Forschungsdaten und Textpublikationen verknüpfen – Potenziale, Umsetzung und Herausforderungen

Julian Naujoks und Patrick J. Droß

Wissenschaftszentrum Berlin für Sozialforschung (WZB)

Als Extended Abstract zum Lightning-Talk „Gut verknüpft – besser auffindbar?" auf den E-Science-Tagen 2021 thematisiert dieser Beitrag die Verknüpfung von Text- und Datenpublikationen vor dem Hintergrund unserer Erfahrungen aus der Kuratierungs- und Veröffentlichungspraxis von sozialwissenschaftlichen Forschungsdaten am Wissenschaftszentrum Berlin für Sozialforschung (WZB). Im Sinne eines Werkstattberichts wird dabei auf Potenziale und Mehrwert, Umsetzung und Herausforderungen eingegangen. Unserer Erfahrung nach spielen bei der Verknüpfung vor allem das richtige Timing, die Heterogenität der Datenformate und die Prozesshaftigkeit eine zentrale Rolle. Dabei geht es nicht zuletzt um das komplexe und teils organisatorisch anspruchsvolle Vorhaben netzwerkartiger Verknüpfungen zwischen traditionellen und neuen Publikationsformaten im Sinne einer qualitätsgesicherten, offenen Wissenschaft.

1 Einleitung

Im Zentrum der Bemühungen um mehr offene Forschungsdaten stehen nach wie vor die Kernforderungen einer grundsätzlichen Zugänglichkeit und eindeutigen Refererenzierbarkeit von Forschungsdaten als eigenständige Forschungsergebnisse. Gleichwohl lohnt es sich, mit der zunehmenden Konsolidierung im Open-Data-Bereich auch einige bisher eher randständige Aspekte in den Blick zu nehmen. Hierzu zählt bspw. die Verknüpfung von Text- und Datenpublikationen. Mit dem folgenden Beitrag wollen wir uns diesem wichtigen Teilbereich im Gesamtprozess der Kuratierung und Veröffentlichung von Forschungsdaten widmen. Wir folgen dabei der Annahme, dass die Verbindung von nachhaltig veröffentlichten Daten mit (im besten Falle sogar frei zugänglichen Open-Access-) Textpublikationen erst die Transparenz schafft, die den Grundgedanken von Open Science, aber auch die Idee einer guten wissenschaftlichen Praxis auszeichnet. Im Rahmen dieses Extended Abstracts werden wir zum einen schlaglichtartig auf die Potenziale und praktischen Herausforderungen der Verbindung von Text- und Datenpublikationen hinweisen. Zum anderen werden wir unser eigenes Vorgehen am WZB vorstellen und wollen mit diesem Einblick zur aktuellen Diskussion beitragen.

Publiziert in: Vincent Heuveline, Nina Bisheh (Hg.): E-Science-Tage 2021. Share Your Research Data. Heidelberg: heiBOOKS, 2022.
DOI: https://doi.org/10.11588/heibooks.979.c13735

2 Potenziale

Die Verbindung von frei zugänglichen Daten mit (möglichst) ebenso frei verfügbaren Textpublikationen stellt hinsichtlich der Umsetzung der FAIR-Prinzipien in puncto Offenheit und Transparenz wissenschaftlicher Ergebnisse einen Idealzustand dar. Dies betrifft insbesondere den Start- und Zielpunkt von FAIRness (Auffindbarkeit und Nachnutzbarkeit). Zwar wird die Findability von Forschungsdaten durch neue Initiativen, Datenbanken und Suchmaschinen Schritt für Schritt erhöht, gleichwohl weisen aktuelle Studien auf „eine gewisse Divergenz zwischen Angeboten zur Datensuche und dem tatsächlichen Suchverhalten“ hin [1]. Demnach stoßen viele Forscher*innen im Alltag auch weiterhin vor allem über Textpublikationen bzw. allgemeine Forschungsliteratur auf die für sie relevanten Forschungsdaten (neben sozialen Kontakten/Netzwerken sowie Webseiten). Der direkte Verweis auf die einer Textpublikation zugrunde liegenden Primärdaten und deren Veröffentlichungsort würde die Auffindbarkeit von Daten entsprechend stark vereinfachen und weitere, ggf. aufwendige Rechercheschritte erübrigen.

Gleichzeitig fördern Textpublikationen, die aus einer Datenveröffentlichung direkt angesteuert werden können, das inhaltliche und methodische Verständnis der Daten (Stichwort Reusability). Sie erhöhen damit nicht nur das Potenzial ihrer direkten Nachnutzbarkeit, sondern leisten durch wichtige Kontextinformationen und -interpretationen insgesamt einen Beitrag zu einer verbesserten Datenqualität. Dies ist insofern von Bedeutung, da die Qualität der Datenveröffentlichungen letztlich mit darüber entscheidet, ob sich das grundlegende Ziel, nämlich die Nachnutzung der Daten, künftig in der erhofften Breite verwirklichen lässt [2].

