

Proeger, Till; Meub, Lukas; Pöler, Hauke

Research Report

Analyse des Digitalisierungsgrads von Bildungseinrichtungen auf Basis von Webscraping - eine methodische Vorstudie

Göttinger Beiträge zur Handwerksforschung, No. 56

Provided in Cooperation with:

Volkswirtschaftliches Institut für Mittelstand und Handwerk an der Universität Göttingen (ifh)

Suggested Citation: Proeger, Till; Meub, Lukas; Pöler, Hauke (2021) : Analyse des Digitalisierungsgrads von Bildungseinrichtungen auf Basis von Webscraping - eine methodische Vorstudie, Göttinger Beiträge zur Handwerksforschung, No. 56, Volkswirtschaftliches Institut für Mittelstand und Handwerk an der Universität Göttingen (ifh), Göttingen, <https://doi.org/10.3249/2364-3897-gbh-56>

This Version is available at:

<https://hdl.handle.net/10419/235250>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



DEUTSCHES HANDWERKSINSTITUT

Till Proeger, Lukas Meub, Hauke Pöler

**Analyse des Digitalisierungsgrads von
Bildungseinrichtungen auf Basis von
Webscraping**

– eine methodische Vorstudie

Göttinger Beiträge zur Handwerksforschung 56

**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:



aufgrund eines Beschlusses
des Deutschen Bundestages



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-56>

Alle Rechte vorbehalten

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 • 2021

Analyse des Digitalisierungsgrads von Bildungseinrichtungen auf Basis von Webscraping – eine methodische Vorstudie

Autoren: Till Proeger, Lukas Meub, Hauke Pöler
Göttinger Beiträge zur Handwerksforschung Nr.56

Zusammenfassung

Diese Vorstudie zeigt beispielhaft, wie mit der Methodik des Webscrapings, d.h. einem systematischen Abruf und statistischer Analyse von Webseiten, Erkenntnisse über Bildungseinrichtungen gewonnen werden. Anhand von zwei Teilstudien über 200 niedersächsische sowie 173 Bildungseinrichtungen aus Nordrhein-Westfalen wird eine strukturelle und inhaltsbezogene Analyse der Webseiten durchgeführt. Der Fokus liegt dabei auf dem Digitalisierungsgrad und der Nutzung digitaler Technologien durch die Bildungseinrichtungen. Übergreifend zeigt die Studie:

- Webscraping ist ein geeignetes Werkzeug zur umfassenden und tagesaktuellen Analyse von Bildungseinrichtungen insbesondere im Hinblick auf deren Digitalisierungsgrad und die Nutzung zentraler Technologien.
- Webscraping weist eine Reihe von Vorteilen gegenüber klassischen Befragungen auf, wobei insbesondere die effiziente Nutzung größerer Stichproben bei begrenztem Aufwand sowie die Möglichkeit zur wiederholten Abfrage von Informationen zentrale Vorteile darstellen.
- Die regelmäßige Analyse der Webseiten ermöglicht ein dauerhaftes regionales bis deutschlandweites Monitoring von Veränderungen im Digitalisierungsverhalten. Dies ermöglicht es, die Wirkung von Förderprogrammen oder Beratungsleistungen, etwa zwischen Bundesländern oder Schulträgern, zu evaluieren.
- Insbesondere können die Webseiten transparent in Bezug auf die Nutzung spezifischer Technologien, Apps, Plattformen oder sozialen Netzwerken analysiert werden. Hiermit lässt sich effizient zeigen, welche Schulen mit welchen Technologien werben, welche Regionen sich in technologischer Hinsicht dynamisch entwickeln und wo Marktpotenziale für einzelne Technologien bestehen.
- Webscraping ermöglicht es, in deutschlandweiten Stichproben bei Bildungseinrichtungen Vorreiter und Nachzügler im Hinblick auf das Digitalisierungsverhalten zu identifizieren, regionale, bundeslandbezogene und deutschlandweite Vergleichsgruppen zu bilden sowie diese kartografisch darzustellen.
- Ebenso kann der Zusammenhang der soziodemografischen Eigenschaften von Regionen mit den Ergebnissen des Webscrapings untersucht werden, etwa im Hinblick darauf, welche systematischen Unterschiede zwischen dem Digitalisierungsverhalten von Bildungseinrichtungen in Stadt und Land oder zwischen Regionen mit unterschiedlichem Wohlstandsniveau bestehen.

Übergreifend stellt Webscraping ein innovatives und effizient einsetzbares Instrument für die systematische Analyse der Entwicklung von Bildungseinrichtungen im Feld der Digitalisierung dar.

Schlagwörter: *Bildungseinrichtungen, Digitalisierung, Webscraping*

Inhalt

1.	Einleitung	1
2.	Studie I: Nutzungsmuster von Webseiten in Bildungseinrichtungen	3
3.	Studie II: Vollerhebung im Regierungsbezirk Detmold	5
3.1	Methodik und Datengrundlage	5
3.2	Ergebnisse	7
3.2.1	Umfang der Webseiten	7
3.2.2	Analyse der Verlinkungen	12
3.2.3	Analyse von Schlüsselworten	14
3.2.4	Digitalisierungsindikatorik	20
3.2.5	Wortkontextanalyse	21
3.2.6	X-Ray-Analyse von Webseiten	22
4.	Studie III: Zufallsstichprobe aller Schultypen in Niedersachsen	24
4.1	Methodik und Datengrundlage	24
4.2	Ergebnisse	26
4.2.1	Komplexität und Heterogenität der Schulwebseiten	26
4.2.2	Explorative Analyse: Sprache und Themensetzung	27
4.2.3	Aktuelle Bezüge: Schulwebseiten und Corona	28
4.2.4	Detailanalyse I: Digitalisierung auf den Schulwebseiten	30
4.2.5	Detailanalyse II: Fortbildung auf den Schulwebseiten	31
4.2.6	Linkanalyse: Social Media Nutzung	32
5.	Ausblick	34
6.	Literatur	36
7.	Anhang	37
7.1	Fragebogen	37
7.2	Regression Determinanten der Worthäufigkeiten	38

Abbildungen

Abb. 1:	Zielstellung und Nutzung der Webseite, Frage zwei des Fragebogens	3
Abb. 2:	Zielstellung und Nutzung der Webseite, Frage drei des Fragebogens	4
Abb. 3:	Schulen in der Stichprobe	6
Abb. 4:	Umfang der Webseiten	8
Abb. 5:	Wortwolke zu den Schulformen	9
Abb. 6:	Absolute Worthäufigkeiten	10
Abb. 7:	Relative Worthäufigkeiten Gymnasium zu Berufskollegs	11
Abb. 8:	Anzahl der Verlinkungen nach Schultypen	12
Abb. 9:	Häufigkeit der Verlinkungen nach Zielen	13
Abb. 10:	Absolute Häufigkeit der Digitalisierungsbegriffe nach Schultypen, aggregiert	15
Abb. 11:	Absolute Häufigkeit der Schulentwicklungsbegriffe nach Schultypen, aggregiert	15
Abb. 12:	Häufigkeit der einzelnen Digitalisierungsbegriffe nach Schultypen	16
Abb. 13:	Häufigkeit der einzelnen Schulentwicklungsbegriffe nach Schultypen	17
Abb. 14:	Häufigkeit der Digitalisierungsbegriffe in Bezug auf Schülerzahl und Schultypen	18
Abb. 15:	Häufigkeit spezifischer Digitalisierungsbegriffe nach Regionen	19
Abb. 16:	Häufigkeit der einzelnen Digitalisierungsbegriffe nach Indikatorwert	20
Abb. 17:	Kontextanalyse des Begriffs „Fortbildung“	21
Abb. 18:	X-Ray-Analyse der Digitalisierungsbegriffe	22
Abb. 19:	X-Ray-Analyse der Schulentwicklungsbegriffe	23
Abb. 20:	Umfang der Webseiten nach Schulform	26
Abb. 21:	Wortwolke für Grundschulen	27
Abb. 22:	Wortwolke für Fachschulen	28
Abb. 23:	Erwähnung von Corona nach Schulform	29
Abb. 24:	Erwähnung von Corona nach Schulgröße	29

Abb. 25:	Erwähnung Digitalisierung nach Schulform	30
Abb. 26:	Erwähnung Digitalisierung nach Schulträger	31
Abb. 27:	Erwähnung Fortbildung nach Schulform	32
Abb. 28:	Social Media-Nutzung nach Schulform	33

Tabellen

Tabelle 1:	Übersicht Stichprobe Regierungsbezirk Detmold	5
Tabelle 2:	Übersicht Datengrundlage Stichprobe Niedersachsen	25

Tabellen im Anhang

Tabelle A 1:	Regression Determinanten der Worthäufigkeiten	38
--------------	---	----

1. Einleitung

Der enorme digitale Entwicklungsschub, den viele Bildungseinrichtungen in den Jahren der Corona-Pandemie erleben, wird kurz- und mittelfristig die Schwerpunkte der Schul- und Unterrichtsentwicklung maßgeblich beeinflussen. Erstmalig stehen Lehrpersonen flächendeckend vor der Herausforderung, den Unterricht im Kontext der herausfordernden Szenarien und Corona-Lagen flexibel als Präsenz-, Hybrid- oder Online-Konzept zu planen. Zugleich zeigen die Erfahrungen der Jahre 2020 und 2021, welche bedeutende Rolle Schulen in der Vermittlung zeitgemäßer Bildung und digitaler Kompetenzen zukommt. Digitalisierung ist nicht erst durch Corona zu einem unmittelbar drängenden Thema geworden. Der zusätzlich aufgebaute Entwicklungsdruck für alle beteiligten Institutionen - von Regierungen über die Schulbehörden und Schulträger bis zur Einzelschule - erfordert jedoch umso mehr die Orientierung an Vorreiter-Schulen und adaptierbaren Best-Practice-Beispielen.

In der Studie „Digitales Potenzial“¹ werden anhand von ICILS-2018-Daten Erfolgsfaktoren „digitaler Optimalschulen“ zusammengestellt. Diese definieren sich letztlich aus dem Zusammenspiel von:

- technischer Ausstattung,
- kontinuierlicher Fortbildung des Kollegiums,
- reflektiertem Einsatz digitaler Medien im Unterricht
- und besonders intensiver Weiterbildung mit fachlichem Schwerpunkt.

Der einseitige Fokus der aktuellen bildungspolitischen Debatten auf die Infrastruktur, insbesondere begründet durch den seit Jahren angekündigten „DigitalPakt Schule“ greift demnach zu kurz. Ebenso müssen Lehrkräfte darin unterstützt werden, digitale Technologien fachbezogen im Unterricht einzusetzen und die Kompetenzen der Schülerinnen und Schüler für „Bildung in der digitalen Welt“² zu fördern.

Während also die Anforderungen klar definiert sind, fehlt ein Überblick über tatsächlich geplante und durchgeführte Entwicklungsmaßnahmen der Einzelschulen - kurz: eine Indikatorik für digitale Schulentwicklung. Von Interesse für die Wissenschaft wären insbesondere Strukturmerkmale von Digitalisierung, die - isoliert oder verbunden betrachtet - als Planungsgrundlage für Beratung, Strategiefindung und Umsetzungskonzepte dienen könnten:

- räumlich (In welchen Regionen sind Bildungseinrichtungen besonders aktiv?)
- soziodemografisch (Lassen sich Rückschlüsse auf die Relation von Einzugsgebiet und schulischer Aktivität ziehen?)
- Schulform (Bestehen Unterschiede zwischen den Schulformen?)

¹ Vgl. Eickelmann, B. & Drossel, K. (2020).

² KMK-Strategie „Bildung in der digitalen Welt“, in: <https://www.kmk.org/aktuelles/artikelansicht/strategie-bildung-in-der-digitalen-welt.html> [Abrufdatum: 01.06.2021].

Die Ergebnisse könnten auf unterschiedlichen Ebenen interessant sein:

- Sind Stadt-Land-Gegensätze zu beobachten?
- Lassen sich Rückschlüsse auf Schulträgerebene ziehen?
- Welche Schulform ist (ggf. in welcher Region) besonders aktiv?
- Lassen sich daraus Modellbereiche (Schulträger, Schulformen, ...) identifizieren, die anderen als Orientierung dienen könnten?
- Welche Schulen können als Best-Practice-Beispiele herangezogen werden?
- Wo besteht Beratungsbedarf?

Da bislang kein umfassender Überblick über Digitalisierungsaktivitäten auf Einzelschulebene existiert, bietet sich für eine Primärauswertung ein Zugriff auf eine unmittelbar verfügbare Datenbasis an: die Webseiten von Schulen. Diese bilden in der Regel zuverlässig und aktuell schulisches Engagement ab, einerseits zur Information der Schulgemeinschaft, andererseits angesichts des zunehmenden Wettbewerbs zwischen den Schulformen auch aus Werbemotiven. Diese Einschätzung bestätigt auch eine Umfrage unter Schulleitungen verschiedener Schulformen, die für diese Vorstudie durchgeführt wurde.

Eine Auswertung aller Einzelschulen nach ihren Strukturparametern und nach Digitalisierungsthemen bedarf allerdings einer Methodik, die trotz potenziell nicht beeinflussbarer Variablen (z.B. individuelle Pflege der Schulwebseite) in der Lage ist, mit angemessenem Aufwand repräsentative Aussagen zu ermöglichen. Eine Möglichkeit, diese Anforderung zu erfüllen, ist das sog. "Web scraping". Dieses besteht aus einem Software-basierten systematischen Abrufen von technischen Informationen und Volltexten einer großen Anzahl an Webseiten und einer anschließenden statistischen Analyse der resultierenden Datensätze. Mit Hilfe dieser Methodik kann eine sehr große Anzahl an Webseiten effizient analysiert werden, es werden lediglich die entsprechenden Internetadressen benötigt; eine darüberhinausgehende manuelle Betrachtung der Webseiten kann die Analyse ergänzen, erfordert aber einen erhöhten Zeitaufwand. Insgesamt können damit sowohl grundlegende Struktureigenschaften einer großen Stichprobe analysiert als auch einzelne Schulen mit spezifischen relevanten oder beachtenswerten Eigenschaften aus einer großen Stichprobe identifiziert werden.

Diese Studie enthält drei Teilanalysen. Eine Web scraping-Analyse kann nur verlässliche Ergebnisse liefern, wenn die zu Grunde liegenden Webseiten relevante Informationen auf einem aktuellen Stand enthalten. Dies wird exemplarisch anhand einer kurzen Online-Befragung überprüft, an der 19 Schulleitungen teilgenommen haben und die als „Studie I“ in Kapitel zwei dokumentiert ist. Im Anschluss wird in Studie II eine Vollerhebung mit anschließendem Web scraping für alle Gymnasien, Gesamtschulen sowie Berufskollegs des Regierungsbezirks Detmold in Nordrhein-Westfalen durchgeführt. In Studie III erfolgt eine Web scraping-Analyse für eine (Zufalls-)Stichprobe von 201 niedersächsischen Schulen verschiedener Schultypen, Träger und Regionen. Beide Stichproben zeigen exemplarisch die Möglichkeiten und Chancen der Methodik Web scraping für die Analyse von Digitalisierungsprozessen in Bildungseinrichtungen.

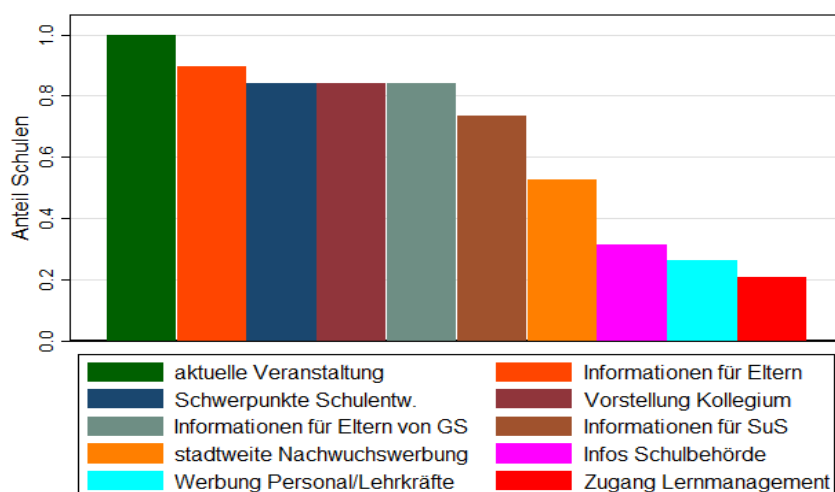
Diese Studie hat die folgende Gliederung: Kapitel zwei dokumentiert die Vorstudie zur Webseitenutzung in Schulen. Kapitel drei enthält die erste Teilstudie zum Regierungsbezirk Detmold, Kapitel vier die Web scraping-Analyse einer Zufallsstichprobe niedersächsischer Schulen und Kapitel fünf eine abschließende Diskussion der Potenziale des Web scrapings für die Analyse von Bildungseinrichtungen.

