ERSTELLEN EINER BIBLIOGRAPHISCHEN DATENBANK

René Riedel, Convis s.c.

Im Rahmen des Projekts Resi'Cow soll eine dynamische Dokumentendatenbank zum Thema „Landwirtschatz und Klimaschutz" erstellt werden. Diese Datenbank soll allen interessierten Nutzern die Möglichkeit geben, Dokumente zu aktuell diskutierten Themengebieten rund um den landwirtschaftsbezogenen Klimaschutz schnell und einfach zu finden und abrufen zu können und sich so zu einem Thema auf den neuesten Stand bringen zu können. Die Datenbank soll dabei visuell ansprechend gestaltet werden.

Zunächst wurde eine Literaturrecherche betrieben. Aus Publikationen, Artikeln, Büchern, Studien, Webseiten, technischen Beschreibungen und Präsentationen wurden 86 Dokumente mit insgesamt mehr als 3000 Seiten zusammengestellt und gelesen. Dabei wurden die diskutierten Themen herausgeschrieben und in zehn Kategorien zusammengefasst, die in Tabelle 1 zu sehen sind und die alle aktuellen Themengebiete des landwirtschaftsbezogenen Klimaschutzes abdecken. Relevante Dokumente lassen sich daher einer oder mehrerer dieser Kategorien zuordnen.

Die Einteilung von Dokumenten in diese Kategorien soll per Textanalyse geschehen, indem die Texte aus den Dokumenten nach Schlüsselwörtern durchsucht werden, die spezifisch für eine Kategorie sind. Als Beispiel stellt das Wort „Politik" ein Schlüsselwort für die Kategorie „Politik und Gesellschaft" dar. Ein Dokument, das das Wort „Politik" enthält, kann somit der Kategorie „Politik und Gesellschaft" zugeteilt werden.

Um für jede Kategorie eine Liste an Schlüsselwörtern zu erstellen, die die jeweilige Kategorie spezifisch beschreiben, wurden alle Wörter in den 86 Dokumenten extrahiert und ihrer Häufigkeit entsprechend sortiert. Füllwörter wie „der", „und", etc. wurden aus der Liste entfernt, so dass nur relevante Wörter übrigblieben. Aus diesen Wörtern wurden für jede Kategorie 20 bis 30 spezifische Schlüsselwörter herausgesucht.

Die Zuteilung der Dokumente anhand dieser Schlüsselwörter gestaltete sich jedoch sehr komplex. Zum einen liegen die Dokumente in verschiedenen

Sprachen (Französisch, Deutsch und Englisch) vor, zum anderen existieren durch Konjugation, Deklination und Synonymbildung viele verschiedene Variationen eines Schlüsselworts. Folglich wurden für jedes Schlüsselwort alle möglichen Variationen herausgesucht und in die Textanalyse eingebaut.

Zur Automatisierung der Textanalyse wurde ein Skript in der Programmiersprache Python verfasst, welches den Text aus jedem Dokument extrahiert und nach den Schlüssel-wörtern samt all ihrer Variationen durchsucht. Die Zuteilung in Kategorien ist dabei an folgende Regeln geknüpft:

1. Die Zuteilung in eine Kategorie erfolgt erst, wenn mindestens drei Schlüsselwörter dieser Kategorie im Dokument gefunden wurden.
2. Ein Dokument kann mehreren Kategorien zugeordnet werden.

Die erste Regel verhindert, dass Dokumente fälschlicherweise einer Kategorie zugeordnet werden, die sie nur erwähnen, aber nicht weiter behandeln. Die zweite Regel ergibt sich aus der Tatsache, dass ein Dokument mehr als ein Thema behandeln kann. Auf diese Art und Weise lassen sich auch neue Dokumente innerhalb von Sekunden zuteilen.

Zur Darstellung der Ergebnisse der Textanalyse dient ein Sankey-Diagramm, wie es in Abbildung 2 zu sehen ist. Hier werden verschiedene Knoten in Spalten sortiert dargestellt und miteinander verbunden. Dies erlaubt einen schnellen Überblick über die Themen und ihre Häufigkeit sowie ihre Relation zueinander und erlaubt es dem Benutzer, durch Klick auf ein Thema auf eine weiterführende Webseite zu gelangen, beispielsweise eine Liste der Referenzen der Dokumente. Ein Pythonskript erlaubt die automatisierte Erstellung dieses Sankey-Diagramms nach der Textanalyse.

Die vorläufige Textanalyse der 86 Dokumente zeigte, dass vor allem die Themen „Bodenmanagement“ und „Betriebstypologie“ diskutiert werden, während „Rentabilität“ und „Autonomie“ weniger stark aufzufinden waren.

Momentane Arbeiten fokussieren sich auf die Optimierung der Textanalyse, etwa durch Anpassen der Schlüsselwörter, sowie der Anpassung der Darstellung der Ergebnisse durch eine interaktivere Gestaltung des Sankey-Diagramms. Die entsprechenden Resultate mit dem Sankey-Diagramm sollen bald auf der Webseite des Projekts Resi’Cow zu finden sein.

