Working Paper

"Ist kooperativ jetzt umsonst? Die Ausweisung von Datenautorenschaft als neue Form wissenschaftlicher Reputation zur Förderung offener Forschungsdatenkulturen"

Vortrag auf der DHd 2018 "Kritik der digitalen Vernunft", Köln 2018

Katrin Moeller, Leiterin des Historischen Datenzentrums Sachsen-Anhalt

Institut für Geschichte, Martin-Luther-Universität Halle-Wittenberg
katrin.moeller@geschichte.uni-halle.de

Ich möchte Sie mit meinem Vortrag in die nahe Zukunft wissenschaftlichen Arbeitens entführen:

[**Folie 1**] Schließen Sie doch einfach mal die Augen und stellen Sie sich vor, wir haben das, was wir heute rund um das Forschungsdatenmanagement so intensiv diskutieren, bereits verwirklicht und geschafft. Es gibt ähnlich wie heute Bibliotheken und Bibliothekskataloge sogenannte Forschungsdatenrepositorien mit ihren Katalogen. Dort haben Sie eine Grundlagen ihrer Forschung - sei es Daten oder annotierten Quellen - publiziert. Weil sie solidarisch und kooperativ sind oder vielleicht auch weil Drittmittelgeber sie dazu ein wenig drängelten, haben Sie sich dem Mehraufwand unterworfen und ihre Quellen wundervoll dokumentiert, bereinigt, mit allen notwendigen Metadaten versehen und schließlich unter einer bestimmten Lizenz in einem solchen Repositorium publiziert. Gerade in den Geisteswissenschaften ist dies kein kleiner Schritt. Schließlich müssen sie dann alle Grundlagen ihres Forschens offen legen und bis zu einer Publikationsreife führen. Dieser Schritt ist wichtig, um die Qualität ihrer Daten zu sichern. Hier entsteht also ein gewichtiger neuer Arbeitsschritt, der die Grundlage der Forschung in ein Produkt - vergleichbar einem Buch - umwandelt. Zugleich wird dieses Datenprodukt also wie jedes andere Buch verwendbar, wenn auch vielleicht unter bestimmten Einschränkungen des Datenschutzes oder lizenzrechtlicher Verwertungsrechte. Den Lohn für diesen Mehraufwand gewinnen Sie mit einer zweiten Veröffentlichung. Andere Forscher können ihre Daten nun finden, herunterladen und damit arbeiten.

Viele fragen sich jedoch: Was passiert von jetzt ab mit meinen Daten?

Jetzt wollen wir wieder in die heutige Wirklichkeit zurückkehren - sie können die Augen wieder öffnen - und ganz im Sinne der Tagung möchte ich die Frage aufwerfen, was müssen wir tun, damit sich diese Wünsche des zukünftigen Forschens erfüllen, denn über die Frage: Was passiert von jetzt ab mit meinen Daten, darauf kann heute eher spekuliert werden.

[**Folie 2**] Wir gehen jetzt einfach mal davon aus, Sie haben Ihre Daten unter einer sogenannten CC-BY-Lizenz veröffentlicht. Dies bedeutet, sie stellen Ihre Daten unter der

Auflage der Namensnennung zur weiteren Nutzung frei. Was heißt Namensnennung hier in der heutigen Praxis genau? Letztlich kann dies auf drei verschiedenen Wegen passieren:

1.) der Nutzer Ihrer Daten zitiert sie

2.) der Nutzer Ihrer Daten kontaktiert Sie und publiziert mit Ihnen gemeinsam als Co-Autor seine auch auf ihren Leistungen basierenden neuen Forschungsergebnisse

