

Der Artikel diskutiert anhand von Beispielen aus der Osteuropäischen Geschichte, inwiefern digitalisierte Archivalien als Forschungsdaten zu begreifen sind. Dabei werden die folgenden beiden Thesen aufgestellt: (1) Retrodigitalisierte Archivalien können in begründeten Fällen als Forschungsdaten angesehen und publiziert werden – hierfür werden im Beitrag entsprechende Anforderungen formuliert. (2) Generell bietet es sich jedoch an, retrodigitalisierte Archivalien unter dem Oberbegriff digitale Daten zu subsumieren.

Taking examples from Eastern European history, the article discusses the extent to which digitised archive materials can be regarded as research data. It puts forward the following two theses: (1) In justified cases, retro-digitised archive materials can be regarded as research data and published; the requirements for this are set out in the article. (2) In general, however, it makes sense to subsume retro-digitised archive materials under the generic term »digital data«.

ARNOŠT ŠTANZEL

Digitalisierte Archivalien: Daten oder Forschungsdaten der (Osteuropäischen) Geschichte?

Ein Klärungsvorschlag zum Datenbegriff in den Geschichtswissenschaften

Einleitung

Eine der Fragen, welche die Forschenden in der Osteuropäischen Geschichte seit dem Angriff Russlands auf die Ukraine im Februar 2022 umtreibt, ist die nach der Zugänglichkeit von Archivmaterialien in Russland. Die Möglichkeiten, in den dortigen Archiven zu forschen, sind für deutsche Historiker*innen derzeit und vermutlich noch für längere Zeit nicht gegeben. Ein Punkt, der bei Diskussionen innerhalb der Deutsch-Russischen Historikerkommission oder auf Workshops wie dem des Max Weber Netzwerks Osteuropa zur Frage nach Forschungsdaten aus russischen Archiven (München, 4.7.2024) immer wieder auftaucht, ist, ob Scans und Fotografien von Dokumenten aus russischen Archiven als Forschungsdaten veröffentlicht werden sollten. Als Trägerin des Fachinformationsdienstes (FID) Ost-, Ostmittel- und Südosteuropa wird die Bayerische Staatsbibliothek (BSB) als zentrale Infrastruktureinrichtung für die deutschen Osteuropastudien¹ ebenso mit diesen Fragen konfrontiert und regelmäßig zu entsprechenden Gesprächsrunden eingeladen.

Ausgehend von der Frage nach der Veröffentlichung von Archivalien als Forschungsdaten diskutiert der vorliegende Beitrag, inwiefern es sinnvoll ist, retrodigitali-

sierte Archivdokumente (analoge und originär digitale Archivalien werden ausgeklammert) als Forschungsdaten zu begreifen und welche Infrastruktureinrichtungen dafür zuständig sein können.² Dabei werden zwei Thesen vertreten: (1) Retrodigitalisierte Archivalien können in begründeten Fällen als Forschungsdaten angesehen und publiziert werden – hierfür werden im Beitrag entsprechende Anforderungen formuliert. (2) Generell bietet es sich jedoch an, retrodigitalisierte Archivalien unter dem Oberbegriff digitale Daten³ zu subsumieren. Forschungsdaten unterscheiden sich von Daten aufgrund ihrer Eigenschaft als Ergebnisse wissenschaftlicher Forschung zu einem bestimmten Thema. Sie sind ein in den Geschichtswissenschaften bislang wenig genutztes Publikationsformat, das zu etablierten Forschungsergebnissen wie Aufsätzen oder Monografien hinzukommt.

Dieser Differenzierungsvorschlag erfolgt vor allem auf der Basis von Erfahrungen des Verfassers aus der Forschungsdatenmanagementpraxis und stellt einen insbesondere pragmatischen und weniger informationswissenschaftlich begründeten Ansatz dar, Kommunikation und Infrastrukturentwicklung zum Thema Forschungsdatenmanagement zu betreiben. Durch diese begriffliche Schärfung fällt es leichter, erstens die Debatten zu

Forschungsdaten in der Fachcommunity der Geschichtswissenschaften zu führen, zweitens passgenaue Forschungsdatendienste für diese aufzubauen und drittens die Vorteile neuer Forschungsdateninfrastrukturen und ihre Unterschiede zu bestehenden Angeboten wie beispielsweise Digitalen Sammlungen mit Retrodigitalisaten zu kommunizieren.

