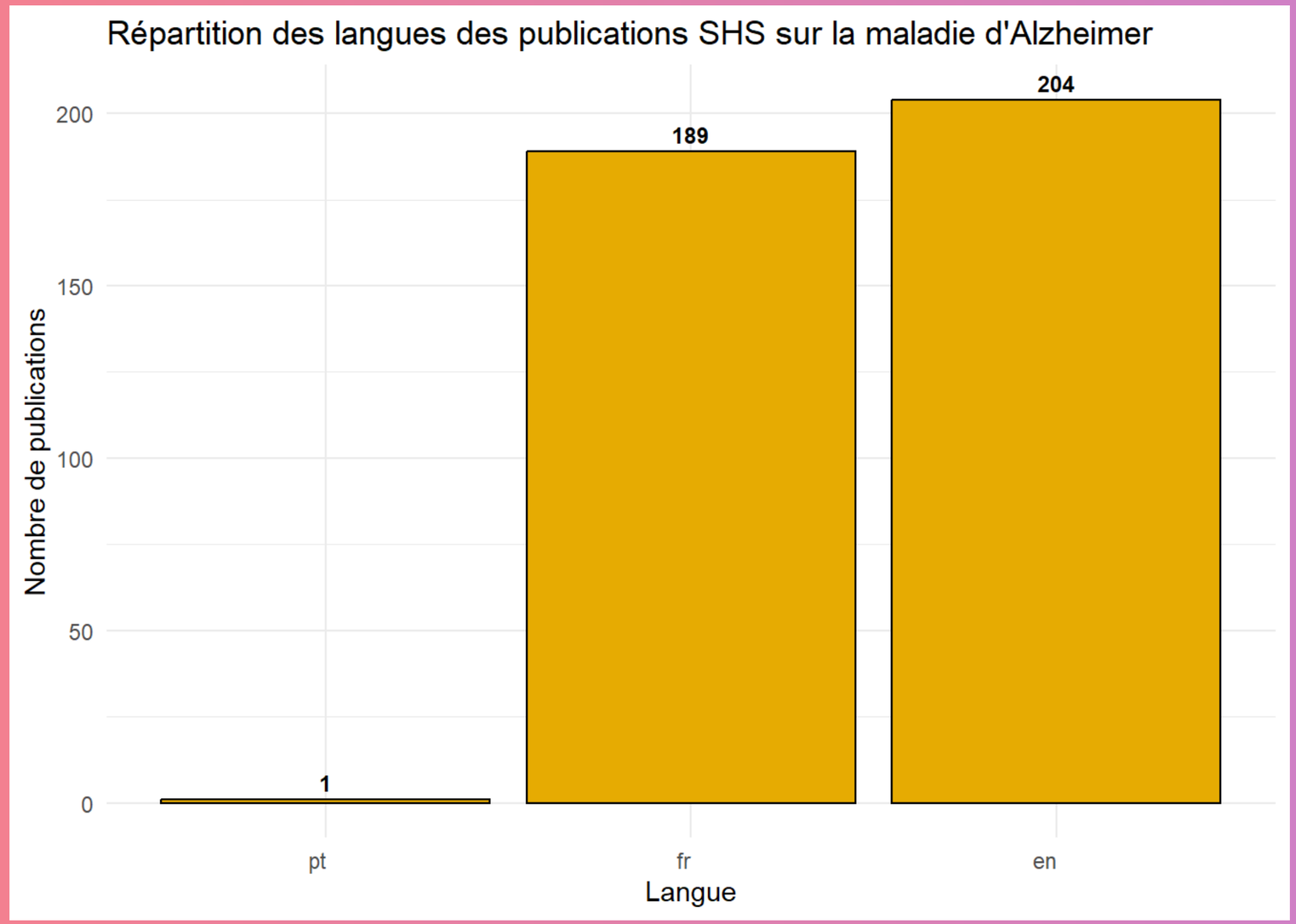
4. Les auteurs les plus actifs

HAL ne donne pas directement les informations relatives aux champs disciplinaires des auteurs. Cette recherche doit se faire manuellement et peut être assez contraignante si on cherche à représenter plus de 20 auteurs. De plus, un auteur peut avoir plusieurs disciplines majeures, ce qui est difficilement représentable sur le graphique.

La psychologie représente la discipline principale de la moitié de ces auteurs. Cela semble assez vraisemblable puisque nous avons remarqué précédemment que les revues en rapport avec la psychologie sont très présentes dans nos recherches. Cette discipline se place avant la gérontologie (l'étude du vieillissement), ce qui semble assez logique étant donné la maladie dont il est question.