2. Studie I: Nutzungsmuster von Webseiten in Bildungseinrichtungen

Die Verlässlichkeit der Methodik des Webscraping ist abhängig von der Qualität und Aussagekraft der jeweils analysierten Webseiten. Indikatoren zum Umfang der Seiten und die Suche nach gewissen Kernthemen bieten hier zwar eine Möglichkeit der Validierung der Qualität, doch eine robuste Legitimationsgrundlage lässt sich nur durch eine qualitative Überprüfung durchführen. Hierbei sind die Schulen als Webseitenbetreiber zu befragen, um zu erfahren, wie aktuell und aussagekräftig die Webseiten gehalten werden und welche Rolle sie für die Außendarstellung spielen. Um diese Herausforderung für die Methodik zu adressieren, wurde im Rahmen dieser Vorstudie eine kurze Befragung unter Schulleitungen durchgeführt. Hierzu wurden insgesamt 25 Schulen gebeten, einen kurzen Online-Fragebogen auszufüllen (der Fragebogen ist in Kap. 7.1 im Anhang dokumentiert). 19 Schulen antworteten, was einer Rücklaufquote von 76 % entspricht. Die Schulen wurden auf Basis des persönlichen Netzwerks der Autoren angefragt und befinden sich in Niedersachsen, Nordrhein-Westfalen und Hessen. Die Stichprobe enthält neun Gymnasien, drei Oberschulen, drei integrierte Gesamtschulen und drei Grundschulen. Die Erhebung erfolgte vollständig anonym, sodass keine Rückschlüsse auf die konkrete Schule möglich sind. Der geringe Stichprobenumfang zeigt, dass hier nur Tendenzen abgeleitet werden können, die anhand umfangreicherer Fragebögen mit mehr Teilnehmern validiert werden müssen.

Die Webseiten der untersuchten Schulen weisen insgesamt eine hohe Inhaltsdichte und Aktualität auf. Abb. 1 und 2 zeigen kompakt die Ergebnisse der Umfrage. Aus Abb. 1 wird ersichtlich, dass alle Befragten angeben, dass die Schulseite für den Hinweis auf aktuelle Veranstaltungen genutzt wird. Rund 80 % der Befragten geben an, dass die Schulseite Informationen für Eltern, Schwerpunkte der Schulentwicklung, die Vorstellung des Kollegiums sowie Informationen für Eltern von Grundschulern (GS) enthält. Rund 75 % geben an, dass Information von Schülerinnen und Schüler (SuS) Inhalte auf der Schulseite haben. Nur rund 50 % nutzen die Webseite für die Werbung von neuen SuS, während die Bereiche Informationen für die Schulbehörde, Werbung von Personal/Lehrkräften sowie der Zugang zu Lernmanagementsystemen nur von 30 % bzw. 20 % der Befragten genannt werden.

Abb. 1: Zielstellung und Nutzung der Webseite, Frage zwei des Fragebogens

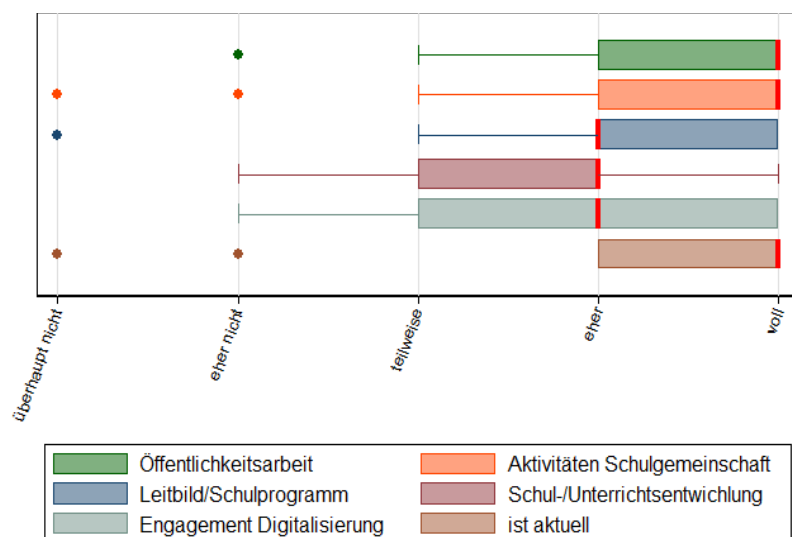


ifh Göttingen

Quelle: Eigene Erhebung

Im Hinblick auf eine übergreifende Bewertung zeigt Abb. 2, dass die Mehrzahl der Befragten die Schulwebseite für die Öffentlichkeitsarbeit sowie für Informationen zu Aktivitäten der Schulgemeinschaft nutzt und diese als aktuell einschätzt. In etwas geringerem Maße, aber dennoch insgesamt stark ausgeprägt, wird die Seite zur Präsentation von Leitbild/ Schulprogramm und der Schul-/ Unterrichtsentwicklung sowie zur Verdeutlichung des Engagements im Bereich der Digitalisierung genutzt.

Abb. 2: Zielstellung und Nutzung der Webseite, Frage drei des Fragebogens



ifh Göttingen

Quelle: Eigene Erhebung

Übergreifend kann somit festgehalten werden: Die Umfrage zeigt, dass Schulwebseiten eine geeignete Informationsquelle für eine Reihe von Aspekten der Schulentwicklung darstellen. Trotz der geringen Teilnehmerzahl und potenzieller Selektionsprobleme der Stichprobe kann dies als ein erster Indikator für die Nutzbarkeit des Webscraping-Instruments für die Analyse von Schulen gewertet werden.

3. Studie II: Vollerhebung im Regierungsbezirk Detmold

3.1 Methodik und Datengrundlage

Ausgangspunkt der Studie ist eine Vollerhebung der Adressen der Schulwebseiten von Gymnasien, Gesamtschulen und Berufskollegs im Regierungsbezirk Detmold. Alle Informationen zu den betreffenden Schulen wurden den offiziellen Seiten des Nordrhein-Westfälischen Schulministeriums entnommen.³ In Summe bilden 173 Schulen des Regierungsbezirks Detmold die Datengrundlage, deren Details in Tabelle 1 gezeigt werden.⁴ Abb. 3 gibt einen kartografischen Überblick für die Datengrundlage. Für das Webscraping entscheidend ist, dass im Durchschnitt 94 % der Schulen eine Webseite haben, sodass eine solide und breite Datenbasis vorhanden ist. Die Anzahl der Schulen zwischen den drei Schulformen ist dabei ausgeglichen, ebenso das Verhältnis zwischen privaten und staatlichen Institutionen.

Tabelle 1: Übersicht Stichprobe Regierungsbezirk Detmold

		Anzahl	Privat		Ganztagsschule		Anzahl Schüler	Website erfasst	
		#	#	%	#	%	Av.	#	%
Schul-form	Gesamt-schule	45	5	11	41	91	953	45	100
	Gymnasium	70	14	20	15	21	826	69	99
	Berufskolleg	58	27	47	0	0	1139	49	84
	Gesamt	173	46	26,6	56	32	964	163	94

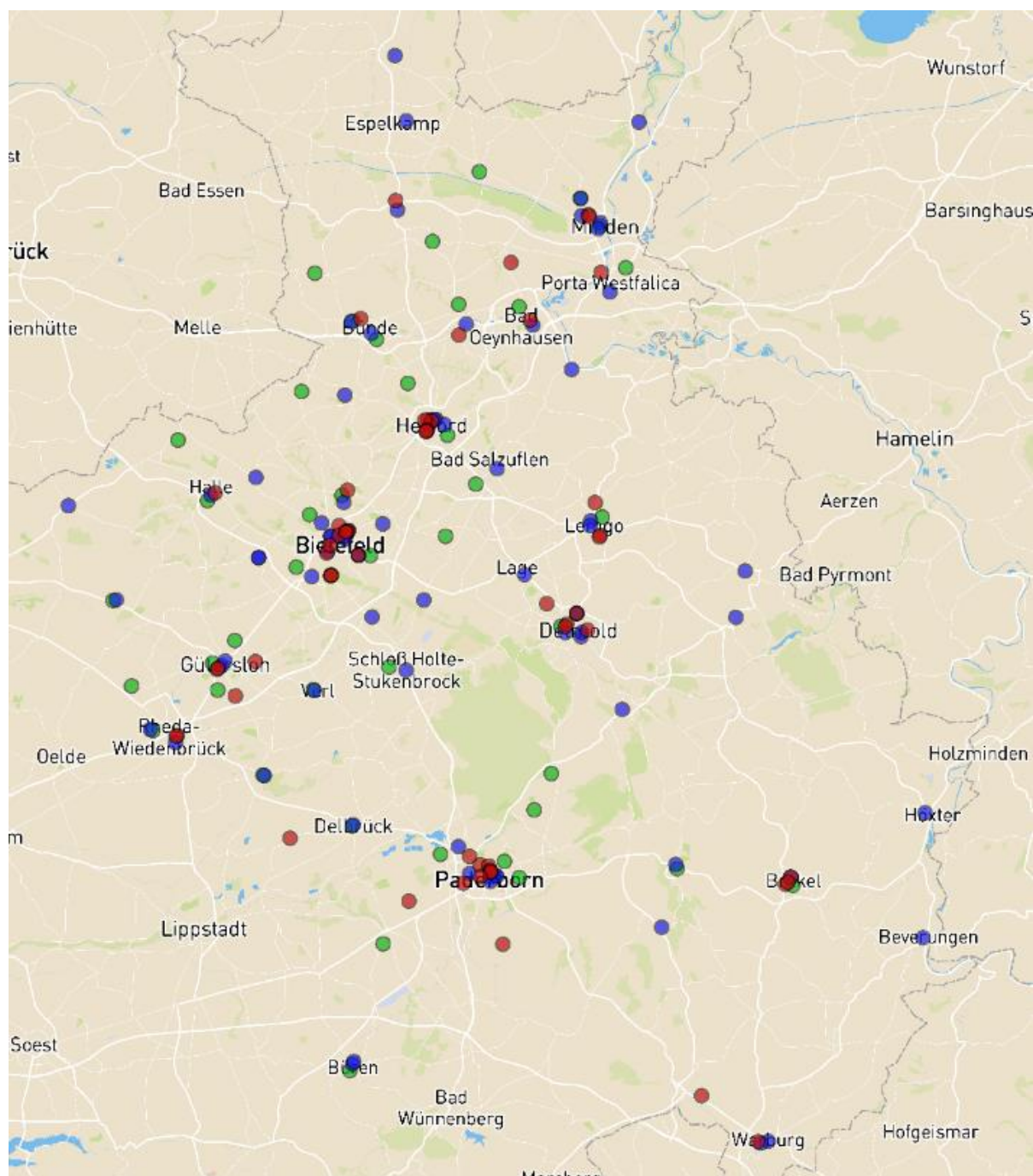
ifh Göttingen

Quelle: <https://www.schulministerium.nrw.de/BiPo/SchuleSuchen/>, [letzter Abruf: 17.06.2021], eigene Berechnungen

³ <https://www.schulministerium.nrw.de/BiPo/SchuleSuchen/> [letzter Abruf: 17.06.2021]

⁴ Die Aufteilung auf die Kreise ist: Herford: 24 Schulen, Bielefeld: 29 Schulen, Lippe: 26 Schulen, Minden-Lübbecke: 22 Schulen, Höxter: 15 Schulen, Paderborn: 27 Schulen, Gütersloh: 30 Schulen.

Abb. 3: Schulen in der Stichprobe



ifh Göttingen

Quelle: eigene Berechnungen

Anmerkung: Grün: Gesamtschulen, Blau: Gymnasien, Rot: Berufskollegs

Die Webseiten selbst wurden im März 2021 mittels des ARGUS-Programms⁵ gescraped und mit den Statistikprogrammen R und Stata ausgewertet.⁶ ARGUS sammelt zum einen die

⁵ Vgl. Kinne, J. & Axenbeck, J. (2018).

⁶ Vgl. Benoit K. et al. (2018).

Volltexte der Webseiten und zum anderen die Verlinkungen auf andere Online-Präsenzen. Für beide Ansätze wird der Zeitpunkt des Zugriffs dokumentiert (Zeitstempel), ob eine Fehlermeldung im Zugriff auf die Seite entsteht (Fehler) oder eine Weiterleitung zu einer anderen Website vorgenommen wurde (Weiterleitung). 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 Schulen der Stichprobe bezieht; „extern“ sind somit alle weiteren Webseiten. Um die Datenmenge des Scrapings zu begrenzen, wurde die Untersuchung auf maximal 50 Unterseiten beschränkt. 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.

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 fallengelassen (z.B. „ein“, „der“, „das“) und manuelle Wortlisten mit kontextspezifischen, inhaltlich nicht aussagekräftigen Begriffen definiert und anschließend entfernt (z.B. „Schule“, „Lehrer“ oder Städte- und Ortsnamen). Erst so entsteht ein geordneter Datensatz, der für inhaltliche Fragen zugänglich ist und zentrale Begriffe sichtbar macht.

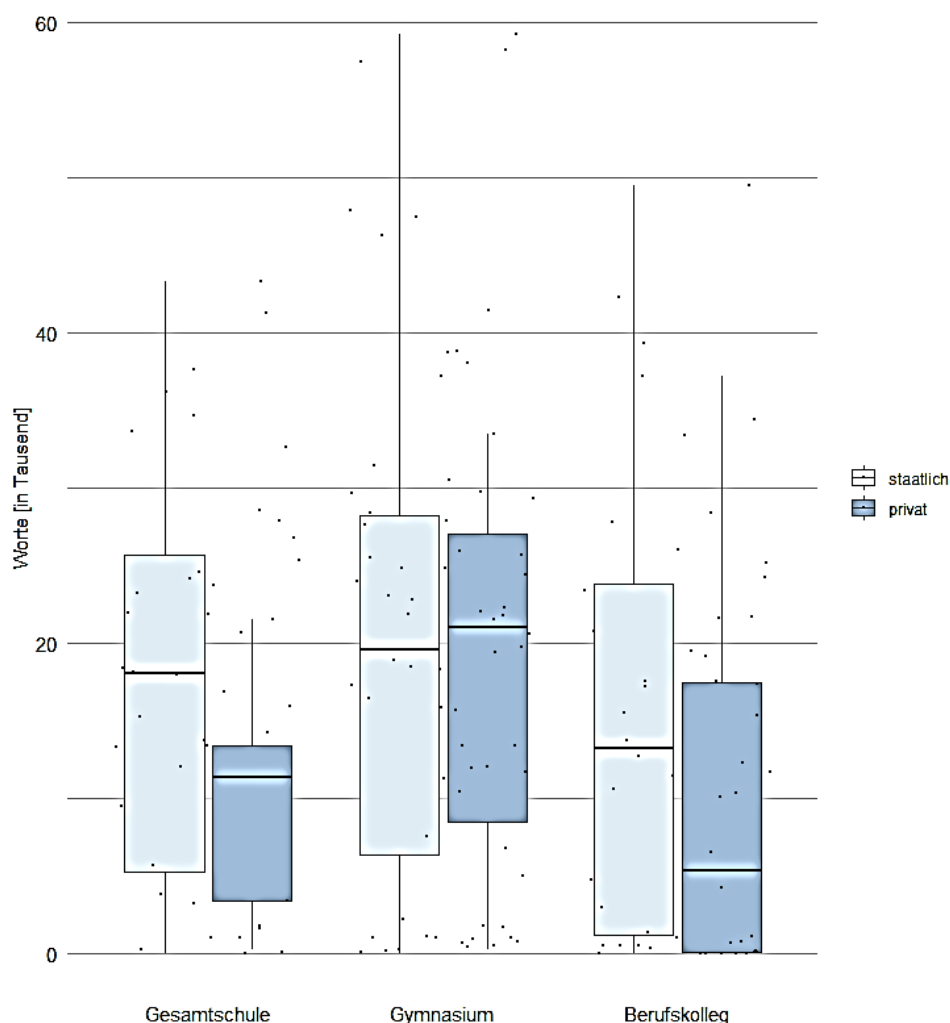
3.2 Ergebnisse

Die Analyse der Stichprobe der Schulen aus dem Regierungsbezirk Detmold wird schrittweise aufgebaut, wobei zunächst die Rahmendaten der Webseiten und im Anschluss daran zentrale Schlagworte im Bereich der Digitalisierung und Schulentwicklung betrachtet werden.

