DEUTSCHES HANDWERKSINSTITUT

Lukas Meub, Till Proeger, Kilian Bizer, Jörg Lahner

Vernetzung von Unternehmen und Forschungseinrichtungen in regionalen Innovationssystemen durch Webscraping

Göttinger Beiträge zur Handwerksforschung 62

Volkswirtschaftliches Institut für Mittelstand und Handwerk an der Universität Göttingen

Veröffentlichung
des Volkswirtschaftlichen Instituts für Mittelstand und Handwerk
an der Universität Göttingen

Forschungsinstitut im Deutschen Handwerksinstitut e.V.

Gefördert durch:

Bundesministerium
für Wirtschaft
und Klimaschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

Die Wirtschaftsministerien
der Bundesländer

DHKT
DEUTSCHER
HANDWERKSKAMMERTAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über

http://dnb.dnb.de

abrufbar.

ISSN 2364-3897
DOI-URL: http://dx.doi.org/10.3249/2364-3897-gbh-62

ifh Göttingen ● Heinrich-Düker-Weg 6 ● 37073 Göttingen
Tel. +49 551 39 174882
E-Mail: info@ifh.wiwi.uni-goettingen.de
Internet: www.ifh.wiwi.uni-goettingen.de

GÖTTINGEN • 2022

Vernetzung von Unternehmen und Forschungseinrichtungen in regionalen Innovationssystemen durch Webscraping

Autoren: Lukas Meub, Till Proeger, Kilian Bizer, Jörg Lahner
Göttinger Beiträge zur Handwerksforschung Nr. 62

Zusammenfassung

Diese Studie zeigt am Beispiel von Südniedersachsen, wie mit der Methodik des Webscrapings effizient Informationen zu regionalen Akteuren wie Forschungseinrichtungen und Unternehmen gewonnen werden können. Webscraping beschreibt das systematische Auslesen der Inhalte von Webseiten und deren anschließende statistischer Analyse z.B. im Hinblick auf spezifische Technologien oder Strukturmerkmale von Unternehmen.

Auf diesem Wege können gemeinsame technologische Schwerpunkte identifiziert werden, wodurch effizient Netzwerkaktivitäten und Projektverbünde zwischen den Akteuren aufgebaut werden können. Akteuren der Wirtschafts- und Innovationsförderung bietet dies ein innovatives Werkzeug, um die regionalen Innovationsnetzwerke strukturiert auszubauen und neue Kooperationen zu etablieren.

Die Studie skizziert verschiedene Anwendungsmöglichkeiten am Beispiel von Südniedersachsen. Zum einen werden regionale Spezialisierungen bei Unternehmen und Forschungseinrichtungen in den Bereichen Wasserstoff und Lasertechnologie aufgezeigt. Ferner erfolgt eine Zuordnung von Unternehmen und Forschungseinrichtungen zu verschiedenen Technologieschwerpunkten.

Übergreifend stellt Webscraping ein innovatives und effizient einsetzbares Instrument für regionale Innovationsakteure dar, das die regionale Koordination und den strukturierten, themenbezogenen Ausbau von Innovationsnetzwerken erleichtert.

Schlagwörter: *Regionale Innovationssysteme, Webscraping, Wirtschaftsförderung*

Inhalt

1. Einleitung 1

2. Methodik und Datengrundlage 2

3. Ergebnisse 4

4. Ausblick 9

5. Literatur 11

Abbildungen

Abb. 1: Häufigkeiten zur Nennung der Schlüsseltechnologien 4

Abb. 2: Matching Unternehmen und Technologien 5

Abb. 3: Matching Forschungseinrichtungen und Technologien 6

Abb. 4: Verknüpfung von Technologiefeldern 7

Abb. 5: „Wasserstoff“ als Technologie in der Region und überregionale Partner 8

1. Einleitung

Erfolgreiche Innovationsvernetzung hat im Kern ein Informationsproblem: zwei oder mehrere (regionale) Akteure könnten voneinander profitieren und Kooperationen initiieren, wenn sie von den Aktivitäten und vom Wissen des jeweils Anderen wüssten. Die Aufgabe von Innovationsintermediären ist es, ausreichend Wissen über verschiedenste Akteure zu generieren, um ein erfolgreiches Match-Making durchzuführen; das heißt, den jeweiligen Akteuren voneinander zu berichten, sie auf den potenziellen Nutzen einer Kooperation hinzuweisen und diese organisatorisch zu unterstützen. Die zentrale limitierende Größe bei der Innovationsvernetzung ist allerdings das Wissen um Forschungs- und Marktinteressen einzelner Akteure des regionalen Innovationssystems.

Eine kosteneffiziente Lösung dieses Problems leistet das sog. Webscraping von Internetseiten. Diese aktuelle Forschungsmethode beschreibt das systematische, regelmäßige Herunterladen und Auswerten von Texten und Links einer großen Anzahl von Internetseiten mit Hilfe passgenauer Software. Deren Inhalte können strukturiert und nach verschiedenen Kriterien analysiert und die Ergebnisse anschließend anschaulich dargestellt werden. Insbesondere innovative Unternehmen unterhalten in der Regel große und informationsreiche Internetauftritte, die gegenüber Kunden und Fachkräften aussagekräftig und attraktiv wirken sollen. Daher ist – unabhängig von der Betriebsgröße – ein hoher Informationsgehalt in Bezug auf neue Technologien und Spezialisierungen des Unternehmens abrufbar. Da insbesondere die innovativen kleinen Unternehmen eine wichtige und schwer zugängliche Zielgruppe für Innovationsvernetzung bilden, stellt Webscraping hier einen optimalen Zugang dar.