3 Umsetzung am WZB

Das WZB war im Rahmen des Projektes SowiDataNet gemeinsam mit dem GESIS-Leibniz-Institut für Sozialwissenschaften, dem Deutsches Institut für Wirtschaftsforschung (DIW) und der ZBW-Leibniz-Informationszentrum Wirtschaft am Aufbau eines sozial- und wirtschaftswissenschaftlichen Fachrepositorium für Forschungsdaten beteiligt. Seit 2018 werden Daten von WZB-Forscher*innen in diesem Repositorium veröffentlicht. Parallel dazu wurde durch das Forschungsdatenmanagement am WZB ein umfangreiches Supportangebot etabliert. Dieses unterstützt die WZB-Forscher*innen von der finalen Aufbereitung der Daten, über notwendige Schritte der Anonymisierung bis hin zur Metadatenbeschreibung und finalen Veröffentlichung von Forschungsdaten im Repositorium. Flankiert wurde dies durch allgemeine Informationsangebote und Schulungen zu diesem Serviceangebot und einer grundsätzlichen Bewusstseinsförderung für die Idee offener Forschungsdaten. Wiederholt zeigte sich dabei, wie wichtig es ist, möglichst frühzeitig mit den Forscher*innen in Kontakt zu kommen. Um so eher die Datenpublikation eingeplant wird, um so besser lassen sich Daten vorbereiten, ggf. auch Bedenken der Forscher*innen aufgreifen und besprechen und im besten Falle auch Daten- und Textpublikationen koordinieren.

Im Sinne eines offenen Zugangs zu Forschungsoutput können die nachhaltig verfügbaren Forschungsdaten häufig auch noch durch Open-Access-Textpublikationen ergänzt werden.

Aus der Perspektive einer wissenschaftlichen Infrastrukturabteilung haben diese neuen Serviceangebote mit ihren vielseitigen Potenzialen natürlich auch diverse Fragen aufgeworfen. Die Verknüpfung von Forschungsdaten und Textpublikationen steht exemplarisch für die Anforderung, neue Services mit bestehenden Nachweissystemen und Publikationsprozessen zusammenzubringen. So waren wir am WZB – wie viele andere Einrichtungen vermutlich auch – damit konfrontiert, neben Bibliothekskatalog, Forschungsinformationssystem und Open-Access-Repositorium mit dem Open-Data-Repositorium ein weiteres System mit neuen Anforderungen in die Arbeitsabläufe zu integrieren.

Aufgrund fehlender Schnittstellen, abweichender (Metadaten-)Standards und nicht zuletzt unterschiedlicher Zuständigkeiten innerhalb unserer eigenen Abteilung war eine vollständig integrierte Lösung aller Informationssysteme vorerst nicht realisierbar. Neben der allgemeinen Bewusstseinsförderung galt es daher, sich um das konkrete organisatorische Zusammenspiel der Infrastrukturangebote zu kümmern, vor allem auch in Bezug auf die Verweise zwischen Text- und Datenveröffentlichungen. Aktuell bedeutet dies zumeist eine systematische manuelle Pflege der Verknüpfungen über die einzelnen Systeme hinweg: Am WZB bemühen wir uns im Zuge von Datenveröffentlichungen um die Aufnahme der Datenzitation inkl. Digital Object Identifier (DOI) direkt in den Zeitschriftenaufsatz oder in andere Textpublikationen (z.B. Monografie, Discussion Paper).
Zentrale Textpublikationen der Primärforscher*innen werden im Gegenzug im Datenrepositorium aufgenommen. Im Bibliothekskatalog erfolgt der Nachweis der Datenpublikation bei den entsprechenden Texten. Gleiches gilt – falls möglich – auch für die Metadaten im Open-Access-Repositorium. Im Idealzustand erhält man so eine wechselseitige Referenzierung, die – unabhängig vom Einstiegspunkt der Suche – Daten und dazugehörige Publikationen oder umgekehrt Publikation und dazugehörige Daten leicht auffindbar macht. Letztlich entsteht so ein komplexes Netzwerk wechselseitiger Verweise zwischen alten und neuen Publikationsformaten. Ein Trend, der sich generell abzeichnet und der Ausdruck einer digitalisierten Wissenschaft ist: „Je mehr sich die Kultur des Data-Sharing verbreitet, desto häufiger wird es damit auch zu netzwerkartigen Verzweigungen zwischen dem Output dieser ehemals prä-publikatorischen Phasen und den traditionellen Formaten der Ergebnispublikation kommen“ [3].