3.2.1 Umfang der Webseiten

Zunächst kann der durchschnittliche Umfang der Webseiten (Worte pro Webseite) in Bezug auf die Schultypen und nach der Unterscheidung privat / staatlich dargestellt werden. Dabei zeigt sich grundlegend, dass private Gymnasien die durchschnittlich höchste Wortanzahl aufweisen, wobei staatliche Gymnasien und staatliche Gesamtschulen einen ähnlichen Wert zeigen. Deutlich geringere Werte weisen private Gesamtschulen und private Berufskollegs auf.

Abb. 4: Umfang der Webseiten



ifh Göttingen

Quelle: eigene Berechnungen

Anmerkung: Hier dargestellt sind Box-Plots, die sich besonders für eine kompakte Darstellung der Verteilung einer Merkmalsausprägung eignen. Die einzelnen schwarzen Punkte repräsentieren immer eine Schule, deren Verteilung in den Boxen zusammengefasst werden. Der untere Rahmen der Box beginnt bei dem ersten Quartil der Verteilung, d.h. 25 % der Werte sind kleiner gleich diesem Wert. Für staatliche Gesamtschulen liegen also 25 % unter einem Wert von etwa 6.000 Wörtern auf der Webseite. Die mittlere Linie in der Box entspricht dem Median, der die oberen und die unteren 50 % einer Verteilung voneinander trennt. Für staatliche Gesamtschulen liegen also 50 % der Schulen über etwa 19.000 Wörtern auf der Webseite und 50 % darunter. Der obere Rahmen der Box entspricht dem dritten Quartil, also 75 % der staatlichen Gesamtschulen liegen unter etwa 26.000 Wörtern auf der Webseite. Die "Antennen" außerhalb der Box ("Whiskers") zeigen die Streuung der Verteilung über die mittleren 50 % hinaus.

Ergebnis 1: Es bestehen deutliche Unterschiede im Umfang der Webseiten, wobei private Gymnasien die umfangreichsten und private Berufskollegs die umfangschwächsten Webseiten aufweisen.

Als Einstieg in eine explorative Analyse, stellt Abb. 5 die wichtigsten Begriffe auf den jeweiligen Webseiten in Form einer Wortwolke dar. Die Schriftgröße korreliert dabei mit der Begriffshäufigkeit, sodass ein erster Überblick zu den verschiedenen inhaltlichen Schwerpunkten der jeweiligen Schulformen gewonnen wird.

Abb. 5: Wortwolke zu den Schulformen



ifh Göttingen

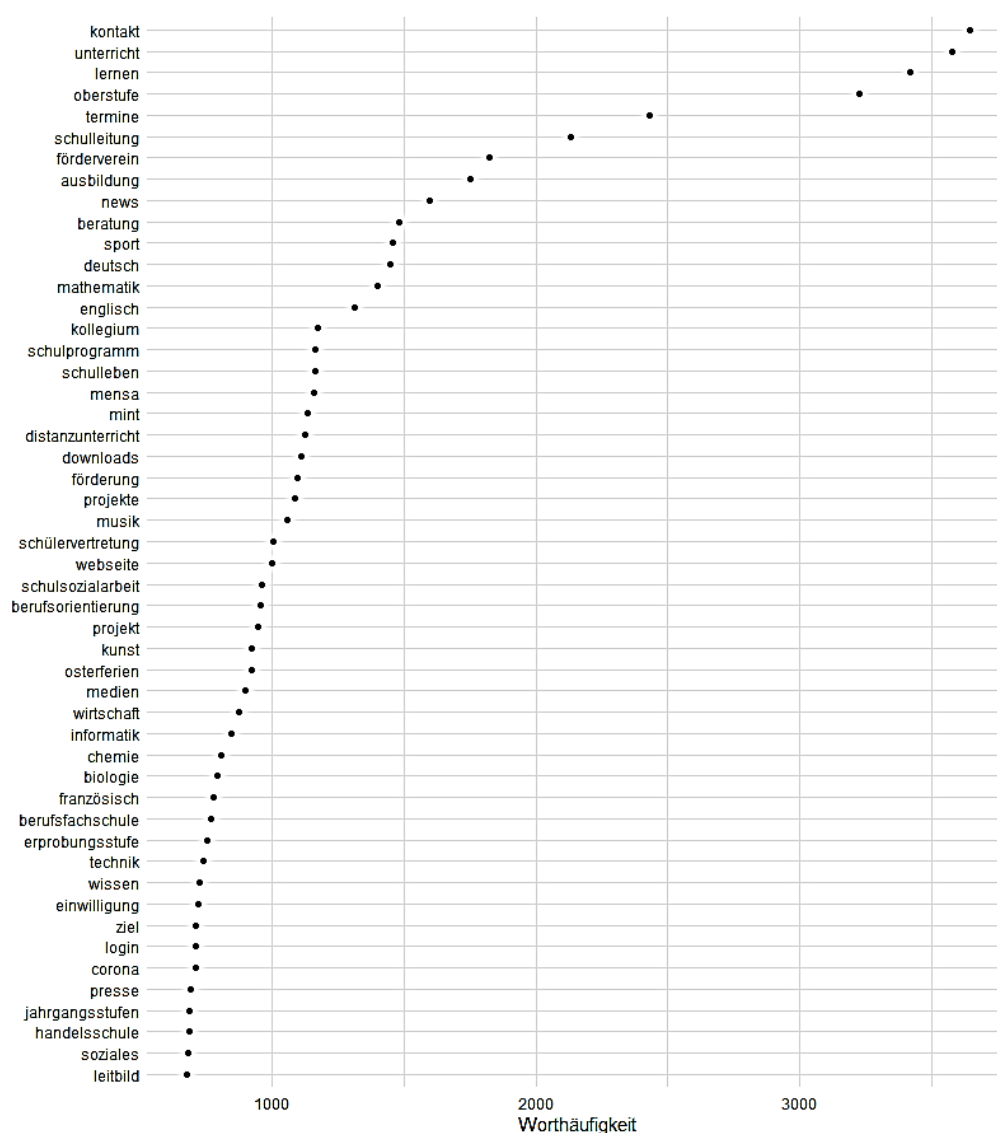
Quelle: eigene Berechnungen

Anmerkung: Je größer die Darstellung der Worte, desto häufiger die Wortnennung

Ergebnis 2: Eine Wortwolkenanalyse zeigt deutliche Unterschiede in den thematischen Schwerpunkten der betrachteten Webseiten für die verschiedenen Schulformen, was auf eine zielgruppenspezifische Ansprache durch die Bildungsinstitutionen schließen lässt.

Die Analyse nach Worthäufigkeiten kann weiter vertieft werden. Abb. 6 zeigt dabei zunächst die Schwerpunkte der Informationsarbeit der Schulwebseiten auf: Im Mittelpunkt steht der Kontakt zur Schule, die Informationen zum Unterricht bzw. zur Ausbildung, zu Terminen und zu News, zum Distanzunterricht, zur Schulleitung und zum Kollegium sowie zum Schulprogramm. Informationen zu einzelnen häufigen Fächern tauchen ebenfalls auf, jedoch relativ seltener als die grundlegenden Informationen, die weitgehend alle Schulwebseiten aufweisen. Diese Informationen sprechen für eine relativ hohe Aktualität und Aussagekraft der betrachteten Webseiten, auch im Hinblick auf aktuelle Entwicklungen und Strategien. Abb. 7 gibt beispielhaft einen Vergleich von Worthäufigkeiten zwischen Gymnasien und Berufskollegs, aus dem die unterschiedlichen Schwerpunkte deutlich werden. Beispielhaft taucht das Wort „Ausbildung“ praktisch nie bei Gymnasien auf, während das Wort „Erprobungsstufe“ relativ häufig in Gymnasien auftritt, aber praktisch nie in Berufskollegs. Insgesamt bietet diese Darstellungsform eine kompakte Übersicht über spezifische Profile zwischen abgrenzbaren Teilgruppen innerhalb der Stichprobe.

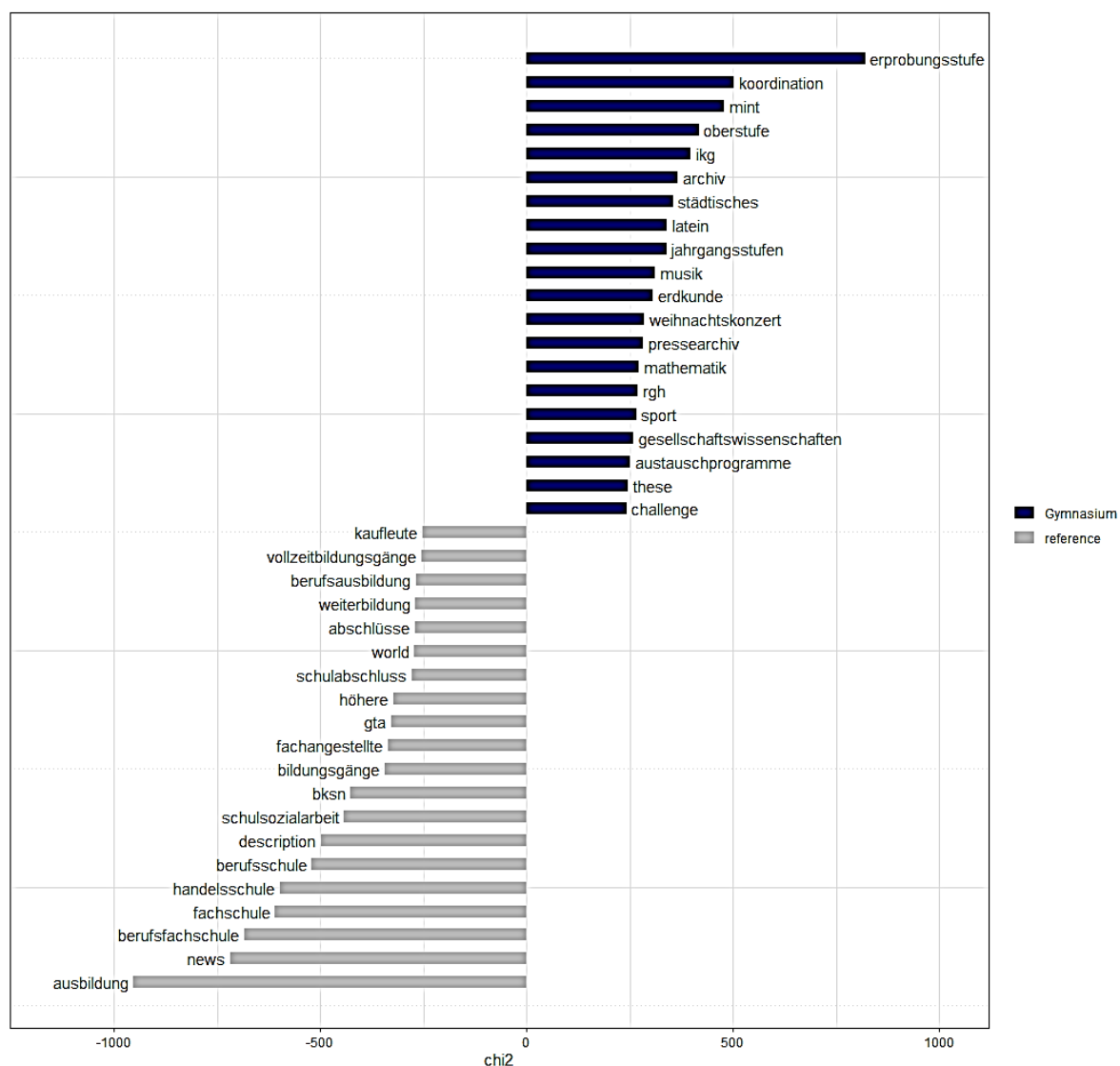
Abb. 6: Absolute Worthäufigkeiten



ifh Göttingen

Quelle: eigene Berechnungen

Abb. 7: Relative Worthäufigkeiten Gymnasium zu Berufskollegs



ifh Göttingen

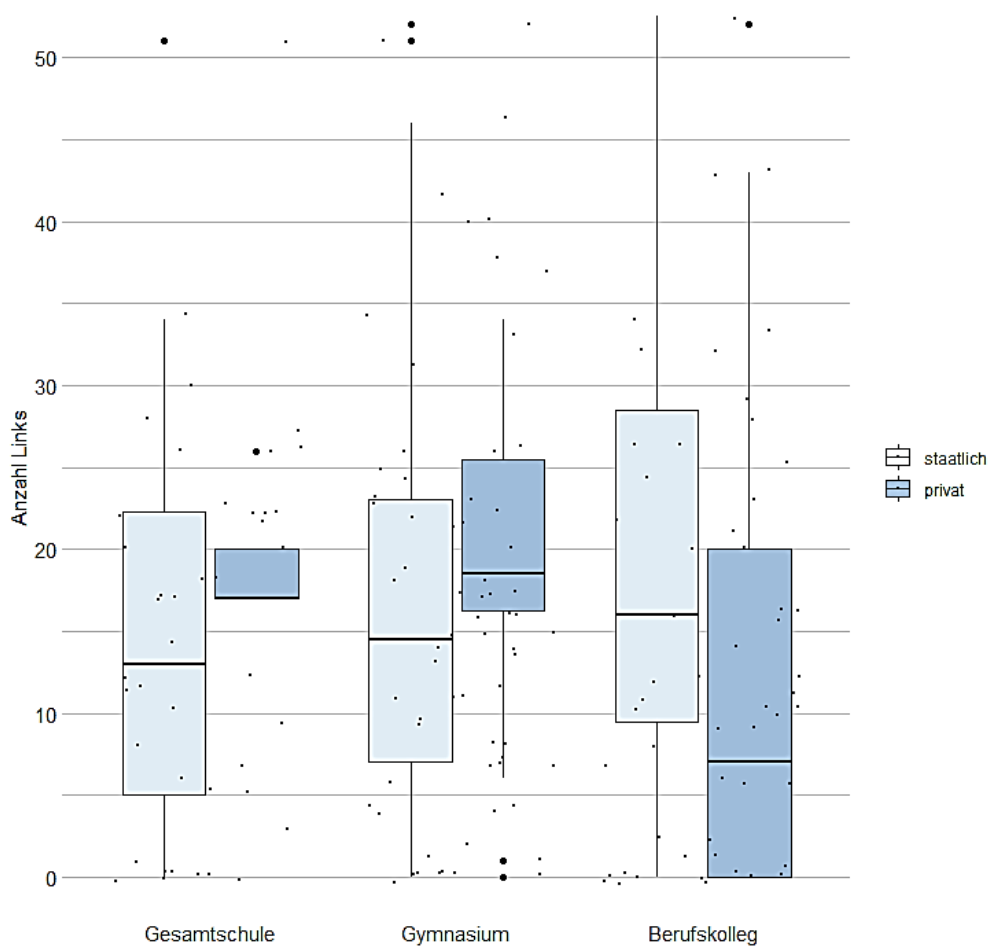
Quelle: eigene Berechnungen

Ergebnis 3: Die Worthäufigkeiten zeigen, dass die Webseiten strukturelle Informationen zu Schulaktivitäten, Strategien und aktuellen Entwicklungen ermöglichen. Die zentralen Begriffe unterscheiden sich dabei deutlich zwischen den Schultypen.

3.2.2 Analyse der Verlinkungen

Im nächsten Schritt können die Verlinkungen der Webseiten betrachtet werden. Hierbei wird erneut die Anzahl der Links zwischen Schultypen verglichen, wobei sich ein ähnliches Bild abzeichnet wie beim Umfang der Webseiten. Deutlich höher liegt jedoch die Anzahl der Links bei privaten Gesamtschulen, die sich auf dem Niveau privater Gymnasien bewegt.