Tabelle 1 - Übersicht über die in der Literatur diskutierten Kategorien

	Kategorie	Themen
1	**Einfluss der Fütterung**	Fütterungsstrategien, Futterergänzungsmittel, Einfluss auf Vorgänge im Darm, Methanreduktion
2	**Stand der Wissenschaft**	Theorie der THG, THG-Messmethoden, Experimente, Stoffkreisläufe
3	**Anpassung an die Zukunft**	Auswirkungen des Klimawandels auf die Landwirtschaft, Anpassung auf Verbraucherverhalten, Klimakiller Kuh?
4	**THG-Einsparungen auf dem Betrieb**	THG-Quellen/Senken auf dem Betrieb, Ökologischer Fußabdruck, Biogasanlagen
5	**Einfluss der Betriebstypologie**	Ökologisch oder konventionell, Einfluss der Typologie, Anpassung der Tierhaltung, Milch- vs. Fleischproduktion
6	**Autonomie stärken**	Proteinautarkie verbessern, Lokaler Anbau, Optimierter Anbau von Pflanzen & Fruchtfolgen, Optimierung der Stickstoffaufnahme und der Kreislaufwirtschaft
7	**Politik & Gesellschaft**	Klimastrategien, Label & Zertifizierungssysteme, Bildung & Beratung für Landwirte, Öffentliche Wahrnehmung, Kommunikation
8	**Rentabilität**	Betriebliche Wirtschaftlichkeit, Wirkung von Zertifikaten, Einsparpotenzial, Subventionen
9	**Bodenmanagement**	Kohlenstoffspeicherung, Lachgasemission durch Düngung, Humus, Landnutzung
10	**Optimierung durch Technologie**	Hilfreiche Software, Künstliche Intelligenz, Technische Hilfsmittel

Abbildung 1 - Wortwolke der Schlüsselwörter, nach Kategorien unterteilt. Die Größe der Schlüsselwörter spiegelt die relative Anzahl der Dokumente wider, die sie enthalten

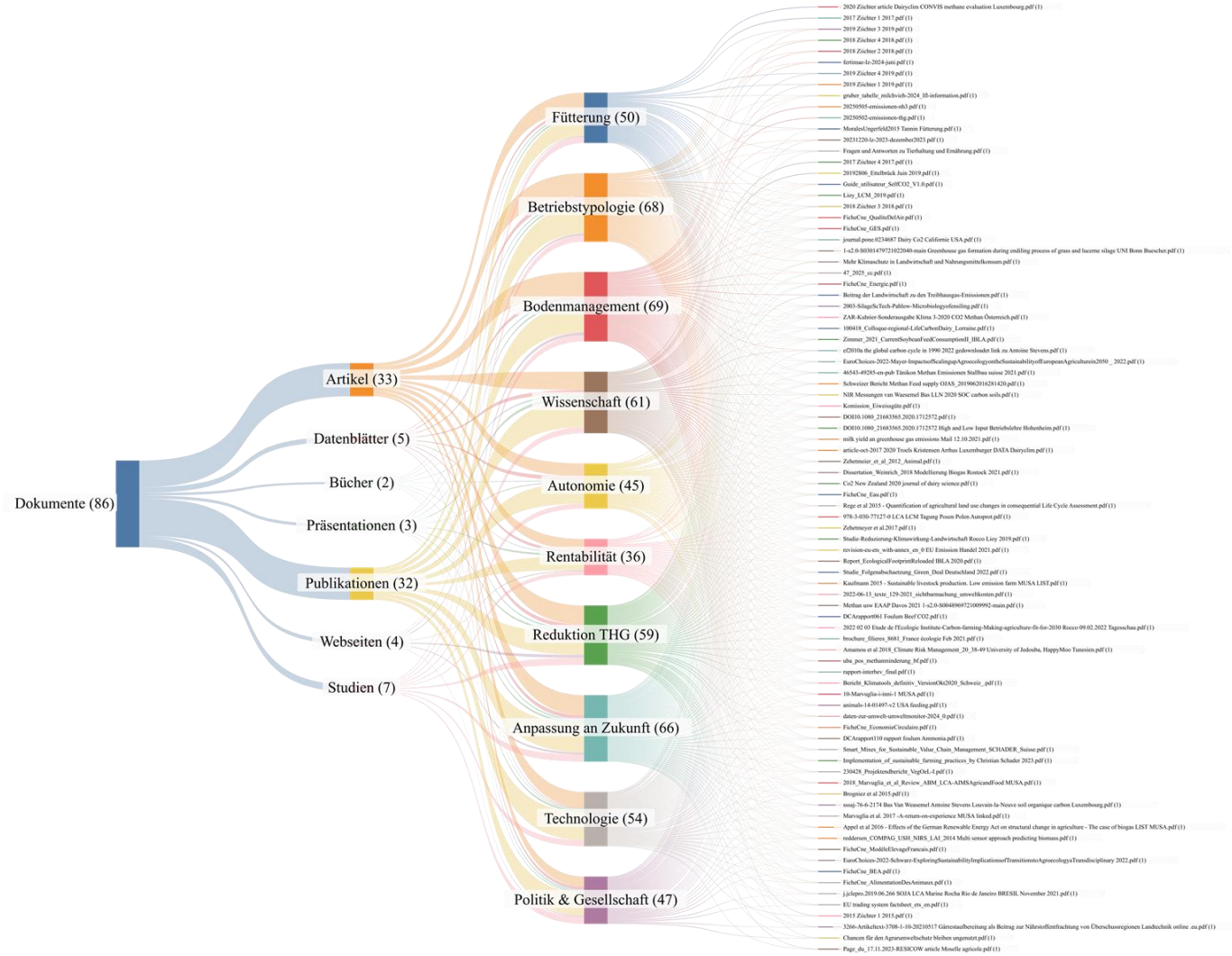

Abbildung 2 - Ergebnisse der Textanalyse als Sankey-Diagramm. Die Art der Dokumente sowie die Kategorien werden als Knoten dargestellt, ihre Relationen zueinander werden durch Verknüpfungen visualisiert. Die Höhe eines Knotens ist proportional zur Anzahl Dokumente in diesem Knoten. In der rechten Spalte befinden sich die untersuchten Dokumente