4 Herausforderungen

Bei der Verknüpfung von Text- und Datenpublikationen stoßen wir in unserer Kuratierungspraxis jedoch immer wieder auch auf Herausforderungen. Drei zentrale Themenbereiche sollen nachfolgend geschildert werden.

Timing der Verknüpfung

Zunächst ist die Verknüpfung von Text und Daten im Wesentlichen eine Frage des Timings. Um die Datenzitation bspw. in einen Journal-Artikel aufnehmen zu können, muss die Datenveröffentlichung naturgemäß bereits erfolgt sein, bevor die finale Version für den Druck an das Journal geht. Aufgrund zum Teil langer Review-Verfahren stehen die Chancen hier grundsätzlich nicht schlecht, aber es bedarf stets einer gewissen Planung. Oft verläuft dies aber auch nicht idealtypisch und das Zeitfenster ist deutlich kleiner, um die Verknüpfung herzustellen. Dann kann es dazu kommen, dass die Daten noch nicht fertig kuratiert und veröffentlicht sind, aber der Artikel bereits final eingereicht werden muss. Diese Fragen stellen sich zudem in der Praxis oft am Ende des Projekts. Dies ist häufig der Zeitpunkt, bei dem die Forscher*innen sich gedanklich oder gar physisch schon an einer ganz neuen Etappe des klassischen Forschungszyklus befinden (neuer Projektantrag, neue Einrichtung, usw.). Wir versuchen dies in zweifacherweise zu entzerren, indem wir erstens seitens des Forschungsdatenmanagements am WZB beraten und Strategien für ein geschicktes Handling der Publikationen vorschlagen. Zweitens nutzen wir die hilfreiche Möglichkeit, den DOI vorab im Repositorium zu reservieren. So kann eine vollständige Datenzitation bereits in den Artikel aufgenommen werden, auch wenn der Datensatz noch nicht publiziert ist. Bei diesem Vorgehen müssen jedoch einige zentrale Punkte berücksichtigt werden: So sollte definitiv klar sein, dass die Daten veröffentlicht werden können und dass dies zeitnah geschehen kann, damit ab Erscheinen des Journals auch der DOI auflöst. Wesentliche Prüfschritte der Kuratierung, wie z.B. die Eignung der Daten, Datenschutz und urheberrechtliche Fragen, müssen folglich geklärt sein.

Heterogenität der Datenformate und Veröffentlichungsorte

Best Practice im Open-Data-Bereich ist nach wie vor zweifellos die Veröffentlichung in einem nachhaltigen Forschungsdatenrepositorium. Hierzu zählt je nach Disziplin die Veröffentlichung eines Rohdatensatzes bzw. aufbereiteten Datensatzes, der mit standardisierten und fachlichen Metadaten beschrieben ist, einen persistenten Identifikator hat und ggf. mit zusätzlichen Begleitdokumenten versehen ist (in den Sozial- und Wirtschaftswissenschaften sind dies häufig Fragebogen und Codebook). In der Praxis bemerken wir aber gerade im Bereich der Verknüpfung von Text und Daten eine zunehmende Ausdifferenzierung der Datenformate und Publikationsorte. Ein Beispiel dafür ist die von wissenschaftlichen Verlagen verstärkt geförderte und teils geforderte Veröffentlichung von Forschungsdaten zu Replikationszwecken auf den Verlagsseiten/-angeboten. Prinzipiell sind verschiedene Wege der freien Verfügbarmachung von Forschungsdaten aus einer Open-Data-Perspektive zu begrüßen, doch leider sind diese in der Praxis nicht immer nachhaltig gestaltet. Dies ist ein Umstand, auf den auch der Rat für Informationsinfrastrukturen (RfII) kürzlich im Zuge der Diskussionen um die sogenannten Enhanced Publications aufmerksam machte [4]. Wir beobachten, dass in vielen dieser Fälle die Daten unvollständig aufbereitet und kaum beschrieben sind. Als Supplementary Material werden sie teils sogar als ZIP- oder Word-Datei auf der Journal-Webseite abgelegt. Häufig handelt es sich um Tabellen

und Auswertungen, aber nicht um die Daten selbst. Bei diesen im Rahmen des Einreichungsprozesses des Artikels eher beiläufig stattfindenden Datenveröffentlichungen fehlt es zumeist an Kuratierung, Standardisierung und Qualitätssicherung durch unterstützende Infrastrukturangebote. Im ungünstigsten Fall verschwinden Daten sogar hinter einer Paywall, so dass sich nur noch schwer von offenen Forschungsdaten sprechen lässt. Sollte dieser Weg gewählt werden, versuchen wir in der Praxis angesichts der Heterogenität von Datenformaten gerade bei den Replikationsdaten darauf hinzuwirken, dass hier ebenso ein Mindestmaß an Qualitätsstandards berücksichtigt wird. Falls möglich, empfehlen wir jedoch bevorzugt, die Daten über ein nachhaltiges Datenrepositorium verfügbar zu machen.