3.) der Nutzer Ihrer Daten dankt ihnen im Vorwort oder Begleittext

Denn momentan gibt es das große Problem, dass Datenproduzenten keineswegs zahlreich zu den Forschungsdatenrepositorien strömen, um Daten einzuspeichern. Bisher bleiben Forschende eher zögerlich, die Repositorien sind noch leer. Daher müssen wir uns fragen, wie wir die Wirklichkeit auf das gewünschte Ziel hin weiterentwickeln. [**FOLIE 3**] Dazu möchte ich die drei Praktiken - Zitieren, Autorenschaft, Vorwort - darauf analysieren, welche Vor- und Nachteilen sie für den Datengeber und den Datennutzer bringen. Mir geht es dabei vor allem um die Effektivität und Einfachheit, aber auch um die wissenschaftliche Reputation. Denn wenn wir kooperative Nachnutzungen ermöglichen wollen, sollten für alle Bedürfnisse von Datenproduzenten, Datennachnutzern und den Institutionen der langfristigen Speicherung Win-Win-Situationen entstehen. Bei dieser Analyse möchte ich aber nicht stehen bleiben, sondern aus diesen Erkenntnissen einen Vorschlag zur separierten Ausweisung von "Textautoren" auf der einen und allen anderen Beiträgern - hier als Datenautoren benannt - auf der anderen Seite unterbreiten. Mit diesem Vorschlag möchte ich die Vorteile aller drei Praktiken - Zitieren, Autorenschaft, Vorwort - nutzen, um uns zum Ziel der angestrebten Wunschvorstellung kooperativen Forschens ein Stück weiter zu bringen. Denn wenn wir ganz ehrlich sind, ist Wissenschaft nicht nur ein System der Kooperation, sondern mindestens ebenso intensiv ein System der Konkurrenz. Bisher fehlen bei den meisten Denkansätzen des Forschungsdatenmanagements jedenfalls weitergehende Ansätze um Kooperation umfassend in wissenschaftliche Reputation zu überführen.

Schauen wir also auf die erste Form:

1.) Das Zitieren der Daten

[**Folie 5**] Das Grundprinzip des Zitats ist die Belegfunktion. Ein Zitat ist nach dem Gesetz dann zugelassen, wenn es eigene Ideen oder Gedanken unterstützt bzw. Ideen anderer aufgreift und in den eigenen Text integriert. Ist dem Recht genüge getan, verliert der Text auch ohne das Zitat nicht an Sinn.[1] Das Urheberrecht regelt bisher, dass in der wissenschaftlichen Forschung "kleine Teile eines Werkes, Werke geringen Umfangs sowie einzelne Beiträge aus Zeitungen oder Zeitschriften" für die wissenschaftliche Forschung ohne Kontakt mit dem Urheber oder separate Vergütung genutzt werden können, indem sie zitiert werden. Nach

[1] Thomas Schwenke, Texte richtig zitieren, statt plagiieren (Anleitung mit Checkliste), in: ILAWit. Blog zum Social Media-, Marketing, Online- und Datenschutzrecht, Berlin 2011 [http://rechtsanwalt-schwenke.de/texte-richtig-zitieren-statt-plagiieren-anleitung-mit-checkliste/].

einer Faustregel der Rechtspraxis durfte ein Zitat maximal ein Drittel eines Textes umfassen.[2] Das nun im März in Kraft tretende Urheberrechts-Wissensgesellschafts-Gesetz sowie vielleicht noch umfänglicher die erst noch zu beschließende europäische Gesetzesinitiative über das Urheberrecht im digitalen Binnenmarkt wollen solche Möglichkeiten stärken. Sie erlauben beispielsweise die Auswertung von Texten in Form von Text Mining und die Benutzung von Teilen von Datenbanken oder ganzer Bildwerke in Form solcher Zitate. Vor allem wird es möglich, Daten und Texte für diesen Zweck zu vervielfältigen. Dabei wurde der Nutzungsumfang zum Zweck der Forschung definiert und bei einer Veröffentlichung in § 60c auf 15 % eines Werke begrenzt. Ob dies nun eher einen Schritt vor oder zurück für die Nachnutzung von Daten in einem großen Umfang bedeutet, möchte ich - als ausgewiesene Nicht-Juristin - hier gar nicht vertieft diskutieren.

Mir geht es ja auch mehr um die rechtlichen Festlegungen zum Zitat allgemein und die daraus zu abstrahierenden Vor- und Nachteile für die wissenschaftliche Reputation.