Archivalien werden im vorliegenden Beitrag als Dokumente verstanden, die von einem haltenden Archiv als unbefristet aufzubewahren bewertet werden. Hinzu kommt die Einschränkung, dass solche Archivalien im Fokus stehen, die ursprünglich in analoger⁴ Form erstellt worden sind (prinzipiell stellen sich die gleichen Fragen auch bei anderen, schwer zugänglichen Materialien wie Handschriften oder Autographen aus anderen Gedächtniseinrichtungen wie Bibliotheken oder Museen). Die Diskussion der eingangs gestellten Frage erfolgt zunächst stark aus der Perspektive der Osteuropastudien, kann aber stellvertretend für weitere Area Studies gelten: Forschende und insbesondere Historiker*innen sind in diesen mit Archiven konfrontiert, zu denen längere Reisen in Kauf genommen werden müssen, oder die teilweise für Forschende gar nicht mehr zugänglich sind und die selbst nur wenige Retrodigitalisate produziert haben oder zur Verfügung stellen. Der Bedarf an diesen digitalisierten Archivalien und die dadurch potenziell mögliche Einsparung von Zeit- und Finanzressourcen ist ungleich höher als bei Beständen in Deutschland, die evtl. noch nicht digitalisiert, aber an sich gut erreichbar und zugänglich sind. Abschließend wird versucht, die Aussagen auf die Geschichtswissenschaften insgesamt zu übertragen und so zu der dort stattfindenden Diskussion nach Archivalien als Forschungsdaten beizutragen.

Archivalien im Spiegel von Forschungsdatendefinitionen

Eine erste Annäherung an die Frage nach der Zuordnung von Archivalien erfolgt mithilfe eines Blicks auf gängige Definitionen von Forschungsdaten. Laut DFG sind »Forschungsdaten [...] eine wesentliche Grundlage für das wissenschaftliche Arbeiten. [...] Zu Forschungsdaten zählen u. a. Messdaten, Laborwerte, audiovisuelle Informationen, Texte, Surveydaten, Objekte aus Sammlungen oder Proben, die in der wissenschaftlichen Arbeit entstehen, entwickelt oder ausgewertet werden.«⁵ Fasst man Archivdokumente als audiovisuelle Informationen (Fotografien der analogen Dokumente) oder Texte auf, ist die Zuordnung klar als gegeben zu betrachten. Die Definition des Rats für Informationsinfrastrukturen (RfII) wiederum legt einen stärkeren Fokus auf den Aspekt der Ergebnishaftigkeit: »Forschungsdaten sind Daten, die im Zuge wissenschaftlicher Vorhaben entstehen, z. B. durch Beobachtungen, Experimente, Simulationsrechnungen, Erhebungen, Befragungen, Quellenforschungen, Aufzeichnungen, Digitalisierung, Auswertungen.«⁶ In diesem Rahmen entstehen Forschungsdaten bei der Analyse von Quellen – also auch Archivalien – diese selbst sind jedoch

keine. Im Forschungsdatendienst OstData,⁷ der an der Osteuropaabteilung der BSB mit weiteren Projektpartnern koordiniert wird, stehen Archivdokumente auch außen vor, da es ebenfalls um neu entstehende Resultate von Forschung geht: »Forschungsdaten stellen all jene Daten dar, die auf Grundlage wissenschaftlicher Methoden gesammelt, erzeugt, entwickelt, beschrieben und/oder ausgewertet sowie angemessen dokumentiert werden.«⁸ Der Fokus bei OstData liegt ganz klar auf neu erstellten, digital vorliegenden Forschungsdaten, insbesondere auch vor dem Hintergrund, dass für diese Art von Publikationen zum Projektbeginn 2019 die entsprechenden Infrastrukturen nicht vorhanden waren, im Gegensatz etwa zu Archivalien.