Prozesshaftigkeit der Forschung

Der stark prozesshafte Charakter der Forschung ist für die Erstellung einer netzwerkartigen Referenzierung zwischen den Publikationsformen eine besondere Herausforderung. So stellt sich bspw. die Frage, auf welche Textpublikationen im Datenrepositorium verwiesen werden soll und ob bzw. wie diese Verknüpfungen aktuell zu halten sind. Dies ist vor allem bei langfristig angelegten Forschungsprojekten der Fall, in denen ein zentraler Projektdatensatz veröffentlicht wird und in Folge noch eine Vielzahl von Textpublikationen anschließen kann. Wie können die Verknüpfungen aktuell gehalten werden, vor allem wenn zwischen der Publikation der Forschungsdaten und den daraus entstehenden Texten längere Zeiträume liegen? Am WZB versuchen wir, auf verschiedensten Ebenen die Verknüpfungen so gut wie möglich zu erfassen und herzustellen. Zunächst entscheiden die Forscher*innen – als Expert*innen für ihr jeweiliges Forschungsprojekt – natürlich selbst, welche relevante Textpublikationen initial im Datenrepositorium aufgenommen werden sollen. Wir weisen zudem darauf hin, dass jederzeit weitere Textpublikationen nachgemeldet werden können. Gleichzeitig versuchen wir aber auch in etwas systematischerer Form, die Angaben im Forschungsinformationssystem zu nutzen, in dem Forscher*innen sowohl ihre Text- als auch ihre Datenpublikationen eintragen und dabei auch wechselseitige Verbindungen angeben können. Schließlich wird durch uns beim Upload von Textpublikationen in das Open-Access-Repositorium auf Verknüpfungen geachtet, um auch dort Verweise auf Daten herzustellen.

5 Fazit

Die Verknüpfung von Text- und Datenpublikationen ist ein wichtiger Teilbereich des Kuratierungs- und Veröffentlichungsprozesses von Forschungsdaten. Auch wenn das technologische Potenzial einer automatisierten Verknüpfung riesig ist, zeigt unsere Erfahrung am WZB, dass wir im aktuellen Stadium noch vielfach mit manuellen Mitteln aktiv sein müssen. Gleichwohl können auch damit schon wichtige Ergebnisse erzielt werden: Die Auffindbarkeit der Daten steigt und die gegenseitige Referenzierung leistet nicht nur der

Transparenz und Reproduzierbarkeit der Forschungsergebnisse Vorschub. Sie steht gleichzeitig auch exemplarisch für Verknüpfungen zwischen alten und neuen Publikationsformaten einer digitalisierten Wissenschaft. Auch wenn einige wichtige Entwicklungen noch abzuwarten bleiben und mit den Fragen des Zeitpunkts, der Ausdifferenzierung der Datenformate sowie der Prozesshaftigkeit von Forschung wichtige Herausforderungen skizziert wurden, zeigt sich: Bei der Thematik geht es nicht nur um die bloße Verknüpfung von Texten und Daten. Vielmehr stehen hier sowohl Fragen der Datenqualität als auch grundsätzliche Aspekte der Openness von Forschung im Fokus.

Literaturverzeichnis

[1] Friedrich, Tanja; Recker, Jonas (2021): Auffindbarkeit und Nutzbarkeit von Daten. In: Markus Putnings, Heike Neuroth und Janna Neumann (Hg.): Praxishandbuch Forschungsdatenmanagement. Berlin: De Gruyter Saur, S. 405-426 (hier S. 423). DOI: http://doi.org/10.1515/9783110657807-023

[2] Blasetti, Alessandro; Droß, Patrick J., Fräßdorf, Mathis; Naujoks, Julian (2017): „Digital ist teilbar. Potenziale und Erfolgsbedingungen von Open Access und Open Data". In: WZB-Mitteilungen, H. 155, S. 34-37. DOI: https://doi.org/10.5281/zenodo.4733943

[3] Breuer, Constanze, Trilcke, Peer (2021): Die Ausweitung der Wissenschaftspraxis des Publizierens unter den Bedingungen des digitalen Wandels, Herausgegeben von der Arbeitsgruppe „Wissenschaftspraxis" im Rahmen der Schwerpunktinitiative „Digitale Information" der Allianz der deutschen Wissenschaftsorganisationen, S. 6. DOI: https://doi.org/10.48440/allianzoa.041

[4] Rat für Informationsinfrastrukturen (RfII) (2019): Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, S. 89. Online verfügbar unter: https://rfii.de/?p=4043