Abb. 8: Anzahl der Verlinkungen nach Schultypen

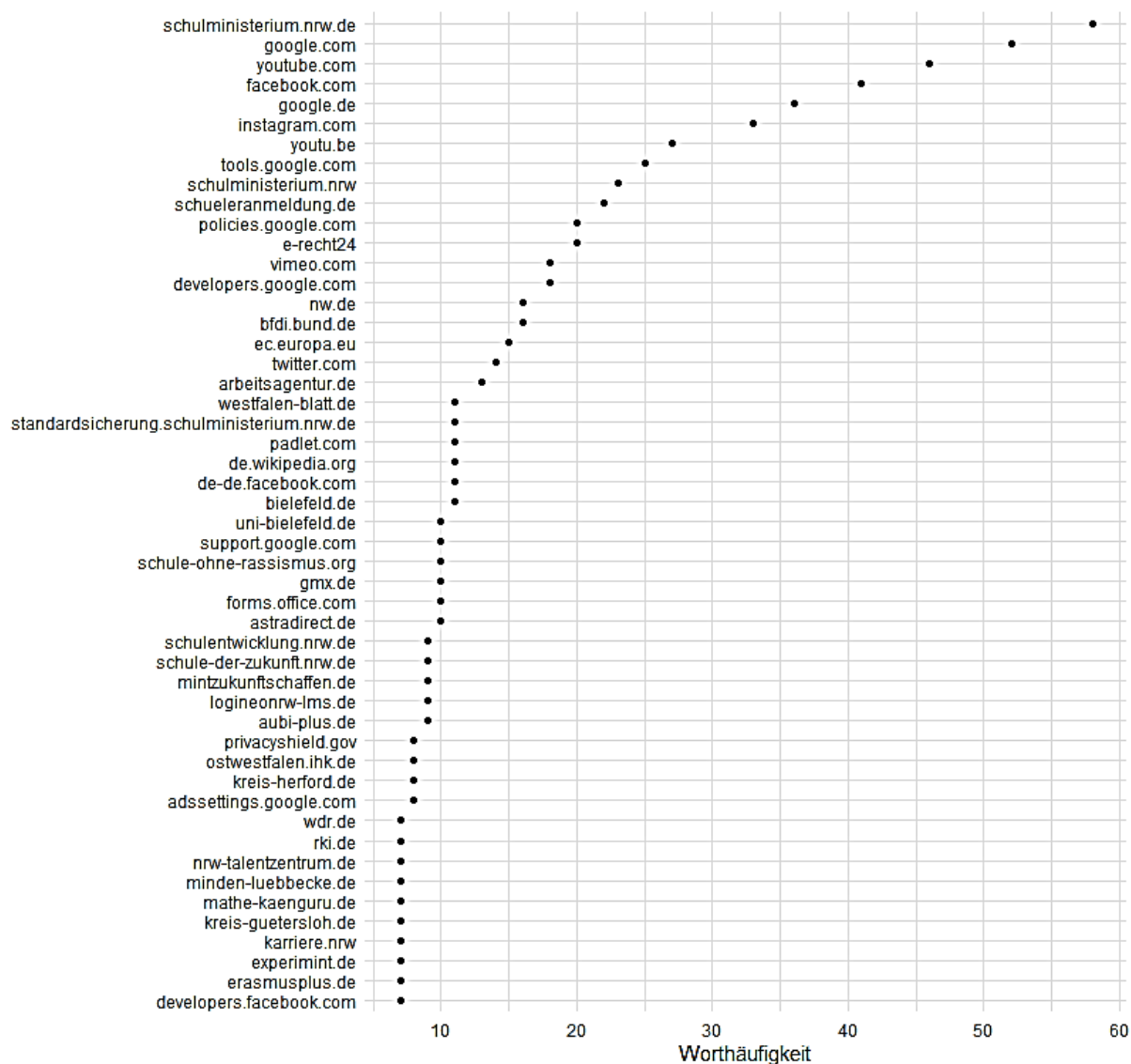


ifh Göttingen

Quelle: eigene Berechnungen

Ebenso können die häufigsten Ziele von Verlinkungen analysiert werden, was in Abb. 9 grafisch abgebildet wird. Dabei zeigt sich der interessante Effekt, dass die häufigste Verlinkung die des nordrhein-westfälischen Schulministeriums ist, anschließend aber bereits die großen Plattformanbieter Google, YouTube, Facebook und Instagram erscheinen. Weitere inhaltlich spezifischere Verlinkungen erscheinen mit deutlich geringerer Anzahl.

Abb. 9: Häufigkeit der Verlinkungen nach Zielen



ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 4: Neben Verlinkungen zum nordrhein-westfälischen Schulministerium dominieren auf den Schulwebseiten Verweise zu Google, YouTube, Facebook und Instagram.

3.2.3 Analyse von Schlüsselworten

Eine inhaltlich tiefere Analyse wird durch die Suche nach Schlüsselworten in den Webseiten ermöglicht. Hierbei wurden beispielhaft zwei Untergruppen von Begriffen gebildet, die Aspekte der Schulentwicklung sowie der Digitalisierung abbilden. Im Folgenden werden die Webseiten der Institutionen sowohl nach der Gesamtanzahl der jeweils auftretenden Begriffe analysiert, als auch nach dem Auftreten einzelner zentraler Begriffe. Die so genutzten Begriffe sind:⁷

Schulentwicklung

"leitbild", "schulprogramm", "digitalpakt", "digitalkonzept", "digitalisierung",
 "medienkonzept", "kooperation*", "netzwerk", "förderverein", "universität", "hochschule",
 "forschungsinstitut", "*bildung", "wettbewerb"

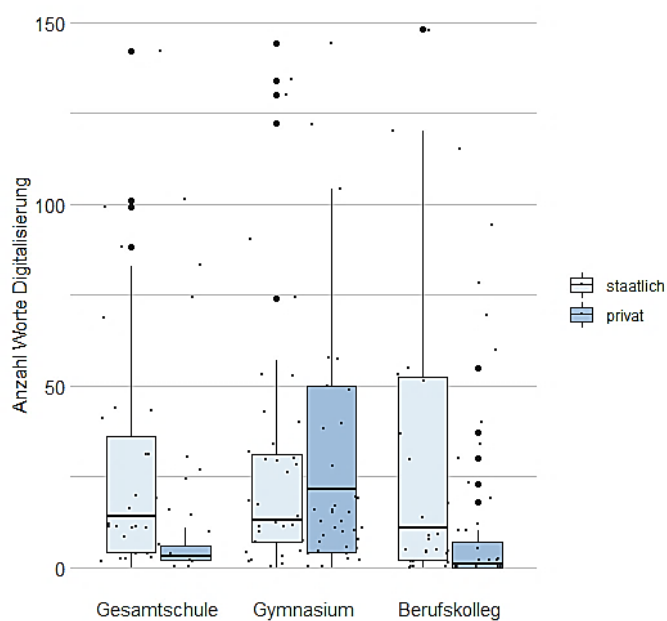
Digitalisierung

"digital*", "*plattform", "computer*", "IT", "tablet*", "ipad", "office", "cloud", "iserv", "download",
 "virtuell", "facebook", "soziale netzwerke", "digitale Medien", "wlan", "beamer"

Als grundlegende Analyse kann die absolute Wortanzahl in den entsprechenden Begriffszusammenstellungen dargestellt werden (Abb. 10 und Abb. 11). Hierbei zeigt sich, dass die gewählten Begriffe in den Schultypen deutlich unterschiedlich auftauchen. Digitalisierungsbegriffe finden sich am stärksten auf Webseiten privater Gymnasien, in ähnlichem Maße bei allen staatlichen Schulformen, hingegen selten in privaten Gesamtschulen und Berufskollegs. Ein anderes Bild zeigt sich für die Häufigkeit von Begriffen, die mit der Schulentwicklung zusammenhängen. Hierbei sind alle staatlichen Schultypen auf einem Niveau, das ebenso von den privaten Gymnasien erreicht wird. Lediglich Webseiten privater Gesamtschulen und Berufskollegs weisen die entsprechenden Begriffe deutlich seltener auf.

⁷ Ein „*“ vor oder nach einem Begriff dient dabei als Platzhalter für einen beliebigen Wortanfang oder eine Wortendung. So erfasst „*bildung“ zum Beispiel sowohl „Fortbildung“ oder „Ausbildung“ als auch „Weiterbildung“.

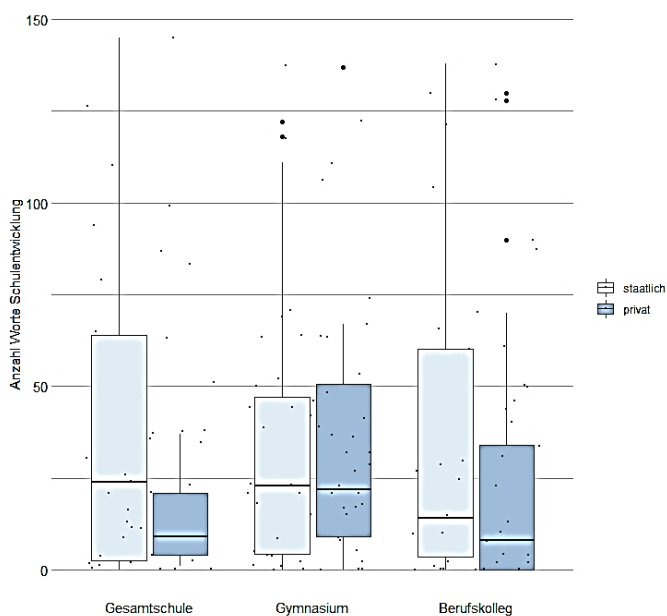
Abb. 10: Absolute Häufigkeit der Digitalisierungsbegriffe nach Schultypen, aggregiert



ifh Göttingen

Quelle: eigene Berechnungen

Abb. 11: Absolute Häufigkeit der Schulentwicklungsbegriffe nach Schultypen, aggregiert



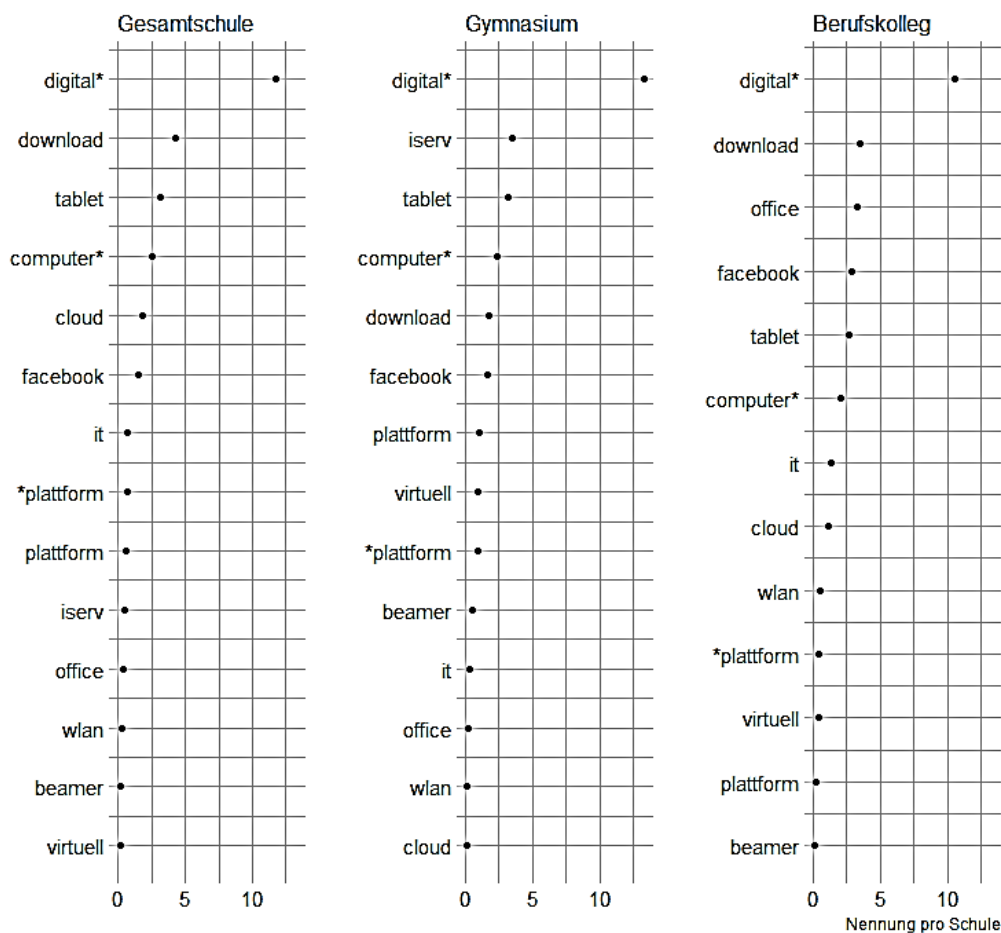
ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 5: Eine differenzierte Suche nach Digitalisierungs- und Schulentwicklungsbegriffen zeigt deutliche Unterschiede, aber auch Gemeinsamkeiten zwischen den Schultypen auf, die insbesondere Rückschlüsse auf den Stand der Digitalisierung zulassen.

Als weitere grundlegende Analyse kann ein Vergleich zwischen den drei Schulformen gewählt werden, um die Häufigkeit und Rangfolge der Begriffe zu zeigen (Abb. 12 und Abb. 13). Hierbei bildet sich zum einen die absolute Häufigkeit der Begriffe ab, wie sie aggregiert schon Abb. 10 und Abb. 11 zu entnehmen ist. Dabei wird die unterschiedliche Verteilung der einzelnen Begriffe deutlich, dass bspw. der Begriff „download“ häufig verwendet wird, während „wlan“ (incl. unterschiedlicher Schreibweisen) praktisch nie vorkommt. Im Vergleich zwischen den drei Schulformen zeigen sich Unterschiede im Hinblick auf die Häufigkeit der Begriffe, allerdings keine grundlegenden Unterschiede. Die ersichtlichen und inhaltlich nachvollziehbaren Effekte sind zum Beispiel die häufige Nutzung von „office“ – im Sinne einer Ausbildung in Microsoft-Office-Produkten in Berufskollegs – und dessen praktisches Nicht-Vorkommen bei Gymnasien oder Gesamtschulen. Insofern sind die Unterschiede gering, jedoch inhaltlich plausibel. Bei unterschiedlicheren Schultypen oder Regionen würde sich das Ergebnis vermutlich deutlich differenzierter darstellen.