Grundsätzlich können – sofern die Internetadressen der relevanten Akteure des regionalen Innovationssystems erhoben wurden – alle öffentlich angezeigten technischen oder forschungsbezogenen Inhalte der Akteure dokumentiert und als regelmäßig aktualisierte Datenbank ausgewertet werden. Sobald die Datenbasis aller Homepages vorhanden ist, kann zunächst eine Inhaltsanalyse der Aktivitätsfelder aller Akteure im regionalen Innovationssystem erfolgen. Entsprechend werden regionale Spezialisierungen und Innovationspfade aufzeigt. Im nächsten Schritt können für die Wissensintermediäre eine Detailsuche und der Aufbau themenspezifischer Datenbanken erfolgen. Hierbei können etwa Technologien oder Forschungsinteressen als Kategorien genutzt werden. Im Anschluss an diese Strukturierung der Unternehmens- und Forschungslandschaft können Vernetzungsformate, Projekte o.ä. konzipiert werden, welche die so identifizierten potenziellen Kooperationspartner adressieren. Ebenso können gewünschte Formate mit den entsprechenden Kontakten hinterlegt werden oder Technologiescouts auf Basis der neuen Datenbank gezielt in die Vernetzung von Akteuren einsteigen.

Die vorliegende Studie illustriert am Beispiel südniedersächsischer Unternehmen und Forschungseinrichtungen die Methodik und ihre Potenziale im Hinblick auf bessere Innovationsvernetzung und skizziert als Ausblick weitere sinnvolle und mögliche Anwendungsfelder von Webscraping in der regionalen Wirtschaftsförderung.

2. Methodik und Datengrundlage

Die Grundlage dieser Studie bildet ein Datensatz mit Unternehmen und Forschungseinrichtungen in Südniedersachsen.

Die Liste der Unternehmen leitet sich aus Bewerbern zum Innovationspreis des Landkreises Göttingen ab.[1] Natürlich bilden die 226 Homepages der Unternehmen eine spezifische Auswahl besonders innovativer und kommunikativer Unternehmen. Eine Vollerhebung der Unternehmen in der Region würde ein vollständigeres Bild liefern, jedoch ist diese Liste zweckmäßig, um das Potenzial der Methodik Webscraping abzubilden.[2]

Mittels Internetrecherche wurden zusätzlich wichtige Forschungseinrichtungen der Region erhoben (Universität Göttingen, HAWK, TU Clausthal). Dabei wurde auf Grund der umfangreichen Webauftritte der Universitäten auf die Fakultätsebene bzw. Institutsebene abgestellt. Zudem sind die Max-Planck-Institute Südniedersachsens einbezogen. Je nach Grad der Differenzierung sind eventuelle Matches zwischen Unternehmen und Forschungseinrichtungen zielgerichteter (z.B. Identifikation eines spezifischen Forschenden). Der hier vorgestellte Ansatz bildet einen Mittelweg, bei welchem eine hinreichende Eingrenzung von Matches erfolgt, um in einem zweiten Schritt Ansprechpartner oder Kontakte zu identifizieren. So wurden insgesamt 94 Webseiten von Forschungseinrichtungen analysiert.[3]

Das Scraping der Webseiten selbst erfolgte im Juni 2021 mittels des ARGUS-Programms[4] und die Textanalyse mit dem Statistikprogramm R.[5] ARGUS sammelt zum einen die Volltexte der Webseiten und zum anderen die Verlinkungen auf andere Online-Präsenzen. Für beide Ansätze wird der Zeitpunkt des Zugriffs dokumentiert, ob eine Fehlermeldung im Zugriff auf die Seite entsteht oder eine Weiterleitung zu einer anderen Website vorgenommen wurde. Für den Textansatz werden neben den Volltexten, sofern vorhanden, Stichworte der Websites und Beschreibungen verzeichnet. Die Dokumentation der Verlinkungen unterscheidet in interne und externe Links, wobei sich „intern“ auf das analysierte Sample, also auf die Unternehmen und Forschungseinrichtungen der Stichprobe bezieht; „extern“ sind somit alle weiteren Webseiten.

Um ein möglichst umfangreiches Bild der Homepages zu erhalten, wurden jeweils bis zu 500 Unterseiten der Websites erfasst, Zudem wurde die Option der Bevorzugung kürzerer URLs aktiviert. Grundlegend ist anzunehmen, dass kürzere URLs mehr und relevantere Informationen enthalten als längere URLs, welche tiefer in der Webseitenhierarchie verortet sind.