Il y a tout de même un aspect philosophique à travers un auteur ce qui montre la diversité, voire la complémentarité, des angles d'approche pour étudier la maladie d'Alzheimer.

5. La répartition des langues des publications



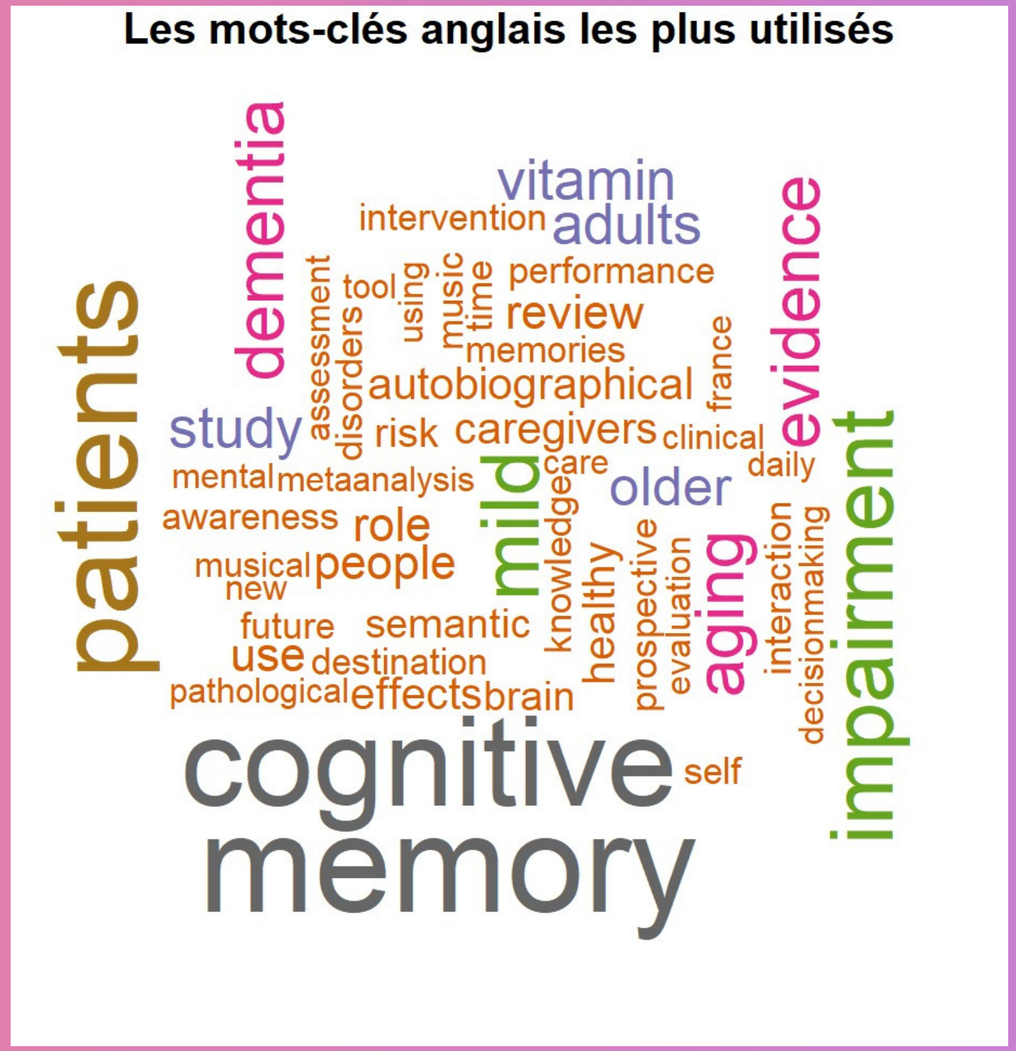
5. La répartition des langues des publications

À partir des textes des publications au sein des revues portant sur la thématique de l’Alzheimer, on peut distinguer les différentes langues.
Alors que le nettoyage réalisé plus haut sur les langues nous montrait leur petite diversité (plus de 5 sur l’ensemble des données), celle-ci se réduit drastiquement lorsqu’il s’agit de publications dans une revue. Les articles sont en anglais principalement, puis en français. Il n’y en a qu’un en portugais.