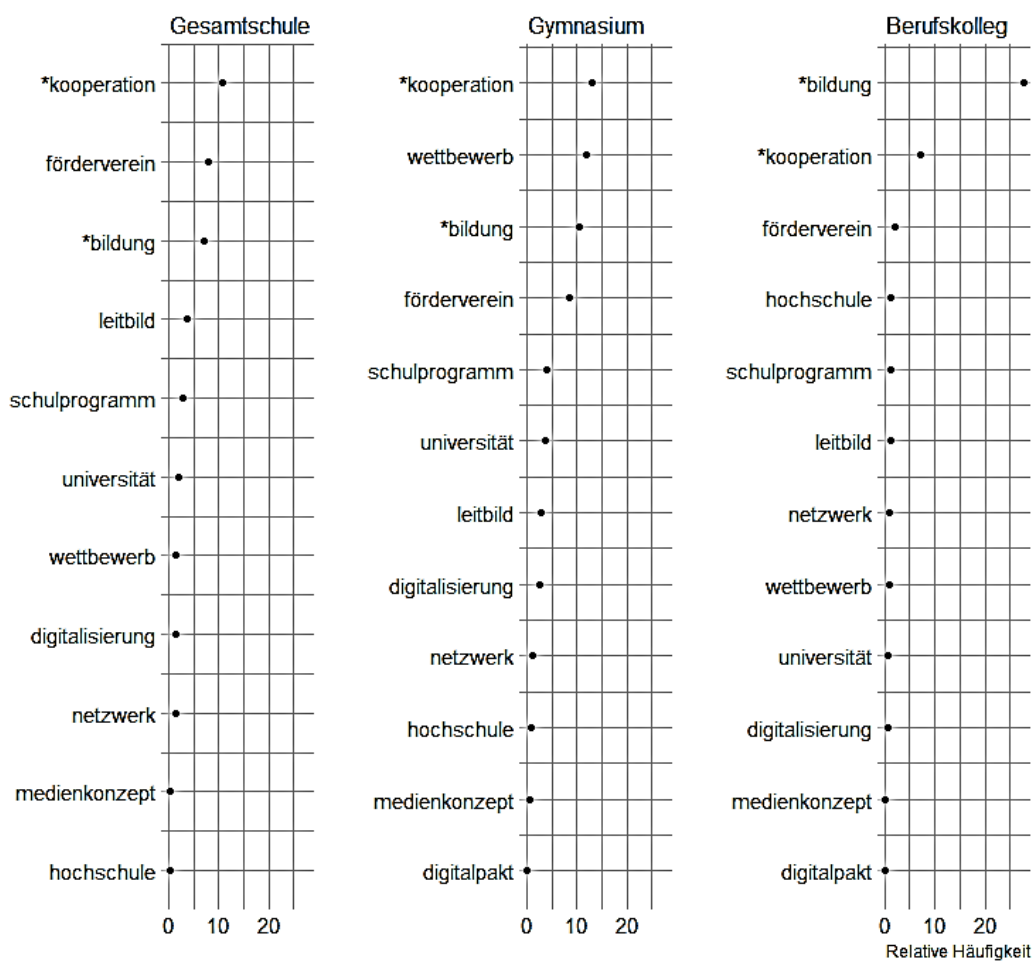
Abb. 12: Häufigkeit der einzelnen Digitalisierungsbegriffe nach Schultypen



ifh Göttingen

Quelle: eigene Berechnungen

Abb. 13: Häufigkeit der einzelnen Schulentwicklungsbegriffe nach Schultypen

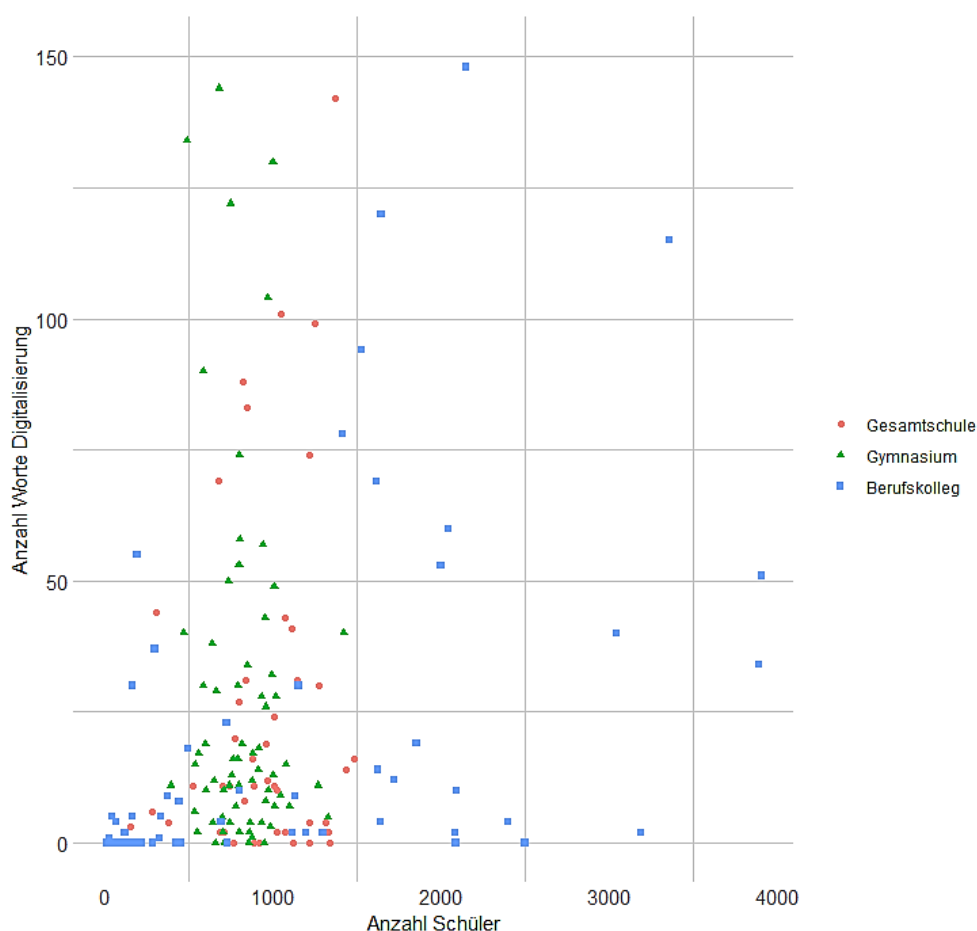


ifh Göttingen

Quelle: eigene Berechnungen

Eine weitere Analyseebene ist die gemeinsame Betrachtung des Auftretens von Begriffen und anderer mit den Schulen verknüpfter Variablen. So wurde beispielhaft die Anzahl der Schüler mit den Schultypen und den Digitalisierungsbegriffen verknüpft (Abb. 14). Hierbei können grundlegende Zusammenhänge grafisch abgebildet werden. Es zeigt sich für die vorliegende Stichprobe kein Größeneffekt, d.h. es gibt keine Hinweise darauf, dass größere Schulen häufiger Digitalisierungsbegriffe auf ihren Webseiten aufweisen. Dies lässt die Interpretation zu, dass die Digitalisierung bzw. deren öffentlichkeitswirksame Dokumentation nicht allein von den Kapazitäten eines großen Kollegiums, sondern in hohem Maße auch von einzelnen handelnden Personen innerhalb der jeweiligen Institution abhängt.

Abb. 14: Häufigkeit der Digitalisierungsbegriffe in Bezug auf Schülerzahl und Schultypen



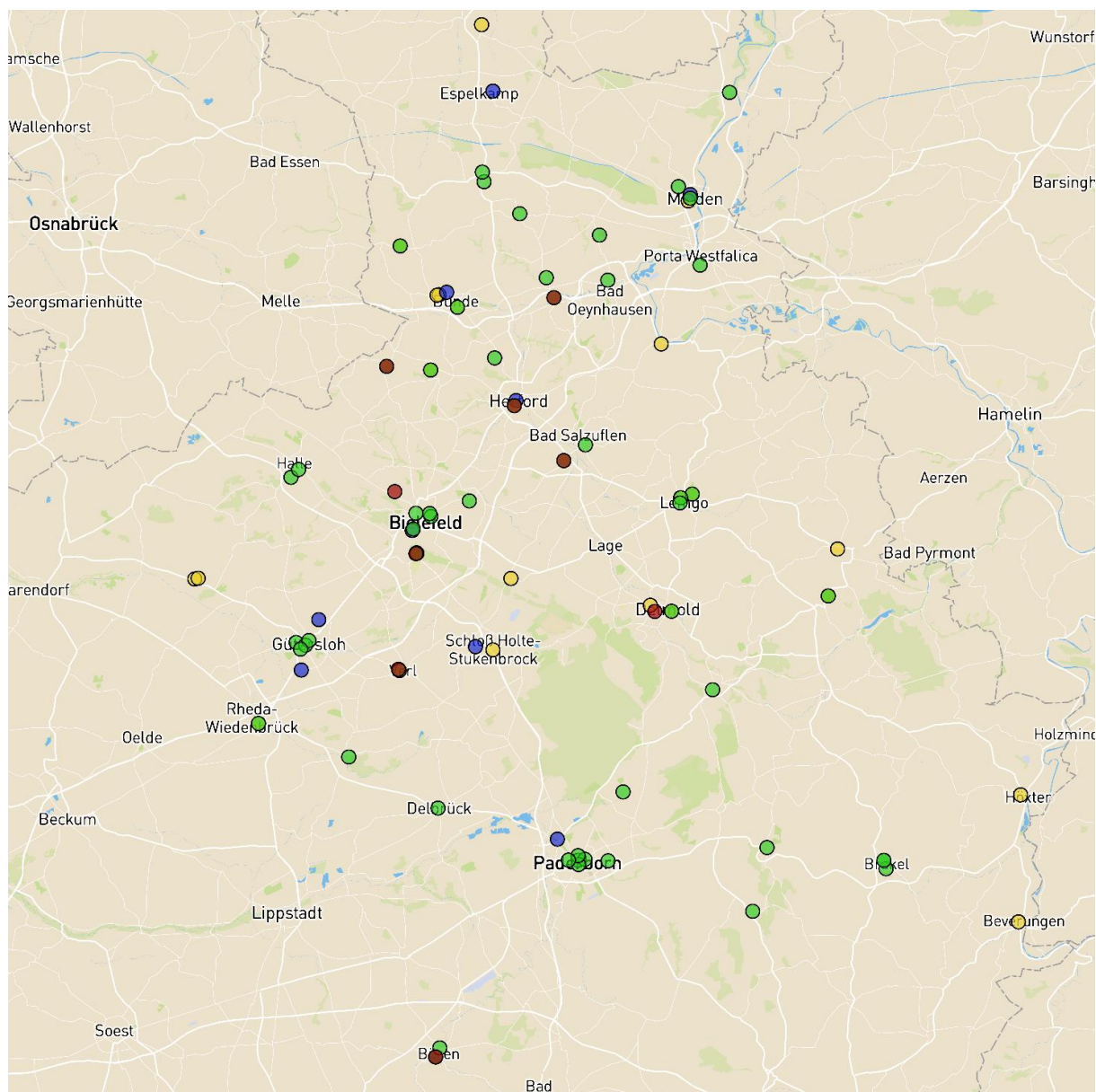
ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 6: Die Stärke der Dokumentation von Digitalisierungsbegriffen hängt nicht von der Schulgröße ab. Dies lässt die Interpretation zu, dass die Digitalisierungsinitiative keine strukturellen Gründe hat, sondern vor allem in der Initiative der jeweiligen Institution begründet ist.

Um regionale Unterschiede und Zusammenhänge deutlich zu machen, eignet sich die kartografische Darstellung zum Auftreten einzelner Begriffe. Abb. 15 zeigt dies beispielhaft für die Begriffe „Cloud“, „Tablet“, „IServ“ und „Lernmanagement“. Hierbei kann direkt die unterschiedliche regionale Häufigkeit der Begriffe gezeigt werden, ebenso wie etwa die stärkere Konzentration der hier ausgewählten Technologien auf größere Städte.

Abb. 15: Häufigkeit spezifischer Digitalisierungsbegriffe nach Regionen



ifh Göttingen

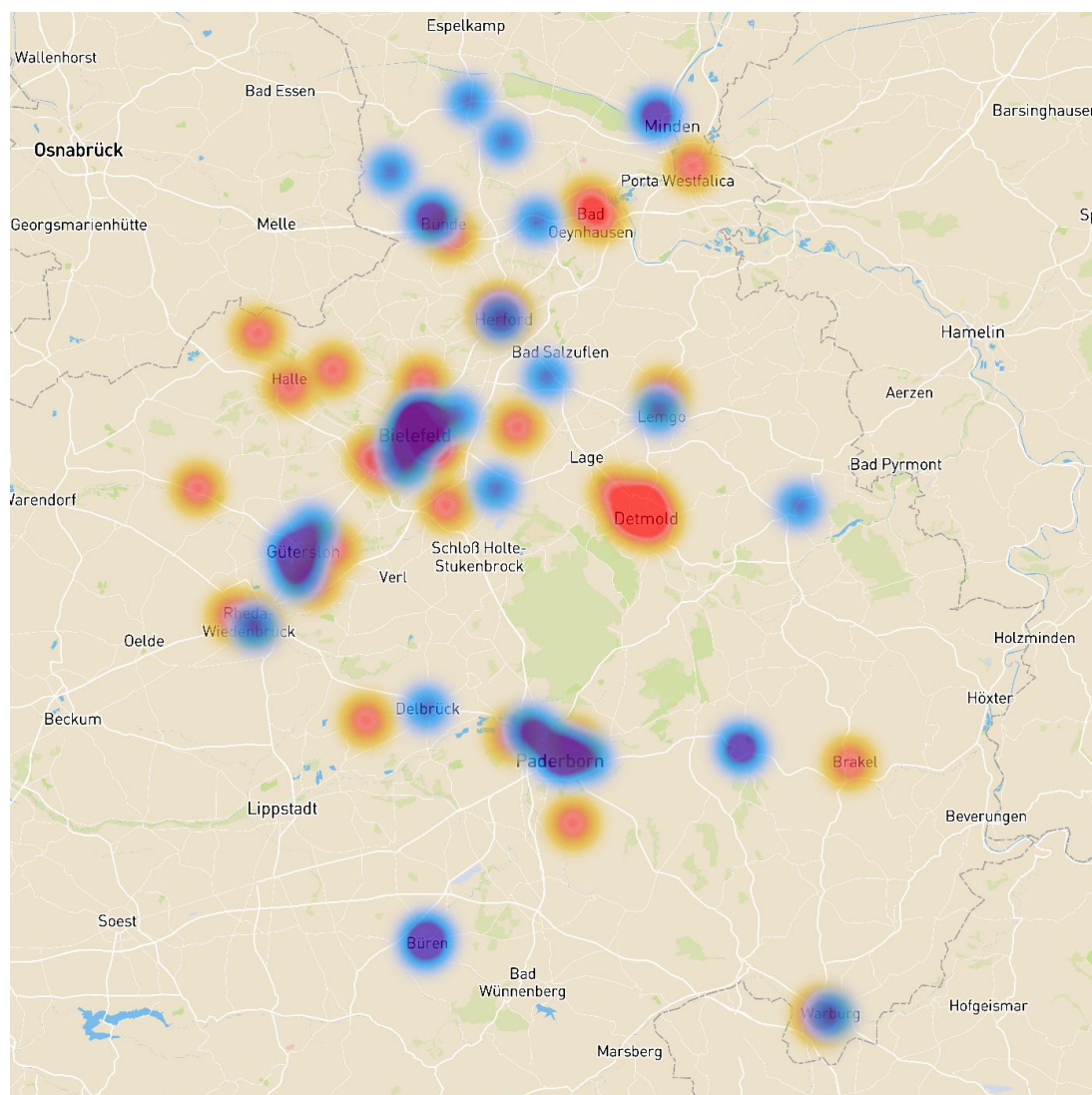
Quelle: eigene Berechnungen

Anmerkung: Cloud: blau, Tablet: grün, IServ: gelb, Lernmanagement: rot

3.2.4 Digitalisierungsindikatorik

Eine Erweiterung der Wortsuche ist eine Heatmap, die alle Digitalisierungsbegriffe in einen Indikatorwert zusammenfasst und so eine Rangbildung über alle Schulen erlaubt. Abb. 16 zeigt beispielhaft ein Ranking im Regierungsbezirk Detmold für die 40 Schulen mit den häufigsten Nennungen und die 40 Schulen mit den wenigsten Nennungen. Hierbei lässt sich deutlicher zeigen, wo regionale Schwerpunkte der Digitalisierungsaktivitäten und wo Nachzügler liegen. Vorreiterschulen zeigen sich in der betrachteten Region in Paderborn, Gütersloh und Bielefeld; gleichzeitig aber auch in ländlichen Regionen. Insbesondere für bundesland- oder kreisübergreifende Analysen kann diese Ansicht regionale Variationen leicht anschaulich machen.