---

1 Siehe https://www.wrg-goettingen.de/inno/innovationspreis-2021/wettbewerb, letzter Abruf 23.12.2021.

2 Das Erstellen umfangreicherer Unternehmenslisten ist sowohl wieder über das Webscraping von Unternehmensverzeichnissen selbst, durch den Erwerb bestehender Datensätze zu Unternehmensregistern von etablierten Anbietern oder durch Nutzung bestehender Listen jeglicher Art möglich.

3 Diese verteilen sich auf vier Homepages der Hochschule für angewandte Wissenschaft und Kunst (HAWK), 40 Institute der TU Clausthal, 13 Fakultäten und 32 Forschungszentren der Universität Göttingen sowie fünf Max-Planck-Institute.

4 Vgl. Kinne & Axenbeck (2018).

5 Vgl. Benoit et al. (2018).

Für die Analyse der Website-Inhalte werden die Volltexte zunächst bereinigt, was dem üblichen Vorgehen im Kontext einer automatisierten Textanalyse (Natural Language Processing) entspricht. Dabei werden Satzzeichen und Zahlen entfernt, häufige Worte entfernt (z.B. „ein“, „der“, „das“) und manuelle Wortlisten mit kontextspezifischen, inhaltlich nicht aussagekräftigen Begriffen definiert und anschließend entfernt (z.B. Städte- und Ortsnamen). Erst so entsteht ein geordneter Datensatz, der für inhaltliche Fragen zugänglich ist und zentrale Begriffe sichtbar macht.

Die vorliegende Studie konzentriert sich allerdings auf Möglichkeiten des Match-Making, sodass inhaltliche Analysen der Forschungseinrichtungen und Unternehmen in der Ergebnisdarstellung weitgehend ausgeklammert werden.[6] Vielmehr wurden beispielhafte Technologien definiert, für welche eine Koordination bzw. Zusammenführung von Unternehmen und Forschungseinrichtungen zum Zweck der Innovationsvernetzung relevant sein könnten. Die Begriffe stehen dabei in inhaltlichem Zusammenhang der im Kontext des Innovationspreises herausgestellten Technologien. Für das Webscraping genutzt werden die Begriffe:

„elektrolyse“, „3D-Druck“, „drohne“, “kreislaufwirtschaft“,

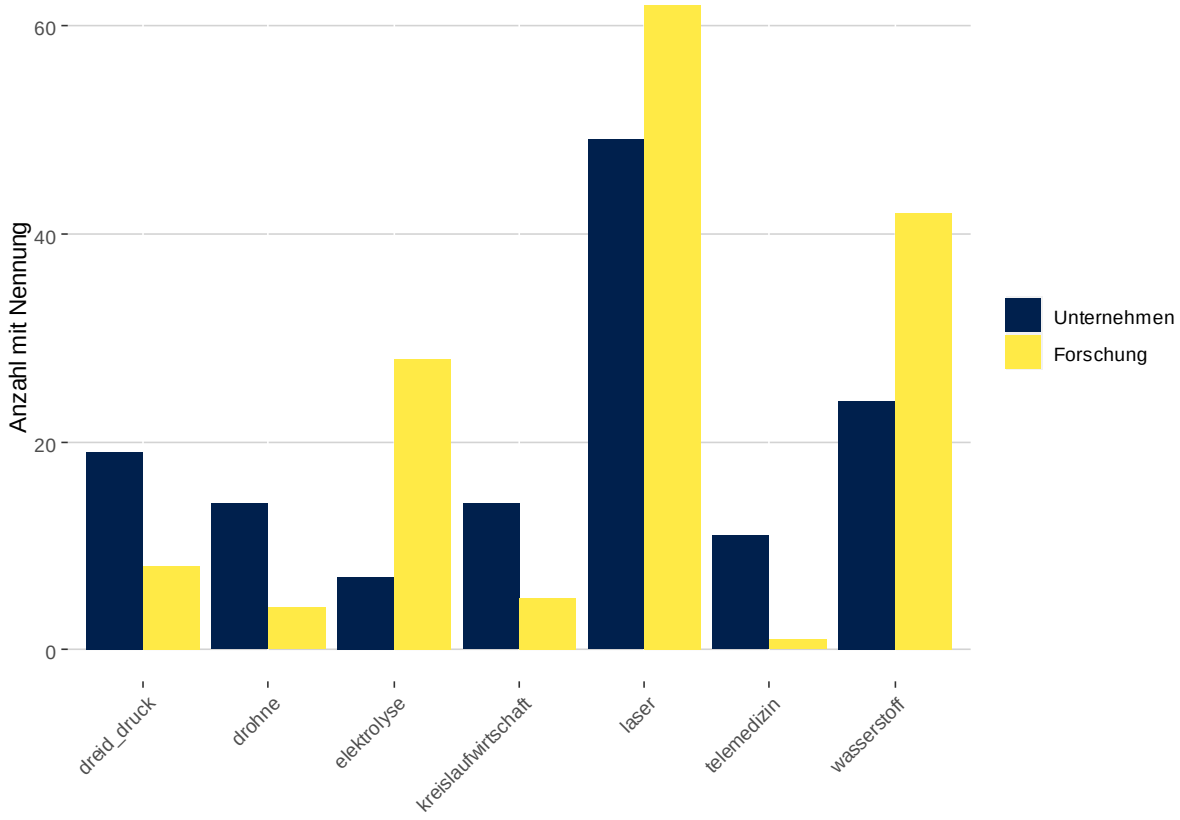
„laser“, „telemedizin“, „wasserstoff"

[6] Für Studien mit einem stärkeren Fokus auf die inhaltliche Analyse von Webscraping-Daten, siehe Proeger et al. (2021a) für Bildungseinrichtungen, Proeger et al. (2021b) für einen Handwerkskammerbezirk sowie Proeger et al. (2019) für einen deutschlandweiten Datensatz aus KMU.