Mit dem NFDI-Konsortium 4Memory gibt es eine Institution, die sich schon qua Arbeitsauftrag mit der genannten Definitionsfrage auseinandersetzen muss und in deren Förderantrag ein breiter Zugang gewählt wird, der auch historische Quellen wie Archivalien als Forschungsdaten betrachtet: »Historians have yet to reach a consensus on defining ›research data‹. [...] The growth of digital RDM has caused many historical researchers to describe the historical record itself – i. e. the sources – as ›research data‹ [...]. The definition of historical sources as ›research data‹ is not without controversy, but it has been tentatively established that digital representations of historical records are fitting subjects for RDM. In essence, ›data‹ now encompasses all forms of representation, from the analogue object to the reused data produced during research [...]. This wider understanding leads to a general definition of research data in the historically oriented disciplines: any information, regardless of its provenance, is considered ›research data‹ as soon as it is collected, described, analysed, evaluated and / or created and stored in machine-readable form for the purpose of maintaining the traceability of research results or for archiving, citation and further processing. Hence, research data include all media formats and digital representations of analogue sources – publications, sources, metadata, controlled vocabularies, statistical or simulation data, etc. – as well as any information that can be used to answer research questions.«

Dieser kurze Blick auf vier Forschungsdatendefinitionen offenbart die Spannweite der möglichen Einordnung von Archivalien als Forschungsdaten. Hauptunterscheidungsmerkmal ist, inwiefern Forschungsdaten Ergebnis einer wissenschaftlich-methodischen Verarbeitung mit einer klar benennbaren Fragestellung sind oder alle Daten, die zur Beantwortung einer Forschungsfrage herangezogen werden.

Datenbegriffe auf retrodigitalisierte Archivalien angewendet

Die Diskussion der vier genannten Definitionen hat gezeigt, dass es aus Sicht der Historiker*innen legitim sein kann, auch digitalisierte Archivalien als Forschungs-

daten anzusehen. Dahinter steht die Betrachtungsweise, dass alle Daten, sobald sie von Historiker*innen gesammelt, beschrieben, analysiert oder evaluiert werden, zu Forschungsdaten werden.

Ohne Frage hat die Retrodigitalisierung der letzten Jahre zu einem enormen Anstieg an nun online einsehbaren Materialien geführt. Diese wurden von der historischen Forschung in analoger Form als Quellen bezeichnet sowie genutzt und werden auch in ihrer digitalen Form weiterhin so genannt. Gleichzeitig können diese Quellen auch als Daten, aus denen historische Erkenntnisse gewonnen werden, angesehen werden. In analoger Form ist dies eher ungewöhnlich und widerspricht dem aktuellen Sprachgebrauch, in digitalisierter Form und mit entsprechenden Verarbeitungsschritten erscheint die Bezeichnung Daten aber durchaus sinnvoll, wie im Weiteren zu sehen ist. Die Nutzung dieser Quellen hat sich im Laufe der Digitalisierungsvorhaben stark vereinfacht: Sei es durch die Möglichkeit für Forschende, digitale Kopien von Dokumenten mit der Handykamera zu erstellen, sei es durch die systematische Digitalisierung ihrer Bestände durch die Archive selbst. So können von Forschenden bei Archivaufenthalten weitaus größere Mengen an Archivdokumenten als früher gesichtet werden, da Exzerpte nicht mehr vor Ort gemacht werden müssen, sondern erst nachträglich erstellt werden können, oder gar direkt vom heimischen Arbeitsplatz im Falle der bereits erfolgten Digitalisierung. Der ganze Prozess ist kostengünstiger geworden, weil keine Kopiegebühren pro Seite mehr anfallen und auch die Logistik und Lagerung digitaler Kopien weitaus einfacher ist als die von Papierkopien – es reicht die eigene Festplatte, und bestenfalls zusätzliche Backups auf USB-Sticks, Netzlaufwerken oder in der Cloud. Auf diesem Wege landet mittlerweile eine beachtliche Menge an Daten mit ihnen innewohnenden Informationen auf den digitalen Laufwerken von Historiker*innen.