Abb. 16: Häufigkeit der einzelnen Digitalisierungsbegriffe nach Indikatorwert



ifh Göttingen

Quelle: eigene Berechnungen

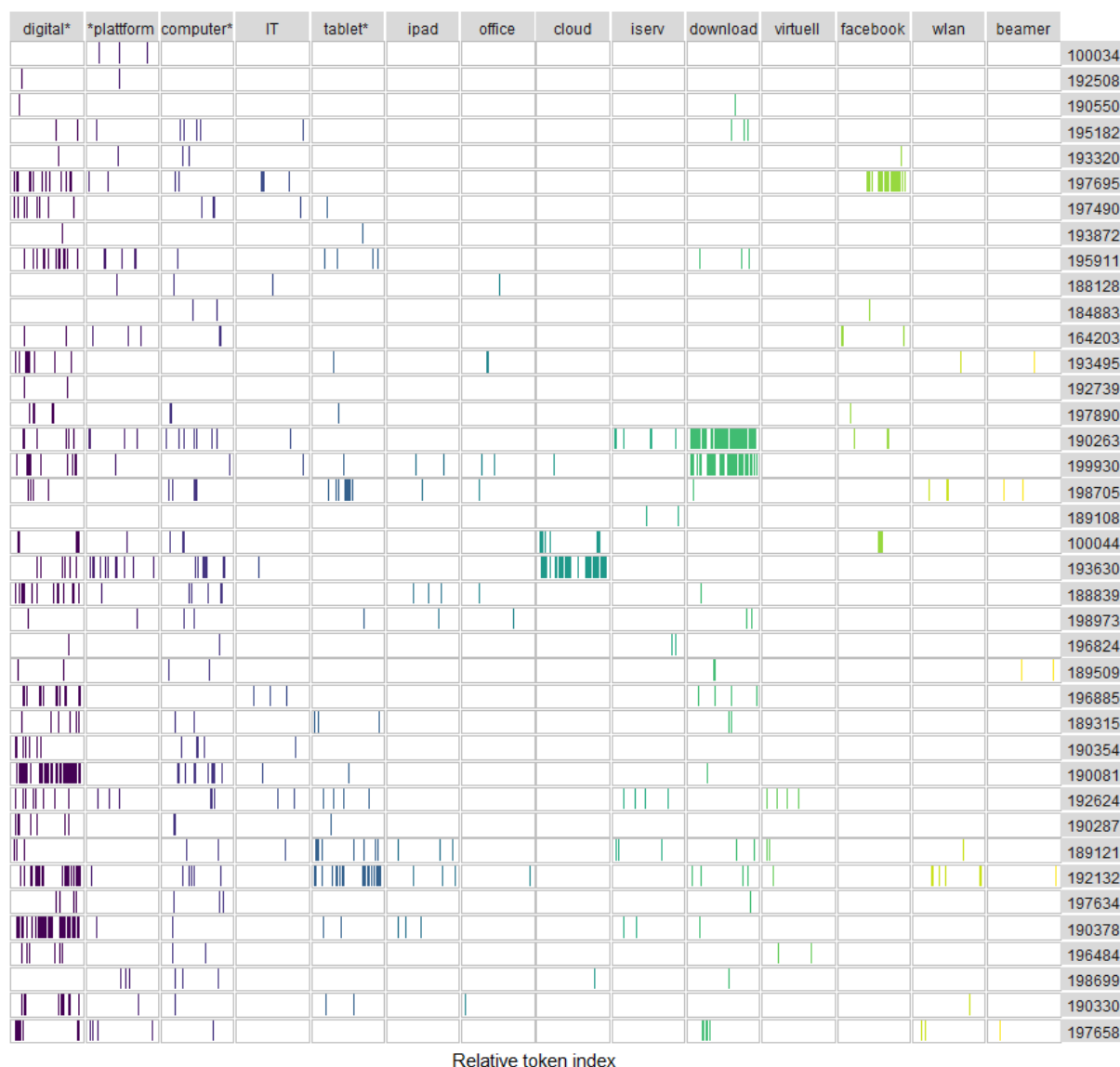
Anmerkung: Top 40: blau, Niedrige 40: Gelb

Ergebnis 7: Kartografische Digitalisierungsindikatoren geben einen effizienten Überblick über regionale Disparitäten, etwa bei der Nutzung spezifischer Technologien.

3.2.6 X-Ray-Analyse von Webseiten

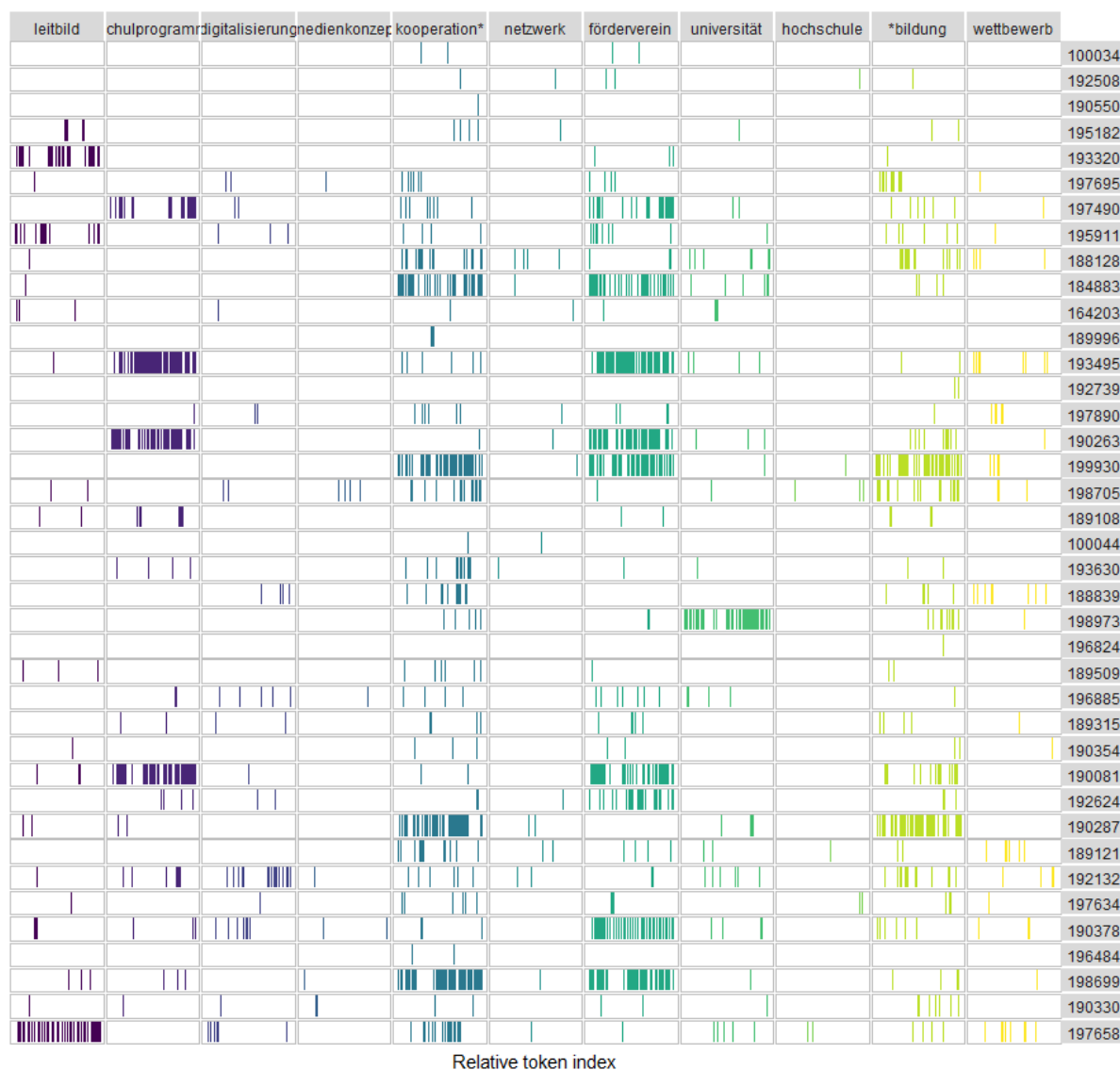
Eine weitere Betrachtung des Auftretens einzelner Worte auf den Schulwebseiten ist die sog. X-Ray-Analyse (Abb. 18 und Abb. 19). Hierbei werden in einer Zeile Schulen dargestellt und in einer Spalte technologische bzw. Schulentwicklungsbegriffe. Eine farbliche Markierung innerhalb der sich ergebenden Zellen zeigt dabei an, dass das betreffende Wort auf der jeweiligen Webseite auftritt. Die Position innerhalb der Zelle zeigt an, ob das betreffende Wort weiter vorne in der Webseitenstruktur oder weiter hinten auftritt. Die daraus resultierende Grafik ermöglicht folglich einen einfachen und schnellen Überblick zur Relevanz zentraler Begriffe für einzelne Schulen. Für Abb. 18 zeigt sich etwa, dass zwei Schulen sehr häufig den Begriff „Download“ verwenden, eine Schule besonders intensiv „Facebook“ nutzt sowie eine Schule häufig „cloud“ verwendet. Ebenso sind zwei Schulen auszumachen, die sehr intensiv „digital*“ (digital mit zusammengesetzten Begriffen) verwenden.

Abb. 18: X-Ray-Analyse der Digitalisierungsbegriffe



Ähnliche Aussagen sind zu den Begrifflichkeiten im Bereich der Schulentwicklung möglich: hier ist zum Beispiel deutlich abzulesen, welche Schulen ein Leitbild bzw. ein Schulprogramm haben, wer einen Förderverein hat, wer Kontakte zu Hochschulen aufweist und wer an Wettbewerben teilnimmt.

Abb. 19: X-Ray-Analyse der Schulentwicklungsbegriffe



Relative token index

ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 9: Die X-Ray-Analyse gibt einen tabellarischen Überblick über das Auftreten und die Positionierung zentraler Begriffe, die hinsichtlich der auf den schulischen Webseiten verwendeten Begriffe einen effizienten Überblick erzeugt.

4. Studie III: Zufallsstichprobe aller Schultypen in Niedersachsen

4.1 Methodik und Datengrundlage

Ausgangspunkt dieser Studie ist eine Zufallsstichprobe von 201 niedersächsischen Schulen, die auf dem Verteilungsschlüssel für die Mittel des Digitalpakts basiert.⁸ Hierbei wurden jeweils die ersten 100 bzw. 101 privaten und öffentlichen Schulen gezogen, die wieder alphabetisch nach Schulträgern aufgeführt waren. Diese Datengrundlage wurde durch eine manuelle Recherche der Webseiten der Schulen ergänzt. Die Stichprobe enthält Informationen zu den folgenden Variablen: *Namen der Schule*, *Anzahl Schülerinnen und Schüler (SuS)*, *Postleitzahl*, *Webadresse*, *Trägerkürzel*, *öffentlicher oder privater Träger* und *Schulform*.

Die Webseiten wurden im November 2020 mittels des ARGUS-Programms⁹ gescraped und mit den Statistikprogrammen *R* und *Stata* ausgewertet. Das ARGUS-Programm 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 (*Zeitstempel*), ob eine Fehlermeldung im Zugriff auf die Seite entsteht (*Fehler*) oder eine Weiterleitung zu einer anderen Website vorgenommen wurde (*Weiterleitung*). 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 Schulen der Stichprobe bezieht; *extern* sind somit alle weiteren Webseiten. Um die Datenmenge des Scrapings zu begrenzen, wurde die Untersuchung auf maximal 50 Unterseiten beschränkt. 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.

Auf Basis dieser Vorgehensweise konnten 190 Webseiten erfolgreich gescraped werden, wobei 8 Webseiten lediglich einen Verweis, z.B. „Seite im Umbau“, enthielten und somit von der weiteren Analyse ausgeschlossen wurden. Die Auswertung erfolgt in den verschiedenen Abschnitten auch nach Schulform, sodass nur die Schulformen mit hinreichenden Beobachtungszahlen in unserer Stichprobe näher untersucht wurden. Insgesamt sechs Schulen der Kategorien Kollegs, Haupt- und Realschulen werden daher nicht näher betrachtet. Zudem wurden Integrierte Gesamtschulen und Oberschulen zu einer Kategorie zusammengefasst. Die Schulen der Stichprobe sind nicht auf einen Träger (der höchste Anteil an der Stichprobe, der auf einen Träger entfällt sind 10 % beim Träger #27, der katholischen Kirche, ansonsten entfallen nie mehr als 5 % der Stichprobe auf einen einzelnen Träger) oder ein Gebiet (Höchstwert von 7 Schulen in Postleitzahl 26160) konzentriert. Insgesamt basiert unsere Analyse damit auf 176 Beobachtungen (siehe Tabelle 2).

⁸ <http://www.schulliste.eu/> [Abrufdatum: 01.12.2020].

⁹ Vgl. Kinne, J. & Axenbeck, J. (2018).

Tabelle 2: Übersicht Datengrundlage Stichprobe Niedersachsen

Schulform	Anzahl	%-Anteil öffentlich	Durchschnitt Anzahl SuS
Grundschule	68	87	160
Gymnasium	19	26	770
Gesamtschule/ Oberschule	29	48	404
Förderschule	22	23	129
Fachschule	38	3	198
Gesamt	176	48	270

ifh Göttingen

Quelle: eigene Berechnungen

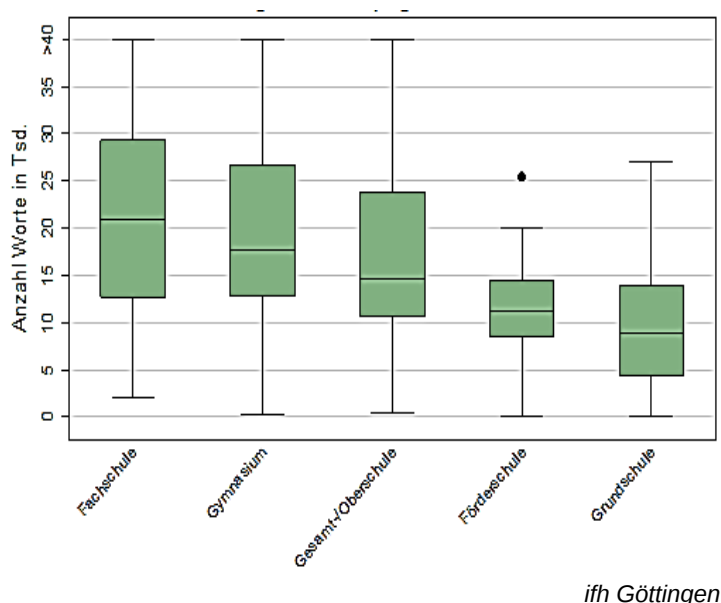
Der Aufbau der Stichprobe ist zunächst stark von den Grundschulen geprägt, die sich überwiegend in öffentlicher Trägerschaft befinden. Die restlichen Schultypen sind mit relativ ähnlichen Anteilen in der Stichprobe vertreten. Dabei fällt auf, dass Gymnasien mit Abstand die höchste Schülerzahl aufweisen, während die Förderschulen die niedrigste durchschnittliche Schülerzahl haben. Gymnasien, Fachschulen und Förderschulen sind überwiegend in privater Trägerschaft, während die Gesamt- und Oberschulen ungefähr zu gleichen Teilen öffentlich und privat getragen werden. Insgesamt zeigen sich in der vorliegenden Stichprobe charakteristisch die Strukturmerkmale der niedersächsischen Schultypen, deren Zusammenhang mit dem Digitalisierungsgrad im Folgenden überprüft wird. Eine deutliche Verzerrung der Zufallsstichprobe stellt hingegen das zu geringe relative Gewicht der öffentlichen Gymnasien dar, was durch die hohe Anzahl privater kirchlicher Gymnasien in der Stichprobe bedingt ist (Schulträger Bistum Hildesheim und Bistum Osnabrück). Dies führt zu einer Verzerrung der Proportionen im Hinblick auf das Gewicht privater Gymnasien, ist jedoch dem geringen Umfang der Vorstudie geschuldet. Bei größerer Stichprobe oder einer Vollerhebung für bestimmte Räume werden solche Verzerrungen vermieden. Insgesamt ist zu betonen, dass der vorliegende Datensatz es ermöglicht, das Potenzial der Webscraping-Analysen von Schulwebseiten im Sinne einer Vorstudie zu zeigen. Aufgrund der sehr begrenzten Anzahl an betrachteten Schulen können allerdings keine repräsentativen Aussagen für Schulformen, Träger oder für das Bundesland Niedersachsen getätigt werden. Die gezeigten durchschnittlichen Effekte können somit mit einer steigenden Stichprobengröße gänzlich anders ausfallen.

4.2 Ergebnisse

4.2.1 Komplexität und Heterogenität der Schulwebseiten

Abb. 20 zeigt den Umfang der Schulwebseiten nach Wortanzahl, dargestellt nach Schulformen. Dies ermöglicht eine erste grobe Einordnung der Relevanz der Webseiten für die Schulformen und den damit zusammenhängenden Aufwand bei der Erstellung und ggf. aktuellen Unterhaltung der Schulseiten.

Abb. 20: Umfang der Webseiten nach Schulform



Quelle: eigene Berechnungen

Grundlegend zeigt die Analyse¹⁰, dass die Unterschiede nach Schulformen sehr groß sind: So haben Grund- und Förderschulen im Durchschnitt weniger als 50 % des Webseiteninhalts verglichen mit Gymnasien oder Fachschulen. Gesamt- und Oberschulen liegen bei nur etwa 83 % des Inhalts gemessen am Median, was aber insgesamt keinen statistisch signifikanten Unterschied darstellt. Die Heterogenität im Umfang der Webseiten ist bei Fachschulen, Gymnasien und Gesamtschulen sehr viel größer als bei Förder- und Grundschulen. So liegt die Spannweite zwischen dem ersten und dritten Quartil der Verteilung (Höhe der Box) für Fachschulen bei etwa 17.000 Worten und für Förderschulen bei nur etwa 6.000.

