Offene Daten für die digitale Philosophie: Anforderungen an eine Datensammlung zur Philosophie und ihrer Geschichte

Heßbrüggen-Walter, Stefan
early.modern.thought.online@gmail.com
Universität Trier, Deutschland

Gegenstand meines Beitrags ist die Erörterung von Anforderungen an eine offene Datensammlung zur digitalen Philosophie und Philosophiegeschichte, insbesondere im Blick auf zwei Fragen. Einerseits ist aus der Sicht der digitalen Geisteswissenschaft zu fragen, welchen Kriterien eine solche Datensammlung genügen sollte, um valide Schlussfolgerungen zu erlauben. Dies betrifft sozusagen ihre "formale" oder "methodische" Seite. Andererseits ist aus der Sicht des Faches Philosophie zu fragen, welche Arten von Daten überhaupt für das Fach relevante Einsichten ermöglichen können. Dies betrifft die inhaltliche oder "materiale" Seite. Hier können beide Fragen natürlich nur exemplarisch erörtert werden, ihre Wichtigkeit für die Projektierung einer Datensammlung zur digitalen Philosophie und Philosophiegeschichte sollte aber auf der Hand liegen.[1]

Zum Begriff der Datensammlung

Wir sprechen im folgenden von „Datensammlungen“,[2] weil die digitale Philosophie und Philosophiegeschichte, wie genauer zu zeigen wird, nicht nur Textdaten, sondern auch Metadaten über Text zu ihren Forschungsgegenständen zählt. Zudem kann die Philosophie des 20. Jahrhunderts und ihre Geschichte in reproduzierbarer Weise aus urheberrechtlichen Gründen weit überwiegend nur mittels „abgeleiteter Textformate“ (Schöch u. a. 2020) analysiert werden. Die Heterogenität dieser Datenformate legt es nahe, den generischen Begriff der Datensammlung anstatt spezifischerer Termini wie "digitale Textsammlung", "Corpus", "Kanon" (siehe dazu (Henny-Krahmer und Neuber 2017)) in Anschlag zu bringen.

Untersuchungsgegenstand

Analysieren möchte ich im folgenden zwei jüngere Studien zur digitalen Analyse von Philosophie im 20. bzw. 21. Jahrhundert (Malaterre u. a. 2021; Noichl 2021).[3] Gegenstand sind zum einen die Philosophie insgesamt, deren Struktur anhand von Kozitationsanalysen erhellt werden soll,[4] zum andern eine Teildisziplin der Philosophie, die Wissenschaftstheorie, deren diachrone Entwicklung anhand von topic models von acht einschlägigen Zeitschriften sichtbar gemacht werden soll.[5] Die zugrundeliegenden Datensammlungen enthalten also einmal ausschließlich Metadaten und einmal Volltexte von Zeitschriftenartikeln mit korrespondierenden Metadaten, v. a. Erscheinungsjahr und die publizierende Zeitschrift. Damit ist schon eine erste Anforderung an eine offene Datensammlung zur digitalen Philosophie und Philosophiegeschichte benannt: Sie sollte nicht nur Volltexte, sondern auch Metadaten einschließen, selbst wenn die den Metadaten korrespondierenden Textträger nicht oder noch nicht digitalisiert sein sollten, sondern, wie bei Noichl, nur Angaben zu den erfassten Texten (wie der Aufsatztitel oder Abstract) sowie die Bibliographie zitierter Werke in die Datensammlung aufgenommen werden, da die eigentlichen Aufsätze selbst noch urheberrechtlich geschützt sind.