An dieser Stelle lohnt es sich mit Blick auf die weitere Argumentation nachzufragen, was für Daten da eigentlich auf den Datenträgern landen. Wie Forschungsdaten ist auch der allgemeinere Begriff der Daten nicht eindeutig zu definieren, je nach Fach gibt es unterschiedliche Herangehensweisen. Laut Wikipedia stehen diese als Pluralwort beispielsweise »für durch Beobachtungen, Messungen u. a. gewonnene [Zahlen]werte sowie darauf beruhende Angaben oder formulierbare Befunde.«⁹ Der RfII sieht in seiner Definition Daten als Ausgangspunkt für die Gewinnung von Wissen an: »Daten sind Grundlage und Teil eines wissenschaftliche Ergebnisse generierenden sowie zugleich ökonomisch wertschöpfenden Kreislaufes aus Daten, Informationen und Wissen. Informationen können aus strukturierten Daten, Wissen aus überprüften Informationen gewonnen werden. Aus wissenschaftlicher Perspektive dienen Daten der Gewinnung und Validierung von (gesichertem) Wissen.«¹⁰ Ein ähnlicher Datenbegriff wird im Rahmen von OstData

herangezogen, wenn Daten als Informationen über Eigenschaften von Analyseeinheiten angesehen werden.¹¹ Aus einer mehr informationstechnologischen Sicht sind laut der Norm des internationalen Technologiestandards ISO/IEC 2382-1 Daten wiederum »a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation, or processing«¹² – in der Informatik und Datenverarbeitung versteht man dementsprechend Daten gemeinhin als (maschinen-)lesbare und -bearbeitbare, in der Regel digitale Repräsentationen von Information. Diese Datenbegriffe eint im weitesten Sinne, dass es um Informationen¹³ als Grundlage von Wissen geht und dass sie in Bezug auf digitale Daten auch die Maschinenlesbarkeit als wichtiges Kriterium hervorheben.

Bei den meist als JPG oder TIFF abgelegten Scans bzw. Fotos von Archivdokumenten handelt es sich aus technischer Sicht um strukturierte Informationen zu den Farbwerten, die die Wiedergabe auf Bildschirmen ermöglichen. Die Retrodigitalisierung kann an dieser Stelle zunächst als bloßer Wechsel des Trägermediums verstanden werden. Die gespeicherten digitalen Daten repräsentieren somit zunächst einmal nicht die für Historiker*innen im Archivdokument enthaltenen interessanten Informationen in maschinenlesbarer Form, nämlich meistens den Text. Dieser ist zwar von Menschen einfach lesbar, sei es auf Papier oder eben digital am Bildschirm. Für die Nutzung mit einem Computer fehlen aber noch entscheidende Schritte, um die für historisch forschende Personen relevanten Informationen in maschinenlesbarer Form zu erhalten, sprich den Text und nicht die Bildwerte. Um Daten zu diesen Informationen zu generieren, bedarf es weiterer Schritte, hierbei sind insbesondere die Methoden der Optical Character und Handwritten Text Recognition (OCR und HTR), ggf. auch der manuellen Transkription zu nennen. Auf diese Weise entsteht letztlich Text in verschiedenen Dateiformaten (.xml, .txt, .docx, etc.). Dies ist keine jüngere Neuerung, die meisten Retrodigitalisierungsvorhaben etwa von Bibliotheken haben von Anfang an nicht nur bildförmige Digitalisate erzeugt, sondern haben auch gleich OCR durchgeführt, um direkt digitale Volltexte für die Nutzenden mit anbieten zu können und so neben der einfacheren Zugänglichkeit einen weiteren Mehrwert zu generieren.

Anforderungen für »archivalische Forschungsdaten«

Diese Betrachtungsweise hilft dabei, zu umreißen, was eigentlich die primär interessanten Daten für die Geschichtswissenschaften sind.¹⁴ Die auf den Volltexten basierende Möglichkeit, Bücher und andere Dokumente komplett zu durchsuchen, bedeutet für Historiker*innen eine große Vereinfachung, sodass diese Art der Digitalisierung einen wahren Quantensprung gegenüber rein analogen Zeiten darstellt. Die so entstandenen Volltexte sind zudem die Grundlage für Methoden der Digital Humanities wie Named Entity Recognition, Tagging-Ver-

fahren, Annotationen oder die Verknüpfung mit anderen Daten (Linked Open Data). Voraussetzung dafür ist aber, dass aus Bildwerten, die Schriftsymbole abbilden, maschinenlesbar codierte Textstrings werden.¹⁵

Anforderung 1: Eine erste Anforderung, retrodigitalisierte Archivalien als Forschungsdaten zu begreifen, ist es, dass die textuellen Informationen in maschinenlesbarer Form vorliegen.