[**Folie 6**] (**flexible Nutzung**) Die Vorteile Forschungsdaten durch das Zitat zu benennen, liegen auf der Hand: Es handelt sich hier um eine sehr einfachen und durchgesetzten wissenschaftlichen Standard! Das ist kein kleines Gut! Es macht uns frei von der Last, einen komplizierten Kontakt zwischen dem Bereitsteller von Daten und dem Nachnutzer von Daten herzustellen. Denn wie soll ich einen Datenproduzenten eigentlich erreichen, der vielleicht nicht mehr im Wissenschaftsbetrieb tätig ist.

Aber dieses schöne einfache System hat zwei Nachteile, die viele Forschende momentan davon abhalten, ihre Daten tatsächlich kooperativ zur Verfügung zu stellen.

[**Folie 6**] a. (**keine oder wenig wissenschaftliche Reputation**) Möchten Forschende zwar nicht unbedingt finanziellen Nutzen und Verwertungsrechte aus Ihren Leistungen ziehen, die ja auch meist mit Steuermitteln finanziert wurden. Sie möchten aber natürlich wissenschaftliche Reputation daraus gewinnen - denn das ist die eigentliche Währung des Forschens. Gerade in den Geisteswissenschaften sind die mit einer klaren Schöpfungshöhe nach dem Urheberrecht versehenen Daten fast so etwas wie ein Lebenswerk. Solche Daten veröffentlicht man dann vielleicht am Ende seines Lebens, um dieses Lebenswerk dauerhaft zu sichern. Genau dies kann man auch in der Praxis beobachten, die Rentner kommen um ihre Daten zu speichern. Solche Daten dann aber in die notwendigen Strukturen zu integrieren ist sehr aufwändig, weil sie eben meist nicht den Standards folgen und Formate der Langzeitarchivierung repräsentieren. Daher wollen wir heute Anreize setzen, solche Insellösungen zu vermeiden und daher Daten von Anfang an in gemeinsame digitale Infrastrukturen mit schnellen Veröffentlichungswegen stecken. Daher müssen wir die wissenschaftliche Reputation solcher kooperativen Leistungen stärken. Dies ist aber in den Geistes-, Kultur- und Sozialwissenschaften bisher kaum der Fall, weil Zitationsindexe nicht

[2] Thomas Schwenke, Texte richtig zitieren, statt plagiieren (Anleitung mit Checkliste), in: ILAWit. Blog zum Social Media-, Marketing, Online- und Datenschutzrecht, Berlin 2011 [http://rechtsanwalt-schwenke.de/texte-richtig-zitieren-statt-plagiieren-anleitung-mit-checkliste/].

für die Bewertung wissenschaftlicher Leistungen zählen. Auch in den STM-Fächern (Science, Technology und Medizin) sind diese nicht unumstritten.

[**FOLIE 7**] b. (**problematische Rechte**) Das zweite Problem ist der Umfang, in dem Daten genutzt werden. Wenn 15 % die Grenze dessen darstellt, was bei einer Nachnutzung in Form des Zitats genutzt werden kann, ist das Zitat mit einer deutlichen Grenze versehen. Eine umfängliche Nutzung der Daten ist damit nicht möglich. In diesem Fall ist also eine richtige Autorenschaft des Datengebers notwendig.