Zur Erstellung von Datensammlungen in der Philosophie: zwei Beispiele

Erster Schritt der Erstellung einer Datensammlung und demnach auch ihrer Bewertung ist nach Schöch 2017, 224 die Angabe, wie Gegenstand und Umfang eingregrenzt werden. Malaterre u. a. 2021, 2885 geben den Umfang ihrer Datensammlung mit 15897 englischsprachigen Aufsätzen an, die zwischen 1934 und 2017 in acht wissenschaftsphilosophischen Zeitschriften veröffentlicht worden sind. Die Volltexte wurden von JSTOR zur Verfügung gestellt. Dass die Vollständigkeit der Digitalisierung und die Korrektheit zugrundegelegten Metadaten mit Hilfe von weiteren Quellen überprüft wurden, ist nicht ersichtlich. Noichl 2021, 5092 nutzt als Ausgangspunkt der Erstellung der zugrundeliegenden Datensammlung die Fachbibliographie „Philpapers“ (Bourget und Chalmers o. J.). Die dort verzeichneten 1.782.816 Aufsätzen werden in zwei Schritten auf eine Datensammlung von insgesamt 68.152 Aufsätzen reduziert, indem zunächst Zeitschriften ausgeschlossen werden, die nach Meinung des Autors nicht zum fachlichen Kern der Philosophie zu zählen sind, allerdings ohne die Anzahl der damit ausgeschiedenen Aufsätze anzugeben. Für die verbliebenen Zeitschriften werden die in der Zitationsdatenbank „Web of Science“ enthaltenen Texte abgefragt. Aufsätze, die nicht mindestens viermal in anderen in „Web of Science“ enthaltenen Aufsätzen zitiert werden, werden ausgeschlossen. Es verbleiben 87.720 Datensätze. Aus dieser Teilmenge werden alle Datensätze entfernt, die nicht mindestens drei Zitationen enthalten, die auch in anderen Aufsätzen angeführt werden. Damit umfasst die zu analysierende Datensammlung am Ende Metadatensätze zu 68.152 Aufsätzen. Die zeitliche Erstreckung des erfassten Schrifttums wird nicht angegeben.

Festzuhalten ist zunächst, dass in beiden Analysen die verwendeten Datensammlungen nicht offen sind, die Ergebnisse somit nicht ohne weiteres überprüft bzw. repro-

duziert werden können. Dass als grundlegende Anforderung für die hier zu projektierende Datensammlung die Erfüllung der FAIR-Prinzipien zugrundezulegen ist, versteht sich eigentlich von selbst, soll aber hier dennoch ausdrücklich hervorgehoben werden.

Schöch 2017, 225 f unterscheidet weiter drei Modi der Festlegung von Datensammlungen: repräsentative Zufallsstichproben, "balancierte Sammlungen", in denen versucht wird, Objekte so auszuwählen, dass Kombinationen aller für die jeweilige Forschungsfrage einschlägigen Merkmale in der Sammlung vorhanden sind, sowie schließlich das Verfahren der ‚opportunistischen Auswahl', die die Verfügbarkeit von Daten als wesentliches Auswahlkriterium an erste Stelle setzt.

Repräsentativität im statistischen Sinne setzt Bestimmung einer ‚Grundgesamtheit' voraus. Für Noichls Ziel, die ‚gegenwärtige Philosophie' als solche abzubilden ist eine solche Grundgesamtheit kaum konstituierbar. Selbst die auf Philpapers verzeichneten mehr als eine Million Aufsätze sind nicht als eine solche zu betrachten: Philosophie wird schließlich auch in Buchform publiziert. Auch aus inhaltlicher Sicht ist es zudem fraglich, ob die für die Bestimmung einer solchen Grundgesamtheit erforderliche Definition der Philosophie als Disziplin überhaupt möglich ist. Weitere Hindernisse für die Bestimmung der Grundgesamtheit selbst einer philosophischen Teildisziplin wie der Wissenschaftsphilosophie sind ebenfalls zu bedenken: selbst wenn es gelingen würde, ein solches Feld in operationalisierbarer Form einzugrenzen, müsste es auch in zeitlicher Hinsicht in einleuchtender Weise abgegrenzt werden. Dass der Beginn der Wissenschaftsphilosophie auf das Jahr 1934 festgelegt werden kann, ist aus fachlicher Sicht eine mit guten Gründen bezweifelbare Annahme.

Ob die von Malaterre et al. und Noichl vorgelegten Datensammlungen als ‚balanciert' gelten können, ist wohl ebenfalls eine strittige Frage. Wie man sie beantwortet, hängt davon ab, welche Merkmale als wesentlich für die behandelte Forschungsfrage anzusehen sind. Malaterre et al. gehen davon aus, dass nicht auf Englisch verfasste Texte für ihre Analyse vernachlässigbar sind bzw. der für deren Modellierung erforderliche Aufwand nicht notwendig ist.[6] Die Sprache, in der ein Aufsatz abgefasst ist, wird also nicht als wesentliches Merkmal aufgefasst, sondern kann für die Untersuchung vernachlässigt werden. Noichls Daten lassen die diachrone Dimension außer acht, hier gilt also das Veröffentlichungsdatum sowohl des zitierenden wie des zitierten Textes nicht als wesentliches Merkmal, das für die Balance der zugrundegelegten Datensammlung erforderlich wäre.