3. Ergebnisse

Diese Studie stellt kompakt die zentralen Möglichkeiten der Innovationsvernetzung („Matching“) auf Basis von Webscraping-Daten dar. Abb. 1 zeigt die Relevanz der beispielhaft gewählten Technologien für Unternehmen und Forschungseinrichtungen einer Region insgesamt. So kann, vor einem spezifischen individuellen Matching, die generelle Relevanz einer Thematik zum Beispiel über die Anzahl der damit befassten Unternehmen und Forschungseinrichtungen erfasst werden. Es können Nischenthemen festgelegt werden, die ein sehr zielgerichtetes Matching erfordern, aber auch Themen mit weitergefassten Interessensgruppen, die ggf. allgemein beworben werden sollten.

Abb. 1: Häufigkeiten zur Nennung der Schlüsseltechnologien



ifh Göttingen

Quelle: eigene Darstellung

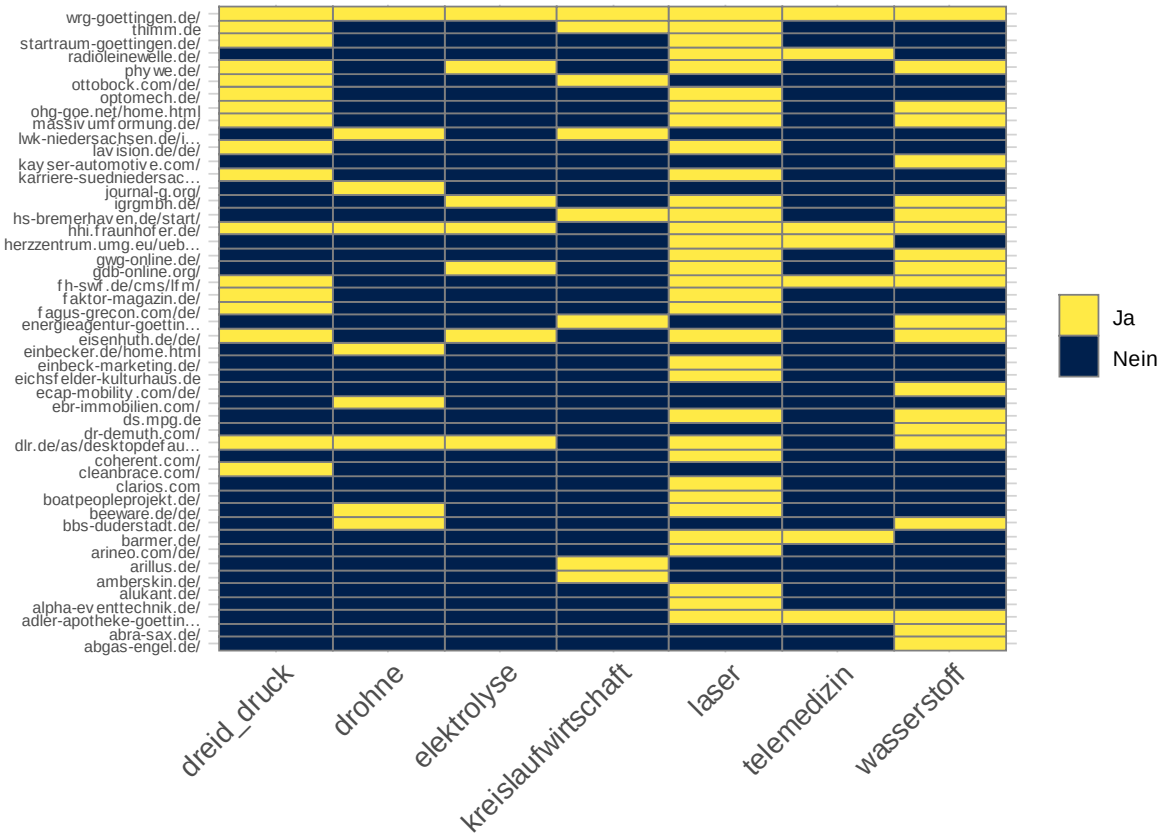
Für unsere Beispieldaten zeigt sich, dass „Laser“ oder „Wasserstoff“ Schlüsseltechnologien sowohl für Unternehmen als auch für Forschungseinrichtungen der Region darstellen. Es besteht folglich ein großes Potenzial, Akteure zu diesem Themenbereich erfolgreich in Netzwerke einzubinden und damit die Technologie in der Region weiter voranzubringen.