Blicken wir daher 2. auf die Co-Autorenschaft

In den STM-Fächern hat sich nämlich nicht das Zitat, sondern die Praxis der Autorenschaft etabliert. Jeder, der irgendeinen Beitrag zu einem Forschungsergebnis leistet, wird als Co-Autor genannt. Vorteil dieser Praxis ist eine Erhöhung der wissenschaftlichen Reputation der einzelnen Forscherinnen und Forscher, weil sie entgegen des Zitierprinzips auch bei einer Nachnutzung ihrer Daten zu Autoren des neuen Forschungsergebnisses werden. Dies ist in allen wissenschaftlichen Disziplinen für die Leistungsbewertung förderlich. Außerdem erhöht diese Praxis die Bereitschaft - Daten zu teilen - erheblich. [**FOLIE 8**] Im Endeffekt dieser Entwicklung steht in den STM-Fächern allerdings ein Resultat, dass mich an dieser Praxis auch wieder zweifeln lässt. Wenigstens müssen wir in den Geisteswissenschaften vielleicht nicht die Fehler der Naturwissenschaften noch einmal wiederholen. Aufgrund der Leistungsevaluation kam es nämlich quasi zur Explosion der Beiträger zu einem Forschungsergebnis. Rekordhalter ist momentan ein 2015 publizierter kollaborativer Artikel aus der Physik, der stolze 5.154 Autoren aufzählt. Wer den eigentlichen Text und das konkrete Forschungsergebnis produziert hat bzw. wer genau welchen Anteil an diesen Leistungen hatte, ist in solchen Texten nicht mehr wirklich nachvollziehbar. [Folie 9] Die DHd-Arbeitsgruppe "Digitales Publizieren" hat in ihrem Working Paper aus dem Jahr 2016 daher vorgeschlagen, bei der Publikation von kollaborativen Texten die einzelnen Rollen von Autoren klarer zu differenzieren und führt insgesamt 29 verschiedene Beitragsarten wie etwa bei der Autorenschaft als Hauptautor, Nebenautor oder Co-Autor an, um zu unterscheiden. Der Vorschlag geht meines Erachtens in die richtige Richtung und stellt für die eigentliche Publikation auch kein Problem dar. Für die Weiterverwendung bei Nachnutzung und auch für die Abbildung in bibliothekarischen Nachweissystemen ist dieser Vorschlag aber zu komplex. Diese Schwäche haben meiner Meinung nach momentan viele etwas idealistisch oder euphorisch entwickelte Standards, die im Praxistest einfach zu viel auf einmal wollen. Sie fordern möglichst alle Details zu erfassen, sind aus diesen Gründen wenig effektiv und treiben so die personellen, zeitlichen und finanziellen Kosten unnötig in die Höhe.

Halten wir die Vor- und Nachteile der Co-Autorenschaft fest:

[Folie 10] a. (**Funktionelle Unterscheidung**) Die Funktion der Beiträger an einem Forschungsergebnis wird nicht transparent. Wichtig ist in dieser Hinsicht vor allem die Unterscheidung der Funktion von eigentlichen Textautoren und aller anderen Beiträger.

[Folie 10] b. (**Wissenschaftliche Reputation**) Zwar besitzt ein Co-Autor eine deutlich höhere wissenschaftliche Reputation als beim Zitat, der eigentliche Textproduzent hat aber nur als Erst- oder Letztautor ein höheres Prestige als alle anderen Beiträger, obwohl bspw. fünf Autoren Hauptautoren sein können. Dies gilt allerdings weitgehend für die STM-Fächer, weil in den geisteswissenschaftlichen Fächern bis heute Co-Autorenschaften aufgrund von wissenschaftlichen Leistungen außerhalb des eigentlichen Textes nach wie vor sehr unüblich sind. Maximal findet eine Ausweisung von Autoren im Rahmen der Beigabe "unter Mitwirkung von" statt. Hier findet sich aber eine interessante funktionelle Unterscheidung verschiedener Rollen am Textbeitrag, die einfach und effektiv ist.

[Folie 10] c. (**Verantwortung, Rechte**) Nächstes Problem der Co-Autorenschaft ist die Verantwortung für einen Text, die der Datenproduzent gleichermaßen trägt wie der Textproduzent. Die Realität bildet dies meist nicht ab, weil bspw. der Ersteller einer Software am Endergebnis der Publikation gar nicht beteiligt sein muss. Mitunter kann diese Praxis, obwohl theoretisch alle Autoren zu einem Ergebnis zustimmen müssen, zu Unstimmigkeiten über den Inhalt von Ergebnissen meist im Nachhinein führen. Gleiches gilt für Nutzungs- und Verwertungsrechten am Text, die aber anders regelbar sind als die Pflichten.

Lassen Sie uns noch einen letzten - sehr kurzen Blick auf die dritte Form - den Dank im Vorwort oder Begleittext tun.