Eng damit zusammen hängt die Güte der Fotografien bzw. Scans: Ist die Qualität der Abbildungen (Belichtung, Bildschärfe und -auflösung) ausreichend für die weitere Nutzung, etwa für OCR-Anwendungen? An dieser Stelle wären dann ggf. Einzelfallentscheidungen zu treffen: Ist die Qualität einer Fotografie ausreichend, um sie mit digitalen Methoden weiter zu verarbeiten oder mit weiteren Informationen anzureichern? Fragen der inhaltlichen Qualität rücken dagegen bei Archivalien eher in den Hintergrund, da diese zum einen von der konkreten Forschungsfrage abhängen, zum anderen bereits entsprechende Selektionsprozesse seitens der Ursprungsarchive stattgefunden haben.

Anforderung 2: Hinsichtlich der Qualität der Scans und Fotografien ist festzuhalten, dass diese bestenfalls so gut ist, dass eine Bearbeitung mit OCR-Technologien möglich ist, um so maschinenlesbare Volltexte zu generieren, deren Fehlerquote niedrig genug ist, um eine sinnvolle Nachnutzbarkeit zu ermöglichen.

Die bisherigen Anfragen an die Osteuropaabteilung der BSB zielten darauf ab, dass Fotografien bzw. Scans von Dokumenten aus russischen Archiven explizit als Forschungsdaten veröffentlicht werden sollen. Damit wollen die Forschenden anderen interessierten Wissenschaftler*innen die Möglichkeit bieten, diese Dokumente für die eigene Forschung nutzen zu können. Verbunden sind diese Anfragen mit der Hoffnung, dass sich diese Art des Teilens von retrodigitalisierten Archivalien weiter verbreitet, sodass die Menge an zugänglichen Materialien außerhalb Russlands beständig steigt.¹⁶ Technisch betrachtet handelt es sich dabei um keine komplizierte Aufgabe, im Kern geht es darum, Dateien (in diesem Fall Bilddateien, ggf. der Bequemlichkeit wegen gebündelt) mit Metadaten zu veröffentlichen, wie dies bei Forschungsdatenrepositorien und Publikationsservern bislang auch schon passiert.

Viel gewonnen wäre damit aber noch nicht, denn Quellenkritik, Interpretation und Nachnutzung sind davon abhängig, dass der adäquate Entstehungskontext bis zu einem gewissen Grad mitgeliefert wird: Aus welchem Archiv stammen die Dokumente? Zu welchem Bestand gehören die Dokumente? Um welches Schriftstück handelt es sich? Was ist der Überlieferungskontext? Welche weiteren Provenienzinformationen gibt es? Diese und andere (Meta-)Informationen zum Kontext eines Dokuments sind essentiell. Sie müssen im besten Fall umfassend übernommen werden, um Metadaten zur Verfügung zu stellen, eine effiziente Indexierung für die Suche

zu ermöglichen und das Problem zu umgehen, dass die Provenienzinformationen der ursprünglichen, bestandshaltenden Institution verloren gehen oder unzugänglich werden.

Anforderung 3: Eine weitere Anforderung ist somit, dass ein ausreichendes Set an sinnvollen Metadaten¹⁷ als Kontextinformation mit angegeben wird.

Wenn sich die Praxis etablieren sollte, digitalisierte Archivdokumente zu teilen, würden die Mengen wachsen und die genannten Erschließungsdaten würden umso wichtiger, um passende Archivalien zu finden, da die Navigation über Archivtekonik (Bestände/Archiveinheiten) ja wegfällt. Als Vorteil könnte sich erweisen, dass solche Archivkonvolute eine tiefere Erschließung erfahren können, als dies in den Findbüchern der Ursprungsarchive der Fall war. So können Teilergebnisse der wissenschaftlichen Analyse, beispielsweise Annotationen, Normdatenverknüpfungen und Schlagworte in den Kollektionen retrodigitalisierter Archivquellen hinzugefügt werden.