Die „Elektrolyse“ zeigt sich dagegen vor allem bei Forschungseinrichtungen relevant, sodass für ein Match-Making eine gezieltere Ansprache auf Unternehmensseite angezeigt wäre. Umgekehrt verhält es sich bei den Themen „Kreislaufwirtschaft“ oder „Telemedizin“, die eine

deutlich größere Relevanz für Unternehmen im Gegensatz zu Forschungseinrichtungen besitzen.

Für ein konkretes Matching sind sowohl Unternehmen als auch Forschungseinrichtungen mit den jeweiligen Technologien in Verbindung zu setzen. Abb. 2 und 3 visualisieren diesen Prozess und zeigen die Treffer, d.h. die Nennung einer Technologie für die analysierten Homepages an. Auf diese Weise kann effizient graphisch abgelesen werden, bei welchen Unternehmen eine Aktivität in den jeweils relevanten Themenbereichen erfolgt.

Abb. 2: Matching Unternehmen und Technologien



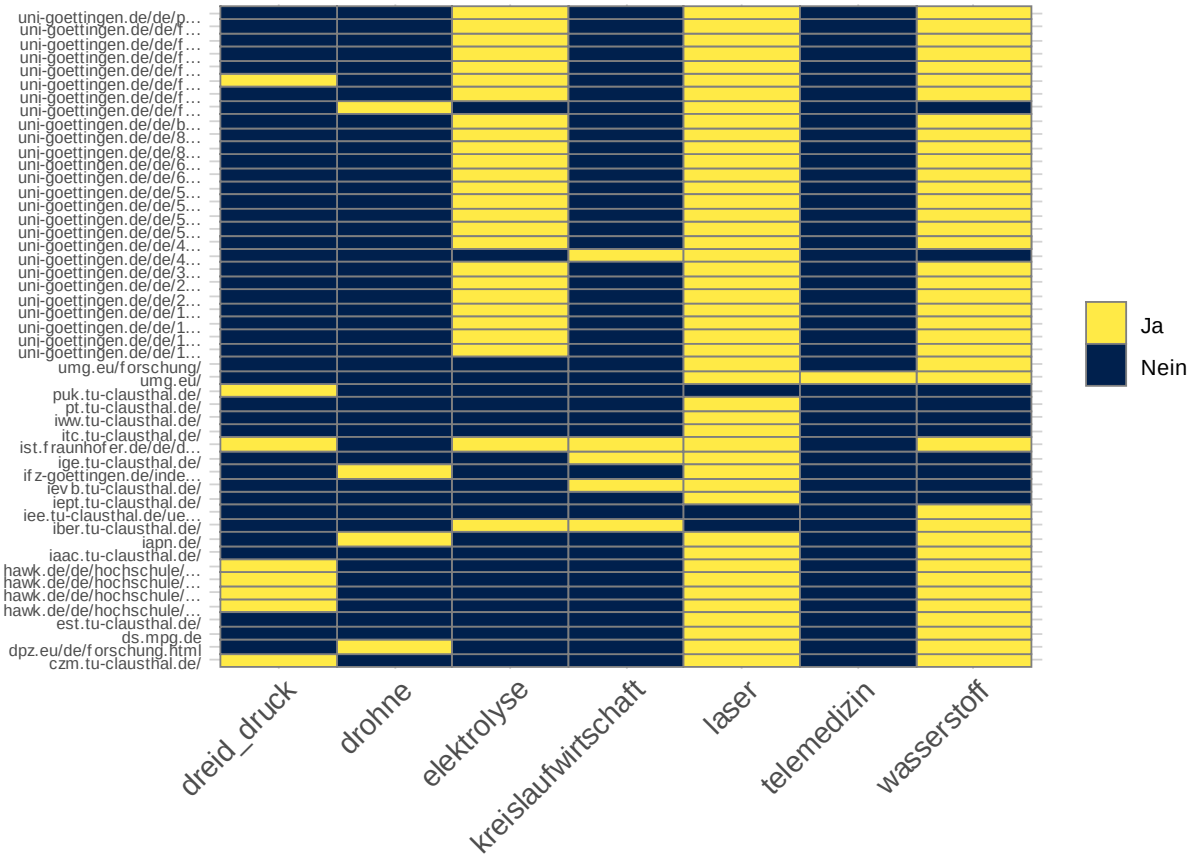
ifh Göttingen

Anmerkung: Darstellung für die 50 Unternehmen mit den meisten Technologienennungen.

Quelle: eigene Darstellung

Es zeigen sich bei einigen Unternehmen begriffliche Mehrfachnennungen über die Beispieltechnologien hinweg. Dabei finden sich zum einen große Unternehmen der Region (wie z.B. „Otto Bock“ oder „THIMM“), aber auch kleinere, innovative Betriebe (wie z.B. „AluKant“ oder „OPTOMECH“). Die zu Grunde liegende Liste der Unternehmen ist eher weit definiert, sodass zum Beispiel auch Organisationen wie „Arillus“ als potenzielle Interessenten für das Thema „Kreislaufwirtschaft“ identifiziert werden können.