¹⁰ Hier dargestellt sind Box-Plots, die sich besonders für eine kompakte Darstellung der Verteilung einer Merkmalsausprägung eignen. Der untere Rahmen der Box beginnt bei dem ersten Quartil der Verteilung, d.h. 25 % der Werte sind kleiner gleich diesem Wert. Für Fachschulen liegen also 25 % der Schulen unter einem Wert von 12.731 Wörtern auf ihrer Webseite. Die mittlere Linie in der Box entspricht dem Median, der die oberen und die unteren 50 % einer Verteilung voneinander trennt. Für Fachschulen liegen also 50 % der Schulen über 20.932 Wörtern auf ihrer Webseite und 50 % darunter. Der obere Rahmen der Box entspricht dem dritten Quartil, also 75 % der Fachschulen liegen unter 29.335 Wörtern auf ihrer Webseite. Die "Antennen" außerhalb der Box ("Whiskers") zeigen die Streuung der Verteilung über die mittleren 50 % hinaus. Sind einzelne Werte noch als Punkt dargestellt, sind diese als Ausreißer identifiziert, also Werte, die weit außerhalb der vorliegenden Verteilung liegen - gemessen durch einen Abstand größer als der 1,5-fachen Spannweite der Box, zum Beispiel bei „Förderschulen“ eine Schule mit 25.427 Wörtern auf ihrer Webseite.

Ergebnis 1: Die Webseiten von Fachschulen und Gymnasien sind deutlich komplexer und heterogener als von Förder- oder Grundschulen. Gesamt- und Oberschulen liegen dazwischen.

Diese Unterschiede sind teilweise auch durch die Trägerschaft der Schulen zu erklären. Insgesamt weisen Schulen in privater Trägerschaft mit etwa 19.400 Worten fast den doppelten Umfang gegenüber Schulen in öffentlicher Trägerschaft mit etwa 11.700 Worten auf. Nach Schulform zeigt sich z.B. für Grundschulen in privater Trägerschaft, dass sie durchschnittlich sehr viel mehr Inhalt auf den Webseiten haben (etwa 16.500 Worte) als Grundschulen in öffentlicher Trägerschaft (etwa 9.000 Worte). Eine spezifische Auflösung der Kausalitäten des Umfangs der Webseiten mit der Trägerschaft wäre mit einer größeren Stichprobe problemlos möglich; der Umfang der Stichprobe dieser Vorstudie ermöglicht diese Überprüfung nicht.

Ergebnis 2: Schulen in privater Trägerschaft haben tendenziell komplexere Webseiten.

4.2.2 Explorative Analyse: Sprache und Themensetzung

Neben der Beantwortung gezielter Fragestellungen kann die Text- und Linkanalyse einen explorativen Charakter annehmen. Hierzu werden zu Grunde liegende Textinformationen entsprechend umgeformt und von wenig aussagekräftigen Begriffen bereinigt (Artikel, Pronomen etc.), um so ein klareres Bild der zentralen Themen, Begrifflichkeiten oder der Sprache insgesamt zu erhalten. Die Methodenvielfalt ist hier enorm, von einfachen Visualisierungen der Worthäufigkeiten bis hin zu ausgefeilten Instrumenten des maschinellen Lernens sind je nach Anwendung und Komplexität der Fragestellungen viele verschiedene Analysen und Darstellungsweisen möglich.

Abb. 21 und 22 zeigen eine grobe Auswertung zur Visualisierung der Worthäufigkeiten der Webseiten-Volltexte von Grund- und Fachschulen.

Abb. 21 Wortwolke für Grundschulen



ifh Göttingen

Quelle: eigene Berechnungen

Abb. 22: Wortwolke für Fachschulen



ifh Göttingen

Quelle: eigene Berechnungen

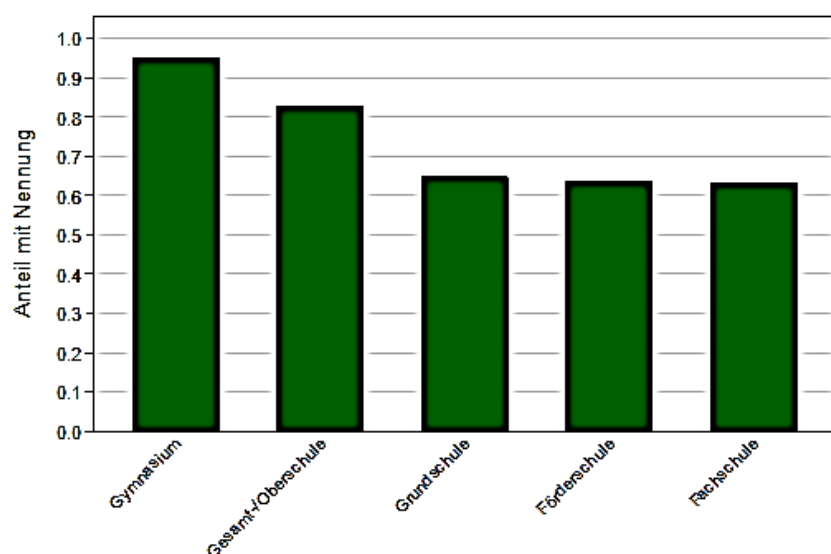
Die Wortwolken zeigen zu erwartende Unterschiede in den Texten von Grund- und Fachschulen. Die Schriftgröße der Worte korreliert linear mit der Häufigkeit des Vorkommens. In Grundschulen dominieren entsprechend Begriffe wie "Kinder", "Eltern", "Schüler" oder "Projektwoche". In Fachschulen häufen sich dagegen Begriffe wie "Ausbildung", "Angebote", "Leistungen" oder "Aktuelles". Mit weiteren Schwerpunktsetzungen und Bereinigungen sind hier auch sehr viel spezifischere Auswertungen und Antworten auf Fragen mit einem präziseren inhaltlichen Zuschnitt möglich.

Ergebnis 3: Eine explorative Textanalyse zeigt die erwarteten Begriffsmuster für Grund- und Fachschulen.

4.2.3 Aktuelle Bezüge: Schulwebseiten und Corona

Die Aktualität ist neben dem Umfang ein wesentliches Merkmal der Kommunikation von Schulen über ihre Onlinepräsenzen. Um eine Approximation der Aktualität der Schulwebseiten zu erzielen, wurde in den Volltexten nach den Begriffen “Corona” und “Covid” gesucht, was eine grobe Einschätzung der Aktualisierungshäufigkeit der Webseiten in 2020 ermöglicht und zeigt, inwieweit die Webseite zur praktischen Kommunikation mit den SuS und Eltern genutzt wird. Die Erwähnung eines der beiden Begriffe wird daher in Abb. 23 nach Schulformen dargestellt. Dabei zeigt sich insgesamt eine starke Thematisierung der Corona-Pandemie auf den Schulwebseiten, wobei über 60 % der Grundschulen, Förderschulen und Fachschulen die Begriffe erwähnen. Deutlich höher mit über 80 % sind die Gesamt- und Oberschulen; Gymnasien erreichen einen Wert von rund 95 %. Entsprechend kann gefolgert werden, dass die Kommunikation mit SuS und Eltern zu aktuellen Themen in Gymnasien sehr stark über die Schulwebseite erfolgt, dies aber mit Abstufungen auch in allen anderen Schulformen erfolgt.

Abb. 23: Erwähnung von Corona nach Schulform

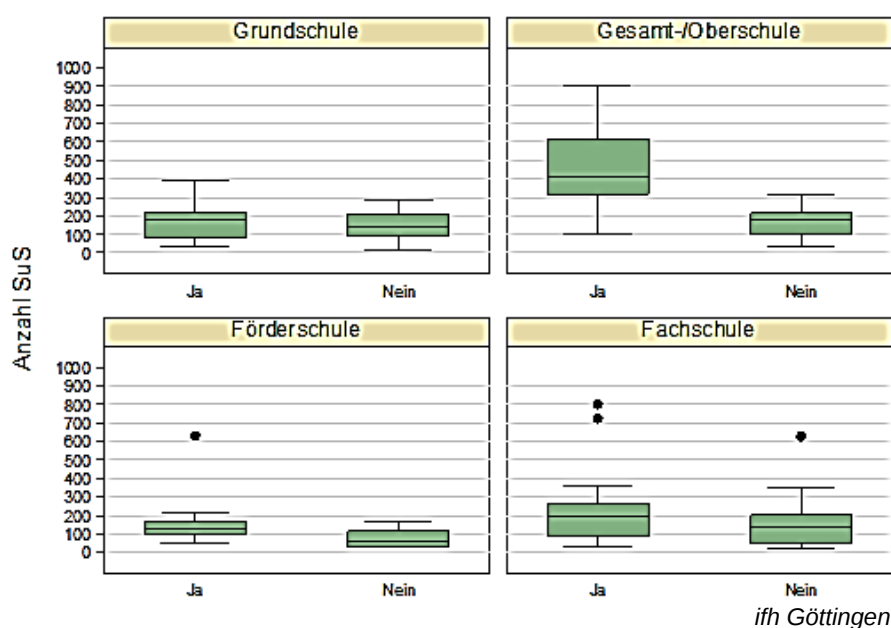


ifh Göttingen

Quelle: eigene Berechnungen

Um zu untersuchen, wie dieser Effekt mit der Größe der Schule zusammenhängt, wird die Erwähnung von Corona in Zusammenhang mit der Schulgröße gebracht (Abb. 24). Hierbei zeigt sich, dass tendenziell größere Schulen eher aktuelle Inhalte über ihre Webseiten kommunizieren. Insgesamt ist der Effekt moderat ausgeprägt und wird vor allem bei den Gesamt- und Oberschulen sehr deutlich. Auf eine Darstellung für Gymnasien wird verzichtet, da mit einer Ausnahme in der Stichprobe alle Schulen Corona erwähnten.

Abb. 24: Erwähnung von Corona nach Schulgröße



ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 4: Die Schulwebseiten werden insgesamt aktuell gehalten, wobei Gymnasien und Gesamt-/Oberschulen sowie allgemein größere Schulen eine höhere Aktualität aufweisen.

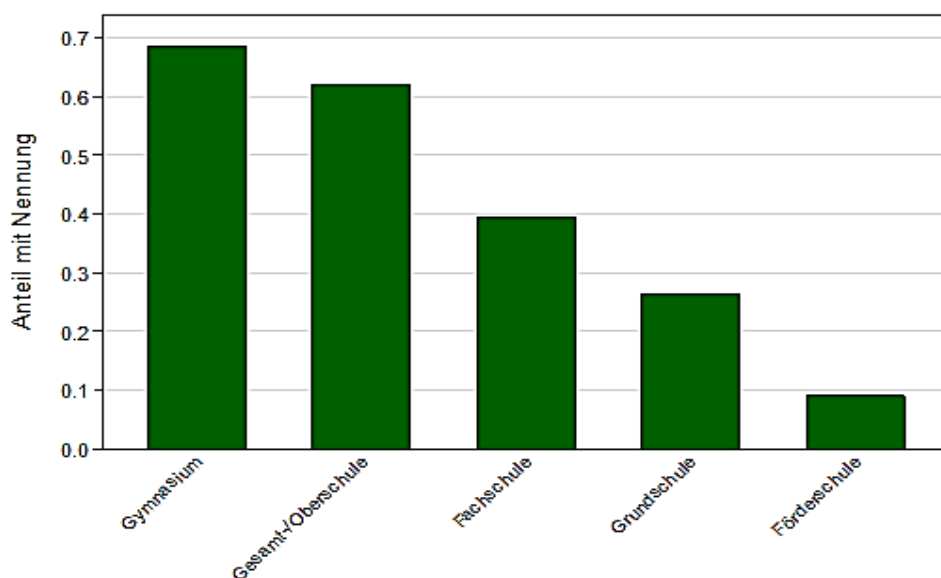
4.2.4 Detailanalyse I: Digitalisierung auf den Schulwebseiten

Wie oben skizziert, bietet Webscraping die Möglichkeit, einzelne Schwerpunkte anhand der Textanalyse über einen großen Datensatz zugleich detailliert zu behandeln. Die einfachste Form ist ein „Wörterbuch“-Ansatz, der im Grunde auf einer Stichwortsuche basiert. Wir untersuchen hierzu beispielhaft zunächst den Themenbereich „Digitalisierung“ mit diesem sehr einfachen Ansatz, d.h. grundlegenden Digitalisierungsbegriffen, die im Schulalltag eine Rolle spielen können. Diese Begriffe sind:

- *digitalisierung, digitalität, digitalpakt, masterplan digitalisierung, lernplattform, tablet, ipad, medienkonzept, digitalkonzept, medienentwicklungsplan* -

Auch für diese Begriffe kann zunächst sehr grob untersucht werden, welche Schultypen einen der genannten Begriffe auf ihren Schulwebseiten haben, um so eine Einordnung der Relevanz von Digitalisierung in der öffentlichen Kommunikation der Schulen bzw. Schulformen zu gewinnen.

Abb. 25: Erwähnung Digitalisierung nach Schulform



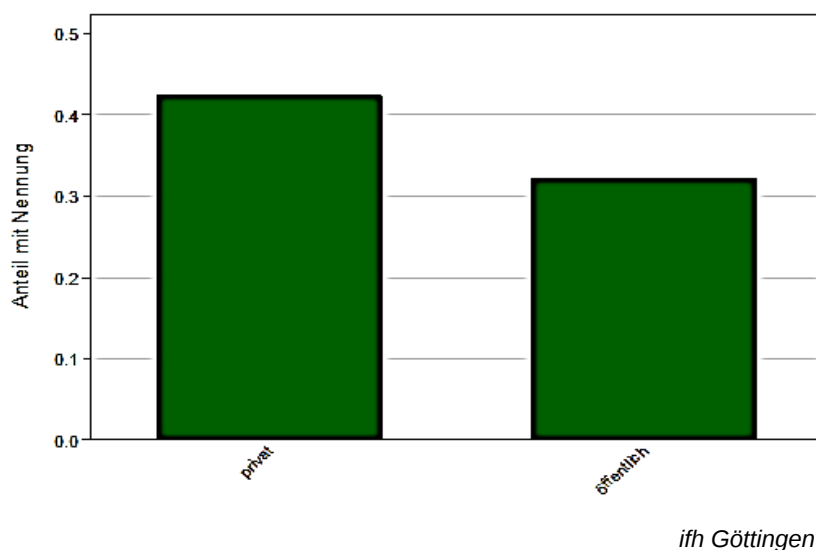
ifh Göttingen

Quelle: eigene Berechnungen

Hierbei zeigen sich erhebliche Unterschiede: Während weniger als 10 % der Förderschulen und rund 25 % der Grundschulen Digitalisierung thematisieren, sind die Anteile bei Fachschulen mit fast 40 %, Gesamt-/Oberschulen mit über 60 % und Gymnasien mit fast 70 % deutlich höher.

Differenziert nach Schulträgern zeigt sich ein leicht erhöhter Wert für private Schulträger von etwa 40 % gegenüber etwa 30 % bei den öffentlichen Schulträgern (Abb. 26).

Abb. 26: Erwähnung Digitalisierung nach Schulträger



Quelle: eigene Berechnungen

Unser quantitatives Vorgehen erlaubt es zudem, Kausalzusammenhänge darzustellen. Hierfür nutzen wir multiple Regressionsanalysen, die Ceteris-Paribus-Effekte aufzeigen, d.h. es werden Korrelationskoeffizienten berechnet unter der Annahme, dass alle anderen Einflussfaktoren unverändert bleiben. Zum Beispiel kann man so den Effekt der Schulform herausarbeiten, ohne die störende Verzerrung der systematisch unterschiedlichen Schülerzahl mit zu erfassen. Tabelle A1 im Anhang zeigt die Ergebnisse einer Probit-Regression, die eine Beschäftigung mit dem Thema Digitalisierung erklärt. Es zeigt sich, dass im Vergleich zu Grundschulen Gymnasien und Gesamt-/Oberschulen statistisch signifikant häufiger das Thema Digitalisierung auf ihren Websites behandeln, Förderschulen dagegen signifikant weniger häufig. Schulen in privater Trägerschaft zeigen keine signifikant höhere Erwähnung von Digitalisierung, was jedoch vermutlich mit der Größe der Stichprobe zusammenhängt und sich mit einer größeren Stichprobengröße relativieren würde. Ein geringer, nicht signifikanter marginaler Effekt wird für die Anzahl der SuS geschätzt, sodass hier kein Zusammenhang vermutet werden muss.