3. (**Anonymität, keine Rechte, keine Pflichten, keine wissenschaftliche Reputation, flexible Nutzung**) Er bietet keinerlei nachweisbare Formen der wissenschaftlichen Reputation, sondern gleicht einer netten menschlichen Geste und ist damit die höchste Ausdrucksstufe für kooperatives Zusammenwirken. Diese Form der Nachweisung bringt dem Datengeber einen entscheidenden Vorteil, es bleibt intransparent was ganz genau er getan hat und quasi bleibt damit auch anonym welche Zuarbeiten genau in ein Forschungsergebnis eingeflossen sind. Letztlich gibt es für diesen Typ der Nachweisung mittlerweile auch eine CC-Lizenz, nämlich CC-Zero. Ob dies dem Ansprüchen an die Transparenz und Nachvollziehbarkeit wissenschaftlichen Arbeitens entspricht, lasse ich mal dahingestellt.

Zusammenfassend lässt sich also die Schlussfolgerung ziehen, dass jedes System seine Vor- und Nachteile besitzt. Da ich mir jetzt zum Ziel setze, möglichst viele wissenschaftliche Datenproduzenten von einer kooperativen Teilung ihrer Daten zu überzeugen, möchte ich also möglichst alle die Vorteile für die Lizensierung von Daten bündeln, die solche kooperativen Anreize stärken und gleichzeitig die ganz unterschiedlichen wissenschaftlichen Gepflogenheiten positiv ausgestalten.

Welchen Ausweg gibt es nun dafür?

[Folie 11] 1. Zunächst kann ich festhalten, dass aus Perspektive der wissenschaftlichen Reputation über alle Fächer hinweg eine Nutzung des Prinzips der Co-Autorenschaft die beste Form für die wissenschaftliche Evaluation bietet.

[Folie 11] 2. Was wir vornehmen müssten, ist zunächst eine funktionelle Trennung des Text- und Ergebnisproduzenten vom "Zulieferer" also etwa Datenautor oder Programmierer bzw. aller anderen Beteiligten an einem Forschungsergebnis. Vermutlich ist es am einfachsten, die Hauptautoren in einer Publikation von allen anderen Autoren zu differenzieren.

[Folie 11] 3. Über diesen Weg kann ich auch eine Ausdifferenzierung von Rechten und Pflichten vornehmen.

Gesetzlich muss also klar geregelt werden, welche Rechte und welche Pflichten durch die Nachnutzung von Daten entstehen. Der Verzicht auf Nutzungsrechte in der Wissenschaft wird momentan über Lizenzen geregelt und ist rechtlich relativ einfach freizustellen. Viel problematischer ist die Freistellung von Pflichten und Verantwortlichkeiten. Den genau diese Verantwortlichkeit für einen Text macht die Kontaktaufnahme zwischen Textautor und Beiträger notwendig. Der Vorteil des Zitierens ist aber gerade die Kontaktlosigkeit. Es muss also klar geregelt werden: Nur der Textautor trägt eine Verantwortung für den Inhalt eines Textes, während alle anderen Beiträger dies genau wie bei einem Zitat nicht tun. Sie haben zwar am Ergebnis einen Beitrag, können aber nicht dafür zur Verantwortung gezogen werden.

Im Ergebnis hätten wir also ein differenziertes, aber sehr einfaches Rollenmodell wie heute etwa bei Sammelbänden, wo zwischen der Rolle von Herausgebern und Autoren der einzelnen Beiträge unterschieden würde.

Wie das aussehen kann, dafür möchte ich kurz noch ein Beispiel präsentieren:

[Folie 12] Unsere Autorin, ich nenne sie mal Maria Musterfrau, möchte zum Thema Berufsverläufe und Karrieren von Reformierten in der Stadt Halle im 18. Jahrhundert forschen. [Folie 12] Durch eine Recherche in einem Forschungsdatenrepositorien konnte sie ermitteln, dass die Tauf-, Heirats- und Geburtsregister einer reformierten Gemeinden bereits vollständig als Daten vorliegen, die mit freien Lizenzen unter der Auflage der Namensnennung (CC-BY) von Dora Datenreich publiziert wurden. Sie muss also vom Wunsch der Namensnennung ausgehen, ansonsten hätte Dora Datenreich die Lizenz CC-Zero gewählt, um anonym zu bleiben. [Folie 12] Dies hat etwa Gustav Geistvoll getan, dessen Biografien sie ebenfalls nutzt, den sie aber deshalb nicht als Autor benennen braucht, weil der Autor durch seine Lizenz darauf verzichtet hat.