Abb. 3: Matching Forschungseinrichtungen und Technologien



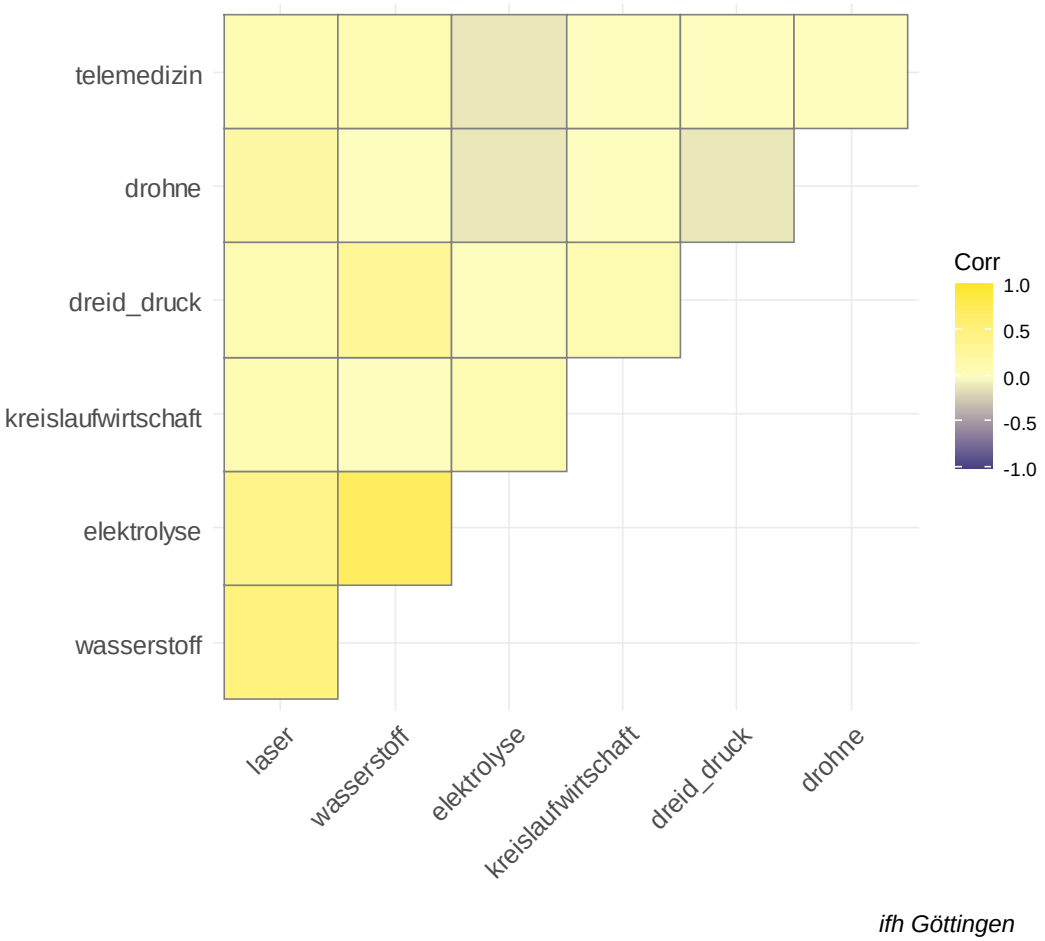
ifh Göttingen

Anmerkung: Darstellung für die 50 Unternehmen mit den meisten Technologienennungen.

Quelle: eigene Darstellung

Neben der direkten Vernetzung von Unternehmen und Forschungseinrichtungen, können auch erweiterte Interessengruppen über korrelierte Technologien erschlossen werden. So zeigt Abb. 4 beispielhaft diese Auswertungsdimension durch die Visualisierung von Technologien, die häufig gemeinsam bei den jeweiligen Forschungseinrichtungen aufgeführt werden. So findet sich zum Beispiel „Elektrolyse“ häufig in Verbindung mit „Wasserstoff“, sodass hier ein potenziell fruchtbares Matching abgeleitet werden kann, bei dem nicht Akteure einer spezifischen Technologie, sondern sich ergänzender Technologien zusammengebracht werden können.

