CRÉATION D'UNE BASE DE DONNÉES BIBLIOGRAPHIQUE

René Riedel, Convis s.c.

Dans le cadre du projet Resi'Cow, il est prévu de créer une base de données documentaire dynamique sur le thème « Agriculture et protection du climat ». Cette base de données doit permettre, à tous les utilisateurs intéressés, de trouver et de consulter rapidement et facilement des documents sur des sujets d'actualité liés à la protection du climat dans le domaine agricole, et ainsi de se tenir informés des dernières avancées sur ces thèmes. La base de données doit être conçue de manière visuellement attrayante.

Dans un premier temps, une recherche bibliographique a été menée. À partir de publications, d'articles, de livres, d'études, de sites web, de descriptions techniques et de présentations, 86 documents totalisant plus de 3 000 pages ont été rassemblés et lus. Les thèmes abordés ont été extraits et regroupés en dix catégories, présentées dans le Tableau 1, qui couvrent l'ensemble des domaines d'actualité liés à la protection du climat dans le secteur agricole. Les documents pertinents peuvent donc être classés dans une ou plusieurs de ces catégories.

Le classement des documents dans ces catégories doit s'effectuer par analyse de texte, en y cherchant des mots-clés spécifiques à une catégorie. À titre d'exemple, le mot « politique » constitue un mot-clé pour la catégorie « Politique et société ». Par conséquent, un document contenant le mot « politique » peut ainsi être classé dans la catégorie « Politique et société ».

Afin de dresser, pour chaque catégorie, une liste de mots-clés décrivant spécifiquement cette dernière, tous les mots contenus dans les 86 documents ont été extraits et classés en fonction de leur fréquence. Les mots de remplissage tels que « le », « et », etc. ont été supprimés de la liste, de sorte qu'il ne restait plus que les mots pertinents. Parmi ces mots, 20 à 30 mots-clés spécifiques ont été sélectionnés pour chaque catégorie.

Le classement des documents à l'aide de ces mots-clés s'est toutefois avéré très complexe. D'une part, les documents sont rédigés dans différentes langues (français, allemand et anglais) ; d'autre part, il existe de nombreuses variantes d'un même mot-clé en raison de la conjugaison, de la déclinaison et des synonymes.

Par conséquent, toutes les variantes possibles de chaque mot-clé ont été identifiées et intégrées à l'analyse de texte.

Afin d'automatiser l'analyse du texte, un script a été écrit en Python ; celui-ci extrait le texte de chaque document et le parcourt à la recherche des mots-clés ainsi que de toutes leurs variantes. Le classement par catégories est soumis aux règles suivantes :

1. Le classement dans une catégorie n'a lieu que si au moins trois mots-clés de cette catégorie ont été trouvés dans le document.
2. Un document peut être classé dans plusieurs catégories.

La première règle empêche que des documents soient classés à tort dans une catégorie qu'ils ne font que mentionner sans l'aborder plus en détail. La deuxième règle découle du fait qu'un document peut traiter de plusieurs thèmes. De cette manière, il est également possible de classer de nouveaux documents en quelques secondes.

Un diagramme de Sankey, tel qu'illustré à la Figure 2, sert à présenter les résultats de l'analyse de texte. Les différents nœuds y sont classés en colonnes et reliés entre eux. Cela offre un aperçu rapide des thèmes et de leur fréquence, ainsi que de leurs relations mutuelles, et permet à l'utilisateur d'accéder à une page web complémentaire en cliquant sur un thème, par exemple une liste des références des documents. Un script Python permet la création automatisée de ce diagramme de Sankey à l'issue de l'analyse de texte.

L'analyse textuelle préliminaire des 86 documents a montré que ce sont surtout les thèmes de la « gestion des terres » et de la « typologie des exploitations » qui sont abordés, tandis que ceux de la « rentabilité » et de l'« autonomie » étaient moins fréquents.

Les travaux actuels se concentrent sur l'optimisation de l'analyse textuelle, notamment par l'ajustement des mots-clés, ainsi que sur l'amélioration de la présentation des résultats grâce à une conception plus interactive du diagramme de Sankey. Les résultats correspondants, présentés sous forme de diagramme de Sankey devraient bientôt être disponibles sur le site web du projet Resi'Cow.