Schlussfolgerungen

Aus diesen Befunden sind meines Erachtens zwei Schlussfolgerungen für die Ausgestaltung einer offenen Datensammlung für die Philosophie und Philosophiegeschichte zu ziehen: erstens sollte man sich wohl von dem Anspruch, mit einer Datensammlung die Disziplin als solche abzubilden, verabschieden. Ziel sollte vielmehr die Zusammenführung von Teildatensammlungen sein, die die Abhängigkeit von den sie leitenden Forschungsfragen offenlegen und damit auch das Gebiet bestimmen, innerhalb dessen aus den in ihnen enthaltenen Datensätzen valide Schlüsse gezogen werden können. Zweitens bedarf die Eingrenzung auf nur einen sprachlich-kulturellen Zusammenhang der forschungsbasierten Begründung und sollte nicht allein pragmatisch motiviert sein.

Nicht nur in methodischer, sondern auch in inhaltlicher Hinsicht kann man aus beiden Arbeiten jedoch wertvolle Hinweise erhalten, in welchen Hinsichten die Ergebnisse digitalbasierter Forschungen für die Philosophie relevant sein können. Dies betrifft zunächst die Unterteilung der akademischen Philosophie in Teildisziplinen, d. h. den Prozess ihrer Spezialisierung. Malaterre et al. legen eine solche Teildisziplin als ‚Analyseeinheit' zugrunde, nämlich die Wissenschaftsphilosophie. Noichl untersucht die Auffächerung der Philosophie in solche Spezialdisziplinen und -diskurse. Die Organisation von Forschungsdatensammlungen zur Philosophie und Philosophiegeschichte wird sich also auch an solchen Einheiten zu orientieren haben.

Zugleich wird die Frage zu beantworten sein, ob, und wenn ja in welchem Sinne, sich solche subdisziplinären Einheiten von gesamtphilosophischen Traditionen abgrenzen lassen. Noichl diskutiert z. B. auch die Unterscheidung zwischen ‚analytischen' und ‚kontinentalen' Traditionen der Philosophie. ‚Kontinentale' Traditionen wie die der Phänomenologie werden sich jedoch kaum als Teildisziplinen definieren, sondern eher als Teiltraditionen der Philosophie. Noichls Analyse spiegelt dies, da zur Abgrenzung der kontinentalen Philosophie von anderen Teilbereichen ein Kanon zitierter Autor:innen herangezogen wird (Noichl 2021, 5094).

Mit dem von Noichl gewählten Werkzeug der Kozitationsanalyse lassen sich beide Dimensionen kaum voneinander abgrenzen. So wie Teildisziplinen Standardtexte zitieren, werden auch in philosophischen Traditionen gemeinsame Referenztexte als Bezugspunkte genutzt. Hier wird an einer inhaltlichen Analyse kein Weg vorbeiführen. Dass topic modeling in dieser Hinsicht eine hilfreiche Methode sein kann, zeigen Malaterre et al.: sie ermöglicht die Erschließung von Begriffskonstellationen und deren diachronen Verlauf. Während also die Identifikation von Traditionen und Schulbildungen wohl über die Betrachtung von Autor:innen und ihren Generationskohorten möglich sein dürfte, also durch Rekonstruktion von Kanonisierungs- und Dekanonisierungsprozessen von Personen, erlauben Verfahren der skalierten Erschließung von Inhalten Einblicke in die Kanonisierung und Dekanonisierung von Begriffen und deren Konstellationen.