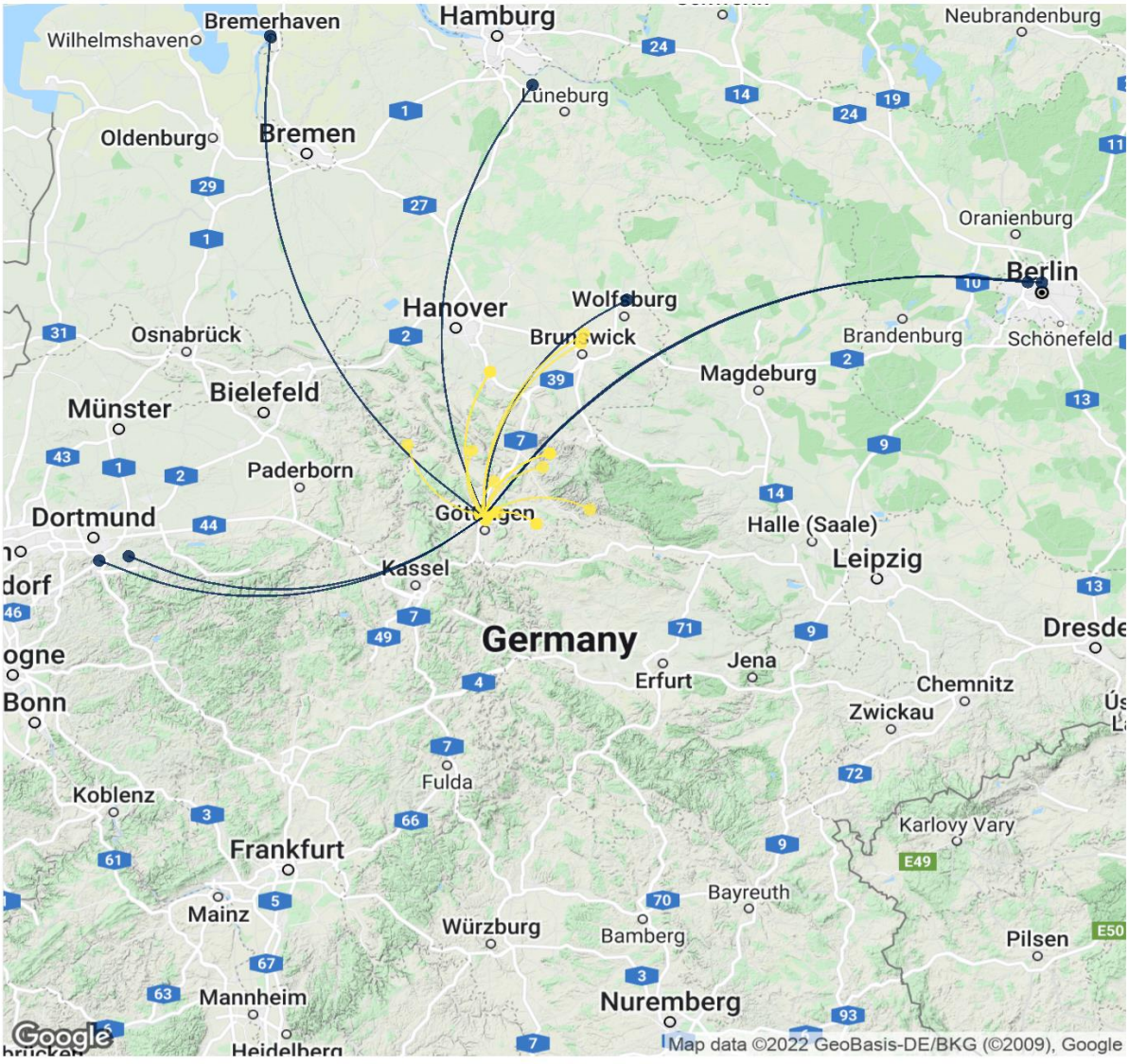
Abb. 4: Verknüpfung von Technologiefeldern



Quelle: eigene Darstellung

Abb. 5 zeigt die Möglichkeit, bestehende Netzwerke durch das Webscraping sichtbar zu machen, hier am Beispiel des Schlagworts „Wasserstoff“. Hier werden regionale Unternehmen und Forschungseinrichtungen dargestellt, sowie überregionale Partner im Kontext der Ausgabe des Innovationspreises zusätzlich eingezeichnet. Diese Netzwerke können auch technologieunabhängig anhand von Querverweisen der Akteure untereinander bestimmt werden. Je nach Qualität der Datengrundlage sind hier vielfältige Ausdifferenzierungen nach Eigenschaften der betrachteten Einheiten möglich.

Abb. 5: „Wasserstoff“ als Technologie in der Region und überregionale Partner



ifh Göttingen

Quelle: eigene Darstellung

4. Ausblick

Die vorliegende Studie zeigt illustrativ das Potenzial regionalen Webscrapings für die Zusammenführung von Firmen und Forschungseinrichtungen auf Basis gemeinsamer Themengebiete. Dabei kann sowohl explorativ untersucht werden, welche Themengebiete bzw. Technologien in einer Region besonders häufig vorkommen, oder aber gezielt Ansprechpartner für spezifische Themen gefunden werden.

Die Studie ist durch die kleine Stichprobe an Unternehmen der Region Südniedersachsen in ihrer Aussagekraft limitiert. Um eine umfassende Recherche und damit effektive Vernetzung von Unternehmen und Forschungseinrichtungen zu ermöglichen, ist eine Vollerhebung aller Unternehmen erforderlich. Ist diese erfolgt und wird regelmäßig aktualisiert, kann sowohl der explorative Ansatz dynamisch weitergeführt werden, um wandelnde Technologieschwerpunkte und Forschung und Unternehmen zu zeigen. Ebenso können neue Themen, die aus Sicht von Wirtschafts- und Innovationsförderung interessant sind, identifiziert oder kurzfristig und mit geringem Aufwand relevante Ansprechpartner aktiviert werden.