Ergebnis 5: Gymnasien und Gesamt-/Oberschulen thematisieren signifikant häufiger das Thema Digitalisierung, Förderschulen seltener. Ein Zusammenhang zur Trägerschaft und Anzahl der SuS zeigt sich nicht.

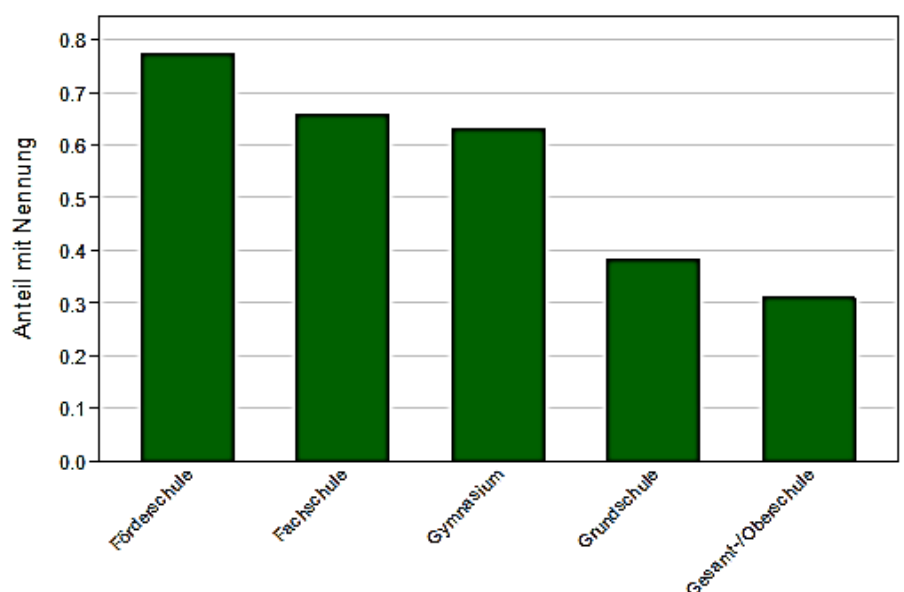
4.2.5 Detailanalyse II: Fortbildung auf den Schulwebseiten

Ein anderes Bild zeigt sich für die Untersuchung von Fortbildungen und ihrem Vorkommen auf den Schulwebseiten. Gesucht werden hierfür die Begriffe:

- *fortbildung, mikrofortbildung, mini-fortbildung* -

Dabei zeigt sich, dass Förder- (75 %) und Fachschulen (66 %) sowie Gymnasien (63 %) relativ häufig Fortbildungen thematisieren, Grundschulen (38 %) und Gesamt-/Oberschulen (31 %) dagegen deutlich seltener.

Abb. 27: Erwähnung Fortbildung nach Schulform



ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 6: Förder- und Fachschulen sowie Gymnasien thematisieren relativ häufig Fortbildungen, Grundschulen sowie Gesamt-/Oberschulen relativ selten.

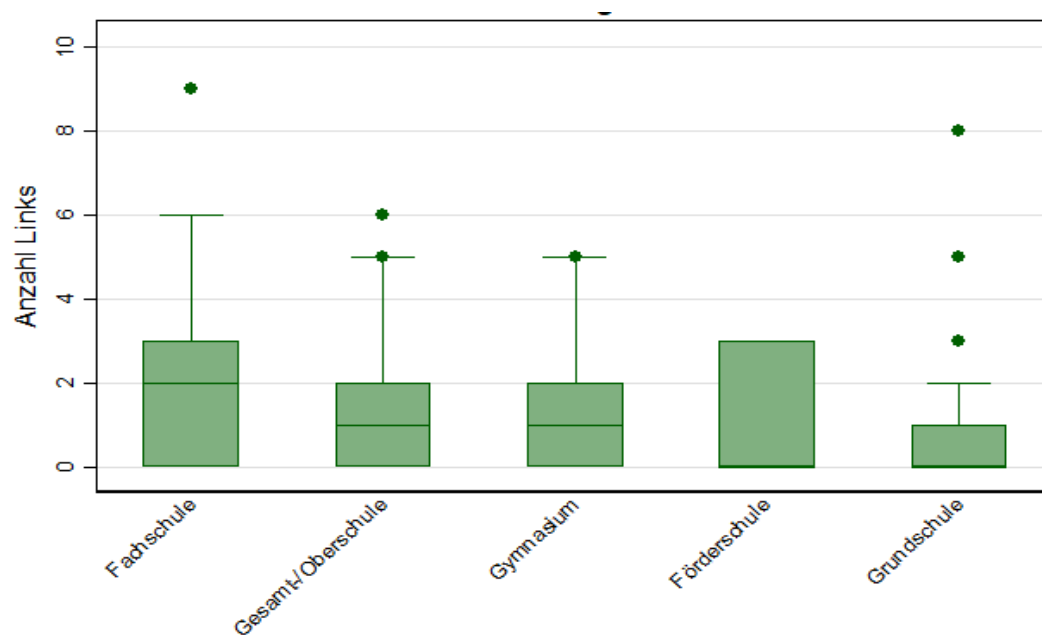
4.2.6 Linkanalyse: Social Media Nutzung

Mit dem Scraping der externen und internen Links aller Webseiten aus der Stichprobe eröffnen sich weitere Auswertungsdimensionen. Angesichts der kleinen Stichprobe zeigt die Betrachtung interner Links - also Links zu anderen Schulen in der Stichprobe - keine relevanten Ergebnisse. Bei größeren Stichproben können dagegen über die internen Verlinkungen Netzwerkstrukturen aufgedeckt und systematisch dargestellt werden. Auch externe Links weisen auf Netzwerke, aber auch zum Beispiel verwendete Instrumente hin. Hier wird das Beispiel der Social Media-Nutzung näher betrachtet. Dabei werden folgende Verlinkungen erfasst:

- facebook, twitter, instagram, youtube, pinterest, linkedin, wikipedia, whatsapp, snapchat, periscope, github, flickr, vimeo -

Grundlegend zeigt sich, dass 45 % bzw. 31 % der Förder- und Grundschulen auf mindestens eine Social Media-Seite verlinken. Gleichzeitig haben mehr als 60 % der Fachschulen, Ober- und Gesamtschulen sowie Gymnasien zumindest eine Verlinkung zu Social Media. Abb. 28 erfasst dabei die Intensität der Nutzung von Social Media nach Schulformen. Die Social Media-Nutzung in Fachschulen ist am stärksten ausgeprägt, wobei im Durchschnitt jede Schule mehr als zwei Links aufweist. Bei Gymnasien und Gesamt-/Oberschulen verweist nur jede zweite Schule auf eine Social Media-Seite, wobei es durchaus Schulen mit deutlich höherer Link-Anzahl gibt.

Abb. 28: Social Media-Nutzung nach Schulform



ifh Göttingen

Quelle: eigene Berechnungen

Ergebnis 7: Grund- und Förderschulen nutzen Social Media weniger häufig als Fachschulen, Gesamt-/Oberschulen und Gymnasien. Hier verweist die Mehrheit der Schulen auf Social Media.

5. Ausblick

Die vorliegende Kurzstudie bietet einen ersten Einblick, wie das Webscraping für den Bildungsbereich eingesetzt werden kann und insbesondere, wie die Digitalisierungsberatung und -förderung von diesem Ansatz profitieren kann. Grundlegend sind vielfältige Anwendungen denkbar, von denen fünf skizziert werden sollen:

1. Übergreifend kann ein deutschlandweites Monitoring von Veränderungen an Bildungseinrichtungen anhand von Aktualisierungen auf den Internetauftritten erfolgen. Hierbei können Nutzungsgewohnheiten, thematische Trends, aktuelle Ereignisse oder ähnliches, was sich auf den Webseiten niederschlägt, abgebildet werden. Ebenso können regionale und schulträgerspezifische Differenzierungen bei den Auswertungen vorgenommen und kartografisch auf verschiedenen Aggregationsebenen dargestellt werden. Dies würde ein dauerhaftes Digital-Monitoring der deutschen Bildungseinrichtungen ermöglichen.

Im Gegensatz zu aufwändigen Befragungen und Wettbewerben bietet Webscraping eine einfache und verhältnismäßig kostengünstige Möglichkeit, ein dynamisches Monitoring aufzubauen. Dieses würde sich nach der ersten Ausgangsanalyse zu einem dynamischen Monitoring entwickeln, das mit jedem weiteren Jahr einen zunehmend fundierteren Längsschnitt mit Blick auf die Entwicklungen in den Einzelschulen und Schulträger-Bereichen erlaubt. Flankierende Befragungen an den Einzelschulen könnten zur Präzisierung der gewonnenen Erkenntnisse beitragen.

2. Darauf aufbauend kann die aus Forschungssicht relevante Frage des Zusammenhangs von soziodemografischen Eigenschaften von Regionen mit den Ergebnissen des Webscrapings abgebildet werden. Hierbei kann etwa untersucht werden, welche systematischen Unterschiede sich zwischen den Webseiten von Bildungseinrichtungen aus Regionen mit unterschiedlich hohem Wohlstandsniveau, zwischen Stadt und Land, zwischen Regionen mit hohem Anteil von SuS mit Migrationshintergrund oder Regionen hohen Wettbewerbs zwischen den Schulformen zeigen. Die Ergebnisse können anschaulich kartografisch aufbereitet werden.

Nicht nur der Beratungs- und Entwicklungsbedarf von Schulträgern und Einzelschulen lässt sich hieraus ableiten, sondern auch die für größere Studien so wichtige Frage der Bildungsgerechtigkeit. Eine Auswertung der hier identifizierten Problembereiche könnte dann von der Bildungsforschung aufgegriffen und mit eigenen Forschungsmethoden intensiv verfolgt werden.

3. Ebenso können über das Webscraping-Verfahren problemlos Vorreiter-Institutionen in Digitalisierungsfragen oder für andere Themen identifiziert werden. Im Falle eines Aufbaus eines deutschlandweiten Datensatzes kann dadurch eine bundesweit gezielte Stichwortsuche zu vielfältigen Themen und die anschließende qualitative Vertiefung der quantitativen Ergebnisse durch Interviews und Schulbesuche bei den Schulen mit relevanten bzw. interessanten Charakteristika erfolgen.

Hier bietet das Webscraping die Möglichkeit, effizient Vorreiter-Schulen (beispielsweise gestaffelt nach Regionen oder Bundesländern) zu identifizieren. Ein Vorhaben, das normalerweise umfassende und kostspielige Ausschreibungen und

Wettbewerbe seitens der einschlägigen Stiftungen und ebenso umfassend angelegte Studien mit hohem Personalaufwand voraussetzt.

4. Webscraping ermöglicht eine Marktanalyse von Schulen, indem die Webseiten der im Wettbewerb um SuS stehenden Schulen nach bestimmten Technologien, Apps, Plattformen oder sozialen Netzwerken analysiert werden. Hiermit kann effizient gezeigt werden, welche Bildungseinrichtungen mit welchen Technologien werben, welche Regionen sich in technologischer Hinsicht dynamisch entwickeln und wo Marktpotenziale für einzelne Technologien bestehen.

Damit bietet der hier entwickelte Untersuchungsansatz das Potenzial, umfassende Marktanalyse und Marktmonitoring - z.B. für Bildungsanbieter im aktuell hochdynamischen Marktgeschehen - im Bildungsbereich zu ermöglichen. Dies könnte sowohl Hardware- als auch Software-Anbietern erstmals einen einheitlichen Überblick über den Markt verschaffen.

5. Ebenso ist eine Schulträgeranalyse und Politikevaluation denkbar: Hierbei können aus wissenschaftlicher Perspektive Bundesländer mit verschiedener Schulpolitik verglichen werden, um zu untersuchen, ob sich die politischen Entscheidungen im Falle ansonsten ähnlicher struktureller Eigenschaften der jeweiligen Bevölkerung in unterschiedlichen Inhalten der Schuldarstellungen niederschlagen. Dies kann insbesondere im Falle differenzierter Digitalisierungsförderungen zu interessanten Evaluationsmöglichkeiten von Schulpolitik führen.

Gerade im föderal organisierten Bildungssystem bietet Webscraping die Gelegenheit, in der Politik- und Schulträgerberatung nicht nur eine differenzierte Bedingungsanalyse vorzunehmen, sondern zugleich auch Handlungsalternativen und Praxisvergleiche schon im Vorfeld des eigentlichen Beratungsgeschehens konzipieren zu können.

6. 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>.
- Eickelmann, B. & Drossel, K. (2020). Digitales Potenzial. Studie der Vodafone Stiftung. <https://www.vodafone-stiftung.de/studie-digitale-optimalschulen> [Abrufdatum: 19.05.2021].
- 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.

7. Anhang

7.1 Fragebogen

Sehr geehrte Damen und Herren,

für eine geplante Studie zu Inhalten von Schulwebsites interessiert uns Ihre Einschätzung bezüglich deren Rolle und Funktion für die Öffentlichkeitsarbeit Ihrer Schule. Mit wenigen Fragen möchten wir uns ein erstes Bild davon machen, wie Sie als Schulleitungen die Bedeutung Ihrer Schulwebsite einschätzen. Schon jetzt vielen Dank für Ihre Teilnahme an dieser vollständig anonymen Umfrage, im Rahmen derer auch keine Rückschlüsse auf die ausfüllende Person oder Schule möglich sind.

1. Welcher Schulform gehört Ihre Schule an?
 - a. Gymnasium
 - b. Integrierte oder kooperative Gesamtschule
 - c. Oberschule
 - d. Grundschule
 - e. Sonstige (Freifeld)

2. Wir nutzen unsere Webseite für...
 - a. ...Hinweise auf aktuelle Veranstaltungen.
 - b. ...die Vorstellung des Kollegiums.
 - c. ...Hinweise auf aktuelle Schwerpunkte der Schul- und Unterrichtsentwicklung.
 - d. ...Informationen für Schülerinnen und Schüler.
 - e. ...den Zugang zum schuleigenen Lernmanagementsystem.
 - f. ...Informationen für Eltern unserer Schülerinnen und Schüler.
 - g. ...Informationen für interessierte Eltern von Grundschulkindern.
 - h. ...die stadtweite (Nachwuchs-)Werbung für unsere Schule.
 - i. ...die Werbung von Personal / interessierten Lehrerinnen und Lehrern.
 - j. ...Informationen die Schulaufsicht (Schulbehörden) betreffend.

3. Wie sehr stimmen Sie auf einer Skala von 1-5 den folgenden Aussagen zu?
Die Website unserer Schule ...
 - a. ...ist wichtig für unsere Öffentlichkeitsarbeit.
 - b. ...informiert über die Aktivitäten der Schulgemeinschaft.
 - c. ...informiert über Leitbild und Schulprogramm.
 - d. ...informiert über Aktivitäten im Bereich Schul- und Unterrichtsentwicklung.
 - e. ...dokumentiert unser Engagement im Bereich Digitalisierung.
 - f. ...wird aktuell gehalten.

4. Offene Antwort: In diesem Feld können Sie uns weitere Rückmeldungen zur Nutzung oder auch zu Problemen in Bezug auf die Schulwebsite geben. Vielen Dank.

7.2 Regression Determinanten der Worthäufigkeiten

Tabelle A 1: Regression Determinanten der Worthäufigkeiten

Iteration 0: log likelihood = -116.43513						
Iteration 1: log likelihood = -102.20665						
Iteration 2: log likelihood = -102.08423						
Iteration 3: log likelihood = -102.08415						
Iteration 3: log likelihood = -102.08415						
Number of obs = 176						
LR ch2(6) = 28.70						
Prob > chi2 = 0.0001						
Log likelihood = -102.08415				Pseudo R2 = 0.1233		
Word_occu_dig	Coef.	Std. Err.	z	P> z	[95 % Conf.	Interval]
Form	1.260963	.5204848	2.42	0.015	.2408319	2.281095
Gymnasium	.9671456	.3330105	2.90	0.004	.3144569	1.619834
Gesamt-/Oberschule	-.8367413	.4399521	-1.90	0.057	-1.699032	.0255488
Förderschule	.2168063	.3444791	0.63	0.529	-.4583604	.891973
privat	.1939229	.2718273	0.71	0.476	-.3388489	.7266947
sus	-.0004272	.0005869	-0.73	0.476	-.0015775	.0007231
_cons	-25906946	.1925907	-3.07	0.002	-.9681656	-.2132237

ifh Göttingen

Quelle: eigene Berechnungen