Fazit: Anforderungen an philosophische Datensammlungen

Abschließend sollen die in diesem Beitrag entwickelten Anforderungen an eine Datensammlung digitaler Philosophie und Philosophiegeschichte kurz zusammengefasst werden. Digitale Philosophie benötigt Metadaten und Textdaten, die auffindbar, zugänglich, interoperabel und reproduzibel sind. Sowohl Textdaten als auch Metadaten sollten je nach Provenienz zumindest stichprobenhaft auf ihre Qualität hin überprüft werden. Die aus ih-

nen zu konstituierende Datensammlung sollte modular aufgebaut sein, um unterschiedlichen Forschungszielen, die den Teildatensammlungen zugrundeliegen, gerecht werden zu können. Datenquellen sind auf mögliche Verzerrungen und Auslassungen hin zu untersuchen. Diese sind, so weit sie sich pragmatisch aus dem zugrundeliegenden Forschungsziel der Teildatensammlung ergeben, zumindest explizit zu machen. Ein wichtiger Aspekt ist hierbei das Streben nach Multilingualität, um die globale Dimension der Philosophie angemessen abbilden zu können. Datensammlungen können dabei sowohl entlang von Teildisziplinen der Philosophie wie auch von Autor:innen, Epochen oder Traditionszusammenhängen konzipiert werden. Sie sollten es aber immer ermöglichen, auch die Entwicklung von Begriffen und Begriffskonstellationen - eines wesentlichen Mediums des Philosophierens - nachzuvollziehen.

Fußnoten

1. Mit der vorläufigen Klärung der hier als ‚material' bezeichneten Fragen ist natürlich, wie in der Begutachtung richtigerweise angemerkt, noch nicht alles gesagt, was aus fachphilosophischer Sicht zu Datensammlungen der Philosophie zu sagen wäre. Diese Debatten sollten jedoch zuerst innerhalb des Faches geführt werden und sind im disziplinären Zusammenhang der digitalen Geisteswissenschaften erst dann von Belang, wenn sie über das Fach hinausweisende Einsichten ermöglichen sollten (was wir naturgemäß erst wissen werden, wenn diese Debatten tatsächlich geführt worden sind).
2. Schöch 2017, 223 definiert sie als „Zusammenführung einzelner […] Datensätze nach einer Einheit stiftenden Systematik".
3. Der sich aus dieser Wahl des Gegenstandes ergebende Fokus auf die philosophische Zeitgeschichte ist kontingent: hier sind eben schon Untersuchungen mit Methoden der digitalen Geisteswissenschaften durchgeführt worden. Mutatis mutandis lassen sich jedoch die hier aufgeworfenen Fragen auf Datensammlungen zur Philosophiegeschichte übertragen.
4. Noichl 2021, 5091: „a visual representation of recent philosophy as a whole".
5. Malaterre u. a. 2021, 2886: „an empirical basis for what might otherwise be informal claims about the discipline and its evolution in the past eight decades as reconstructed from the perspective of its major journals".
6. Malaterre u. a. 2021, 2888: „Whenever journals were published in several languages, we retained only those articles that were written in English due to algorithmic linguistic constraints."

Bibliographie

Bourget, David, und David Chalmers. o. J. „PhilPapers: Online Research in Philosophy". Zugegriffen 1. August 2022. https://philpapers.org/.

Henny-Krahmer, Ulrike, und Frederike Neuber. 2017. „EDITORIAL: Reviewing Digital Text Collections – RIDE". *RIDE* 6. https://doi.org/10.18716/ride.a.6.0.

Malaterre, Christophe, Francis Lareau, Davide Pulizzotto, und Jonathan St-Onge. 2021. „Eight Journals over Eight Decades: A Computational Topic-Modeling Approach to Contemporary Philosophy of Science". *Synthese* 199 (1): 2883–2923. https://doi.org/10.1007/s11229-020-02915-6.

Noichl, Maximilian. 2021. „Modeling the Structure of Recent Philosophy". *Synthese* 198 (6): 5089–5100. https://doi.org/10.1007/s11229-019-02390-8.

Schöch, Christof. 2017. „Aufbau von Datensammlungen". In *Digital Humanities: Eine Einführung*, herausgegeben von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein. Stuttgart: J.B. Metzler. http://dx.doi.org/10.1007/978-3-476-05446-3.

Schöch, Christof, Frédéric Döhl, Achim Rettinger, Evelyn Gius, Peer Trilcke, Peter Leinen, Fotis Jannidis, Maria Hinzmann, und Jörg Röpke. 2020. „Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_006.