Ein weiterer Punkt ist die Frage, ob die vorliegende Gesetzeslage (Urheberrechte, Datenschutz, etc.) und die Regularien der einzelnen Archive es erlauben, digitale Kopien von Archivdokumenten zu erstellen und zu veröffentlichen. Nur zwei von einer ganzen Reihe zu klärender Fragen sind etwa: In Deutschland entstehen keine Urheberrechte auf staatliche Dokumente, aber wie ist dies in anderen Ländern wie Russland? Und wie sieht es mit personenbezogenen Daten aus? Bei älteren Archivalien ist dies eher unproblematisch, aber bei neueren Dokumenten der zeithistorischen Forschung stellen sich hier zum Teil heikle rechtliche (und ethische) Fragen.

Anforderung 4: Die Veröffentlichung retrodigitalisierter Archivalien muss rechtlich möglich sein und durch die Regeln des Ursprungsarchivs gedeckt sein.

Eng mit den rechtlichen Aspekten sind Fragen der Datensouveränität und -ethik verbunden. Bei den großen Anteilen von staatlichen Überlieferungen und administrativen Aktenbeständen in öffentlichen Archiven ist davon auszugehen, dass insbesondere autoritäre Staaten wie Russland es nicht gutheißen, wenn Quellen zur eigenen Geschichte im Ausland gesammelt, publiziert und somit der staatlichen Kontrolle entzogen werden. Geschichts- und Erinnerungspolitik ist eine der Stützen von Putins Herrschaft in Russland, und somit sind zusammenhängende Aspekte wie der Zugang zu Archivalien von großer politischer Bedeutung. In der Mikroformensammlung der BSB finden sich zum Beispiel eine Reihe von Verfilmungen russischer Archivalien, die in den russischen Archiven – nach einer Öffnungs- und Transparenzphase in den 1990er-Jahren – wieder unter Verschluss stehen. Mit Blick auf den Globalen Süden stellt sich zudem für andere Area Studies die Frage, inwiefern die Bereitstellung von Archivdokumenten im Globalen Norden eine Replikation alter kolonialer Muster darstellt. Aus einer demokratietheoretischen Perspektive kann es zwar auch

Gründe geben, die dafür sprechen, Archivalien im Ausland zu veröffentlichen, um so Transparenz für staatliches Handeln herzustellen – sozusagen ein WikiLeaks für Archivalien. Es handelt sich hierbei allerdings um ethisch durchaus anspruchsvolle Fragen, die dringend und im Einzelfall zu klären sind.

Anforderung 5: Fragen der Datensouveränität und Datenethik sind gründlich abzuwägen.

An dieser Stelle kann festgehalten werden, dass im Kontext der Osteuropaforschung ein großes Interesse am Zugang zu digitalisierten, im Original nur schwer erreichbaren oder gar nicht zugänglichen Archivalien besteht. Werden die fünf genannten Anforderungen – Maschinenlesbarkeit, Bildqualität, Metadaten, Rechteklärung und Datensouveränität / Datenethik – erfüllt, können die digitalisierten Archivalien pragmatisch als Forschungsdaten betrachtet werden und in entsprechenden Infrastrukturen veröffentlicht werden. Realistischerweise wird es dabei aber um digitalisierte Archivalsammlungen bzw. -konvolute gehen, die eng mit einer Forschungsfrage verbunden sind und sich dabei auf eine begrenzte Zahl an Archivfonds und -einheiten beschränken werden. Zum einen, weil einzelne Forschende nur eine begrenzte Zahl an Archivalien sichten können. Zum anderen, weil das Vorliegen einer wissenschaftlichen Fragestellung vor dem Hintergrund der meisten Definitionen die Voraussetzung ist, um von Forschungsdaten zu sprechen. Eine umfassende Digitalisierung ganzer Archive würde im Sinne der hier genannten Ausführungen in die Rubrik der Datengenerierung fallen und erfordert bei mehreren Millionen Seiten Dokumenten auch eine andere technische Infrastruktur, als dies Forschungsdatenrepositorien bieten.

Anforderung 6: Das Vorliegen einer wissenschaftlichen Fragestellung, die der Sammlung und Analyse der Archivalien zugrunde liegt, ist die sechste Bedingung.¹⁸

Nach dieser vorläufigen Klärung der Forschungsdatenhaftigkeit von Archivalien stellt sich natürlich auch die Frage nach der Umsetzung.