On peut penser que HAL est davantage populaire dans la recherche francophone et anglophone. Tout en sachant que HAL ne permet pas de faire ces segmentations facilement et de manière pratique. En effet, il y a beaucoup d’erreurs de catégorisation. Des ressources en français se retrouvaient dans la partie anglaise et inversement.

Il a fallut corriger ces erreurs manuellement.

6. Les mots récurrents dans les titres de revus SHS



6. Les mots récurrents dans les titres de revus SHS

En approfondissant cet aspect linguistique, nous avons généré 2 nuages de mots à partir des termes utilisés dans les titres des publications.

Il a été décidé de séparer les titres anglais de ceux français pour pouvoir comparer les 2. En excluant les mots “Alzheimer”, “maladie” et “disease”, ces deux graphiques nous montrent les termes les plus fréquents.

Pour obtenir ce résultat lisible et interprétable, un nettoyage était nécessaire pour enlever les nombres, la ponctuation, les stopwords (les “mots vides” comme les articles, conjonctions, etc...). On remarque qu’il reste toutefois certains mots qui ne sont pas toujours pertinents : “quand”, “chez”.

Ce nettoyage des mots est assez fastidieux et demande une intervention en partie manuelle.

6. Les mots récurrents dans les titres de revus SHS

Dans le nuage de mot français, on remarque que le terme “mild” est toujours présent. En effet, les anglicismes dans les titres français n’étaient pas détectés et apparaissent donc dans notre nuage de mots français. Il est pertinent de les garder puisqu’ils sont utilisés dans des titres français. D’autres mots anglais peuvent apparaître malgré le filtrage de la langue, dû au fait que certains auteurs mettent leur titre en français et en anglais.

Dans ce cas, il est nécessaire d’avoir recourt à un nettoyage manuel, parfois au cas par cas.

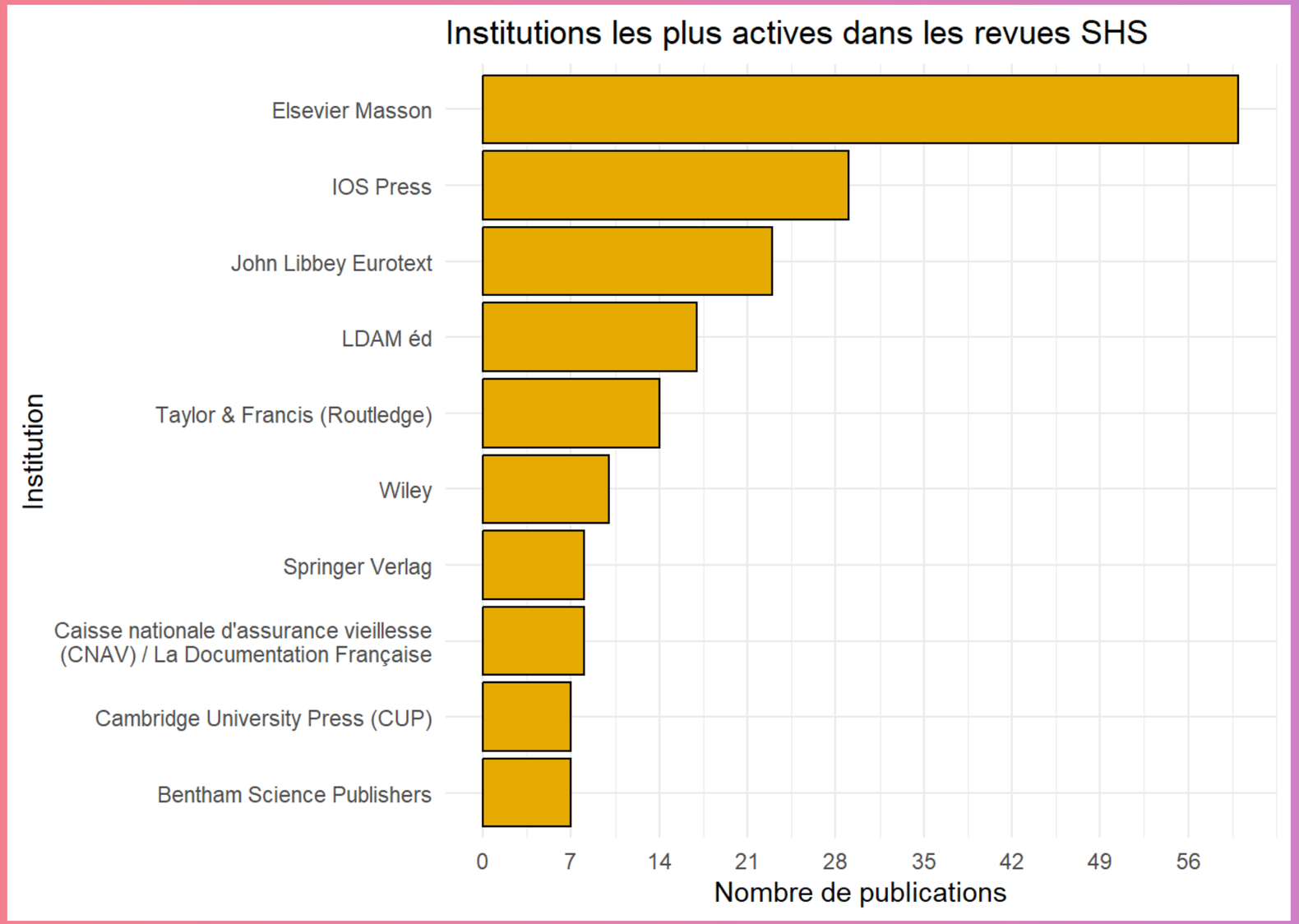
6. Les mots récurrents dans les titres de revus SHS