Das Potenzial der vorgestellten Methodik liegt folglich in der dynamischen und flexiblen Generierung von relevanten Informationen über ein regionales Innovationssystem. Die Nutzung kann durch regionale Wirtschaftsförderer und andere Netzwerkakteure erfolgen, die mittels Netzwerkaktivitäten, gemeinsamen Anträgen oder Projekten eine bessere Zusammenführung an spezifischen Themen interessierter Personen und Institutionen erreichen können. Die Nutzung von Webscraping stellt folglich eine deutliche Effizienzsteigerung des regionalen Match-Makings dar, eines der Hauptziele regionaler Innovationsförderung. Während dessen Effektivität in der Regel von persönlichen Interessen und langfristig aufgebauten Netzwerken der handelnden Personen abhängt, kann das Webscraping den Aspekt der Informationsgewinnung und Vernetzung deutlich effizienter und themenunabhängiger gestalten.

Von der vorgestellten Methodik des Webscrapings können eine Reihe weiterer Anwendungsfelder im Rahmen der regionalen Wirtschaftsförderung abgeleitet werden.

1. **Identifikation regionaler Experten:** Wissensintermediäre planen eine Vernetzungsveranstaltung beispielsweise zum Thema 3D-Druck mit möglichst allen relevanten wissenschaftlichen und unternehmerischen Akteuren der Region. Klassischerweise werden diese direkt über bestehende Mailinglisten sowie diffus über Soziale Medien eingeladen. Ggf. werden einzelne, persönlich bekannte Personen separat eingeladen. Die Reichweite der Einladung ist damit von vorneherein eng begrenzt und die Einladung wenig fokussiert auf tatsächlich interessierte Akteure. Im Gegensatz dazu liefert der Webscraping-Ansatz auf der Basis der Unternehmensdatenbank eine aktuelle Suche auf allen Homepages nach dem jeweiligen Begriff, um die so identifizierten Betriebe mit Verweis auf ihre Homepage einzuladen. Dasselbe gilt für Forscher.

2. **Ad-hoc Suchen nach Themenfeldern** vereinfachen die Suche nach regionalen Experten für spezifische Themen: Wenn von Seiten der Politik, von Wissensintermediären, von Unternehmensseite oder aus den Hochschulen Ansprechpartner zu bestimmten Themen benötigt werden, kann auf eine breite und aktuelle Basis zurückgegriffen werden und mögliche Ansprechpartner über die von

ihnen digital veröffentlichten Inhalte identifiziert werden. Dadurch verhindert man das häufig anzutreffende Phänomen, dass immer dieselben Akteure eingeladen werden.

3. **Vermittlung Kooperationspartner für angewandte Forschung:** Ebenso denkbar ist die Suche eines Wissenschaftlers nach Kooperationspartnern aus der Wirtschaft für ein anwendungsnahes Drittmittelprojekt. Er wendet sich an den Wissensintermediär, der ihm – erneut auf Basis eines aktuellen Webscrapings – Auskunft erteilt und entweder die Internetadressen der Unternehmen weitergibt oder auf deren Basis eine weitere Selektion über Technologieberater vornimmt. Der Wissenschaftler kontaktiert die ausgesuchten Unternehmen mit Verweis auf ihre Homepage und kann so deutlich fokussierter auf Kooperationspartner zugehen.

4. **Exploratives Vorgehen zur Identifikation von Trends:** Weiterhin können aktuelle wissenschaftliche und unternehmerische Trends frühzeitig erkannt werden. Beispielsweise kann wöchentlich überprüft werden, ob bestimmte aktuelle Begriffe verwendet werden und welche Branchenreaktionen sich daraus ergeben. Denkbar wäre zum Beispiel, die Thematisierung von Wasserstoff-Technologien in Südniedersachsen regelmäßig abzubilden. Hierdurch könnte der Aufbau von Projektverbünden unterstützt werden sowie Politik und Öffentlichkeit zum Fortschritt regionaler technologischer Ziele informiert werden.

5. **Analyse regionaler Innovationsschwerpunkte:** Eine branchenspezifische Analyse von Begriffsclustern ermöglicht es, die regionale Spezialisierung der Unternehmens- und Forschungslandschaft besser abzubilden und die regionale Innovationsförderung passgenauer auf diese Spezialisierungspfade abzustimmen. Die Webscraping-Analyse erreicht dabei eine Tiefe, Aussagekraft und Aktualität, die im Rahmen von Umfragen oder Workshops nicht erzielt werden kann.

5. Literatur

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S. & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software, 3 (30), 774. DOI: 10.21105/joss.00774, https://quanteda.io.

Kinne, J. & Axenbeck, J. (2018). Web mining of firm websites: A framework for web scraping and a pilot study for Germany. ZEW-Centre for European Economic Research Discussion Paper, No. 18-033.

Proeger, T., Meub, L. & Bizer, K. (2021a). Webscraping als Instrument zur tagesaktuellen und umfassenden Strukturanalyse des Handwerks. Göttinger Beiträge zur Handwerksforschung (Heft 55). Göttingen.

Proeger, T., Meub, L. & Pölert, H. (2021b). Analyse des Digitalisierungsgrads von Bildungseinrichtungen auf Basis von Webscraping – eine methodische Vorstudie. Göttinger Beiträge zur Handwerksforschung (Heft 56). Göttingen.

Proeger, T., Thonipara, A. & Bizer, K. (2019). Homepage-Nutzung im Handwerk – Eine sektorale und regionale Analyse. Göttinger Beiträge zur Handwerksforschung (Heft 27). Göttingen.