Welche Informationsinfrastrukturen übernehmen Hosting und Nachweis dieser Art von Forschungsdaten?

Deutsche staatliche Archive verfügen über die Erfahrung und das Wissen zur Erschließung von Archivalien, es liegt daher nahe, zunächst an diese zu denken und die digitalisierten Archivalien dort als Forschungsdaten bereitzustellen. Allerdings ist es unwahrscheinlich, dass diese angesichts ihrer begrenzten Ressourcen und ihres eigentlichen gesetzlichen Auftrags Mittel für diese Aufgabe bereitstellen werden. Zumal die deutschen Archive derzeit noch damit beschäftigt sind, überhaupt die eigenen Bestände zu digitalisieren. Bislang ist nur ein Bruchteil der Archivfonds digitalisiert¹⁹, mit Normdaten beschrieben und in Archivportalen recherchierbar. Zudem ist zu fragen, ob seitens der Archive eigentlich gewollt ist, dass Forschende selber Archivalien digitalisieren und dann auf einer Forschungsdateninfrastruktur ihrer Wahl

veröffentlichen, ggf. in Konkurrenz zu den eigenen Angeboten und eventuell überquer mit den langjährig etablierten archivalischen Praktiken und Qualitätsstandards?

Als weitere Option kommen Forschungsdatenrepositorien an Universitätsbibliotheken und anderen wissenschaftlichen Bibliotheken in Frage. Bei diesen fehlt wiederum das archivarische Fachwissen, welches sicherlich von Vorteil ist, um solche Art von Forschungsdaten mit Metadaten zu beschreiben, die eine gute Auffindbarkeit ermöglichen. Und zu ausländischen Archivdokumenten aus regionalwissenschaftlicher Forschung fehlen meistens auch die fachspezifischen Kenntnisse aus den Area Studies. Mit den Universitätsarchiven gibt es Einrichtungen, die mit archivalischem Fachwissen beitragen könnten. Hier wäre es interessant, in Erfahrung zu bringen, inwiefern es auf dieser Ebene Überlegungen oder bereits konkrete Zusammenarbeit mit den jeweiligen Forschungsdatenrepositorien der Universitäten gibt und ob auch der Fall digitalisierter Archivalien schon diskutiert wurde.

Größere Forschungseinrichtungen, etwa innerhalb der Leibniz-Gemeinschaft, verfügen über eigene, institutionelle Forschungsdatenrepositorien. Im Falle der Osteuropaforschung ist beispielsweise das Leibniz-Institut für Ost- und Südosteuropaforschung (IOS) zu nennen, das mit dem Datenportal LaMBDa²⁰ über eine entsprechende technische Infrastruktur verfügt, die den eigenen Forschenden offen steht. Sofern in LaMBDa digitalisierte Archivalien veröffentlicht werden würden (dies ist derzeit noch nicht der Fall), stellt sich zusätzlich die Frage des überregionalen Nachweises dieser Forschungsdaten. An dieser Stelle sind sicherlich regionalwissenschaftliche FID zu nennen. Der FID Ost-, Ostmittel- und Südosteuropa an der BSB etwa plant im Rahmen seiner Beteiligung an OstData einen zentralen Nachweis von Forschungsdaten der Forschung zum östlichen Europa.

OstData baut zudem ein fachspezifisches Forschungsdatenrepositorium auf und bietet sich daher auch als Publikationsort für diese Art von Forschungsdaten an. Allerdings sollten in den ursprünglichen Infrastrukturplanungen retrodigitalisierte Archivdokumente nicht veröffentlicht werden: Zum einen, weil der Fokus von OstData auf Forschungsdaten als Resultate wissenschaftlicher Analysen in digitaler Form liegt, wie die oben zitierte Forschungsdatendefinition zeigt, und zum anderen, weil Archive mit ihrer Erfahrung und ihren Infrastrukturen von OstData als besser geeignete Institutionen angesehen werden. Vor dem Hintergrund der in der Einleitung genannten Anfragen von Osteuropaforschenden wurde diese Haltung bei OstData kritisch hinterfragt (daraus resultiert unter anderem dieser Artikel). Die Position von OstData wäre demzufolge in dem Sinne anzupassen, dass, wenn die genannten Anforderungen erfüllt sind, digitalisierte Archivdokumente für die Forschung dauerhaft zugänglich gemacht werden können. Dies ist ein Weg, in Einzelfällen den Zugang zu Dokumenten zu ebnen, wenn das Archiv nicht mehr einfach