Maria Musterfrau verwendet diese Daten in großen Teilen, ergänzt sie durch zahlreiche eigene Informationen und schreibt jetzt eine Publikation zum Thema. [Folie 12] In diesem Fall würde das von mir vorgeschlagene Prinzip der getrennten Ausweisung von Textautor - Maria Musterfrau - und Datenautor im Fall der - Dora Datenreich - greifen. In der neuen Publikation würde Maria Musterfrau als die eigentliche Textproduzentin und Dora Datenreich als Beiträgerin auftauchen.

Maria Musterfrau hat aufgrund der Lizenzen alle Verwertungsrechte an dieser Publikation und auch die vollen Pflichten. Dora Datenreich erscheint daher auch abgesetzt (im Sinne des

heutigen "unter Mitwirkung von) als Datenautorin. Möglich wird dies nur, weil sie ihre Daten in einem öffentlichen Repositorium mit klaren Lizenzen und Qualitätsmaßstäben veröffentlicht hat. Da in solchen Systemen einzelne ForscherInnen über ID-Systeme persistent nachgewiesen werden können - etwa über ORCID - kann Dora Datenreich später auch erfahren, dass sie zur Autorin eines neuen Textes geworden ist. Sie kann also diese Leistung ihrer wissenschaftlichen Bibliografie hinzufügen, trägt aber an dieser Veröffentlichung keine Verwertungsrechte oder Verantwortlichkeiten.

[Folie 12] Neben dieser Möglichkeit bleibt es darüber hinaus natürlich auch weiterhin üblich Werke und Daten ganz normal zu zitieren, nämlich immer dann, wenn sie nur in kleinerem Umfang benutzt werden. Dies kann etwa beim Text Mining der Fall sein, wenn hier nur Teilinformationen aus Daten und Texten zu Korpora zusammengestellt werden und hier im Fall von Bruno Basiswissen visualisiert ist.

[Folie 13] Letztlich verbinden wir mit dieser Form der differenzierten Ausweisung von Text- und Datenautoren und der Festlegung funktionaler Prinzipien von Verantwortlichkeit und Rechten die Vorteile des Zitierprinzips mit den Vorteilen der Autorenschaft. Durch die Lizenzierungsmodelle der Datenrepositorien können Daten frei nachgenutzt werden, ohne dass Datengeber und -nehmer direkt in Kontakt treten müssten. Gleichzeitig ist ein differenzierter Ausweis des jeweiligen Beitrags von Datengebern und -nehmern möglich. Verantwortlichkeiten werden strikt geklärt. Überdies werden Datengeber in einer adäquaten Form gewürdigt, die in allen Forschungsdisziplinen wissenschaftliche Reputation herstellen. Denn durch die Ausweisung des Datengebers als Datenautor benennt der Textautor die Herkunft seiner Quellen nach ihrem verwendeten Umfang exakt und macht hier die Unterschiede zum Zitat klar. Diejenigen, die explizit mit der Textabfassung betraut waren und Verantwortung für den Inhalt des Forschungsergebnisses tragen, werden nun auch deutlich als solche benannt und erhalten - selbst bei einem größeren Forscherkollektiv - das gleiche Prestige zugewiesen. Der Datenautor erhält gleichfalls wissenschaftliche Reputation, wenn auch etwas abgestuft. Gefördert werden damit Anreize zum kooperativen Arbeiten, weil Leistungen dennoch zur vollen wissenschaftlichen Reputation führen und Kooperation zusätzlich belohnen - und dies übergreifend zur Verbesserung der bisherigen Verfahrensweisen sowohl in den STM-Fächern wie auch in den Humanities.