Tableau 1 - Aperçu des catégories abordées dans la littérature

	Catégories	Thèmes
1	**Impact de l'alimentation animale**	Stratégies d'alimentation, compléments alimentaires, influence sur les processus intestinaux, réduction des émissions de méthane
2	**État des connaissances scientifiques**	Théorie des GES, méthodes de mesure des GES, expériences, cycles des matières
3	**S'adapter à l'avenir**	Impacts du changement climatique sur l'agriculture, adaptation au comportement des consommateurs
4	**Réduction des émissions de GES à l'échelle de l'exploitation**	Sources et puits de GES à l'échelle de l'exploitation, empreinte écologique, installations de biogaz
5	**Impact de la typologie des exploitations**	Agriculture biologique ou conventionnelle, influence de la typologie, adaptation de l'élevage, production laitière vs production viandeuse
6	**Renforcer l'autonomie**	Améliorer l'autosuffisance en protéines, culture locale, optimisation des cultures et des rotations culturales, optimisation de l'absorption d'azote et de l'économie circulaire
7	**Politique et société**	Stratégies climatiques, labels et systèmes de certification, formation et conseil aux agriculteurs, perception du public, communication
8	**Rentabilité**	Rentabilité des exploitations, impact des certificats, potentiel d'économies, subventions
9	**Gestion des sols**	Stockage du carbone, émissions de protoxyde d'azote dues à la fertilisation, humus, utilisation des terres
10	**Optimisation grâce à la technologie**	Logiciels utiles, intelligence artificielle, aides techniques

Figure 1 - Nuage de mots-clés, classés par catégories. La taille des mots-clés reflète le nombre relatif de documents qui les contiennent.

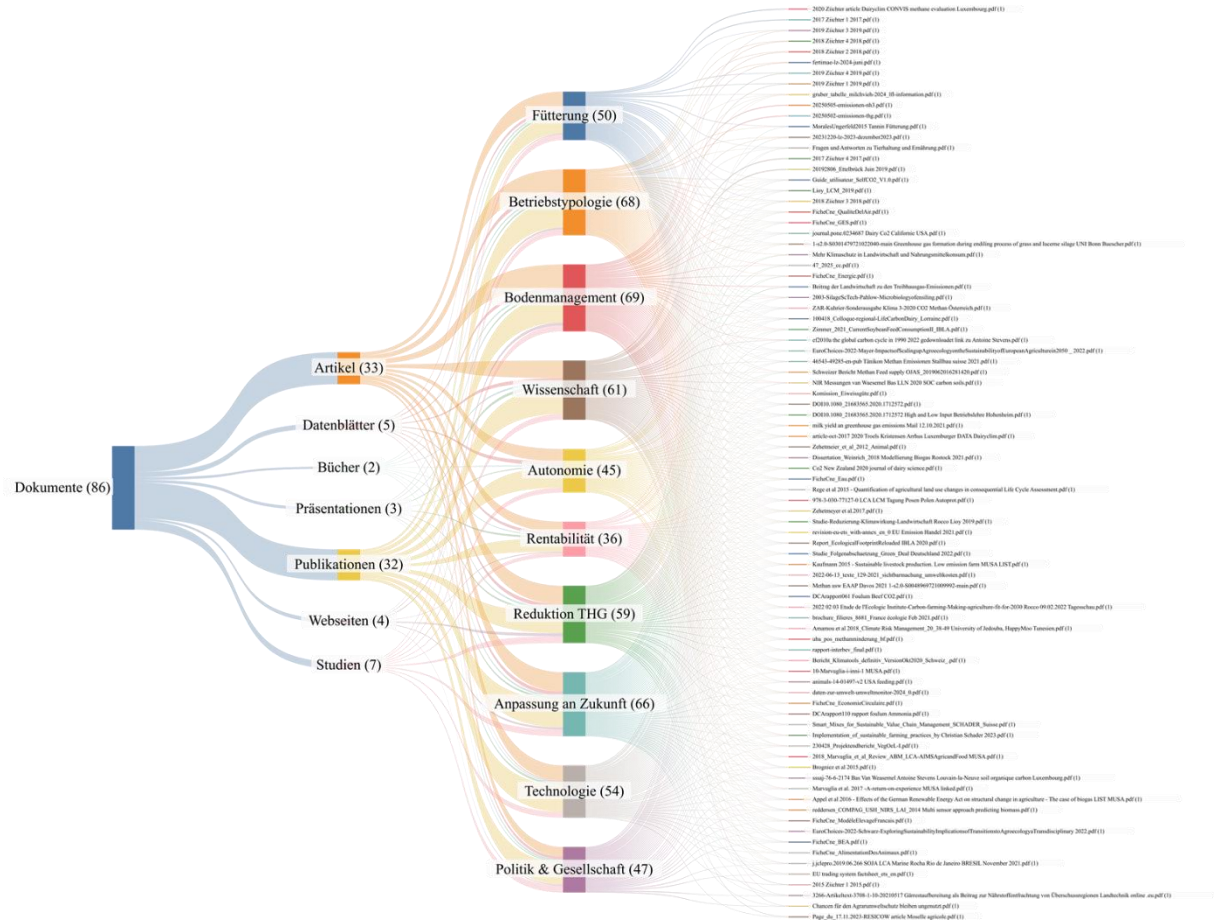

Figure 2 - Résultats de l'analyse de texte sous forme de diagramme de Sankey. Le type de documents ainsi que les catégories sont représentés sous forme de nœuds, et leurs relations entre eux sont visualisées par des liens. La hauteur d'un nœud est proportionnelle au nombre de documents qu'il contient. La colonne de droite présente les documents analysés.