En comparant ces deux graphiques, on peut voir que du côté francophone, le terme “personnes” est au même niveau que “cognitive” et “memory”. Ceci peut montrer que la recherche francophone s’intéresse avant tout à l’individu ou aux personnes aidant les malades. Tandis que la recherche anglaise explore les phénomènes relatifs au cerveau.

Les termes équivalents en français : “cognitive” et “mémoire” n’apparaissent qu’en 4ème position sur le nuage de mots. Avant cela, les mots “troubles” et “aidants” sont davantage récurrents. Puis, en rose, des mots comme “diagnostic”, “prise”, “clinique”, “éthique” laissent penser que la recherche francophone se concentre davantage sur l’accompagnement et la gestion des personnes atteintes de la maladie d’Alzheimer.

Du côté anglophone, les termes qui ressortent sont relatifs à l’aspect concret de la maladie, ce qui se passe de manière technique et scientifique chez les personnes concernées par cette maladie.

7. Les institutions les plus actives



7. Les institutions les plus actives

La recherche s’est finalement portée sur la participation plus ou moins active de différentes institutions.

Afin d’établir un graphique montrant les 10 institutions les plus actives sur HAL sur le sujet de l’Alzheimer, il faut compter le nombre d’occurrences.

L’institution la plus active sur le sujet montre deux fois plus de publications que celle en deuxième position. Il s’agit d’une grande maison d’édition installée depuis plusieurs siècles. On peut imaginer que les financements sont plus conséquents que parmi d’autres institutions ce qui lui permet d’être très présente dans le domaine.

Conclusion

Ce travail à partir de l’API de HAL nous a permis de voir la richesse d’informations qu’il est possible de mobiliser. Toutes ces données sont précieuses pour mener à bien des projets de recherches et permettent d’approfondir de nombreux sujets.

Cependant, dans toutes les étapes de notre projet, nous avons été confrontés aux difficultés de lisbilité des données. En effet, des nettoyages sur plusieurs variables et à différents moments ont été plus que nécessaires pour bien comprendre les informations et pouvoir les rendre tout aussi compréhensibles dans la restitution de nos analyses. Une grande partie a été nettoyée à la main, ce qui est chronophage et fastidieux.

Selon les projets, il est plus ou moins possible d’utiliser l’API sans avoir recours à un nettoyage profond, mais dans d’autres cas, il est impératif de passer par là pour obtenir des données propres et exploitables. Cela limite l’accessibilité des données pourtant très nombreuses dans cette immense base de données.

Conclusion

Un autre point important à soulever est la difficulté de la représentation.

HAL, étant une archive ouverte, n’est alimentée que si un individu souhaite y déposer un document. Ainsi, la plateforme ne reflète pas l’étendue de la recherche actuelle et passée dans de nombreux domaines et sujets spécifiques. Il serait nécessaire, pour faire un état des lieux plus complet, de croiser les données avec d’autres plateformes, ou d’inciter les auteurs à publier systématiquement sur HAL.

Le croisement de données soulève un autre problème : la correspondance des valeurs et des variables. Les bases de données extraites via différentes sources peuvent montrer de nombreuses différences de catégorisations. Cela suppose, à nouveau, un très grand temps de nettoyage pour éviter les doublons par exmple.