erreichbar ist. Durch den Nachweis von Forschungsdaten, welche OstData zusammen mit dem FID Ost im Recherchetool *osmikon.search*²¹ anbieten will, wird ein überregionaler Nachweis von Forschungsdaten erfolgen, allerdings in einem Discovery Service, der bislang nicht für Archivalien bekannt ist. Müsste also parallel auch ein Nachweis im Archivportal-D²² erfolgen? Das Forschungsdatenarchiv *Discuss Data*,²³ das sich ebenfalls an die Osteuropaforschung wendet, hat sich im Nachgang auf den oben erwähnten Workshop des Max Weber Netzwerks Osteuropa auch schon dafür entschieden, eine eigene Kategorie für Archivmaterialien einzurichten, allerdings ist diese auf Metadaten beschränkt und umfasst bislang auch nur einen Eintrag.

Als letzter potenzieller Infrastrukturanbieter sind die NFDI-Konsortien zu nennen, insbesondere das für die Geschichtswissenschaften, 4Memory. In der Selbstbeschreibung der im Aufbau befindlichen Forschungsdateninfrastruktur sind digitalisierte Archivalien Teile des Arbeitsauftrags. Zudem bringt das Konsortium gute Voraussetzungen mit, sich dieser Aufgabe anzunehmen, da es mit den Forschenden der Fachcommunity, Archiven und Bibliotheken die zentralen Akteure und insbesondere deren Kenntnisse vereint. Allerdings befindet sich ein eigenes Forschungsdatenrepositorium von 4Memory noch in einer frühen Planungsphase. Zudem würde es, entsprechend der genannten Anforderungen, automatisierte OCR- und HTR-Pipelines für die unterschiedlichsten Sprachen und Schriftsysteme (nicht-lateinische Schriften wie das Kyrillische) brauchen, um die Volltexte zu den Archivscans zu generieren – wenn diese nicht durch die Forschenden selbst erstellt werden würden.

Trotz dieser Einschränkungen sind potenzielle Infrastrukturen vorhanden, die retrodigitalisierte Archivalien als Forschungsdaten sammeln und auffindbar machen können.

Retrodigitalisierte Archivdokumente: Daten und nicht Forschungsdaten

Unter bestimmten Gesichtspunkten kann es also durchaus denkbar sein, digitalisierte Archivalien als Forschungsdaten in entsprechenden Repositorien zu veröffentlichen. Die Argumente für ein solches Vorgehen ergeben sich insbesondere im Rahmen von Area Studies wie der Ost-, Ostmittel- und Südosteuropaforschung und dort speziell der Osteuropäischen Geschichte, in welchen der Zugang zu Archivalien mit prohibitiv hohen Kosten verbunden sein kann und damit eine digitale Kopie auf einem (deutschen) Forschungsdatenrepositorium von unschätzbarem Wert wäre.

Aber macht es generell Sinn, diese Art von Materialien in den Geschichtswissenschaften als Forschungsdaten zu bezeichnen? Die eingangs aufgestellte These postuliert, dass es gewinnbringender ist, von Daten und Forschungsdaten zu sprechen: Daten sind die Grundlage für die historische Forschung, publizierte Forschungsda-

ten (neu erstellt und digital) sind davon eine Teilmenge. Letztere würden sich vor allem dadurch auszeichnen, dass sie das Ergebnis der wissenschaftlichen Arbeit mit vorhandenen digitalen Daten wie beispielsweise retrodigitalisierten und maschinenlesbaren Archivalien sind. Aus der Perspektive der im Rahmen von OstData gesammelten Erfahrungen sprechen die folgenden Gründe für eine solche Kategorisierung.

Diese ist etwa anschlussfähig an den bestehenden und etablierten Sprachgebrauch in den Geschichtswissenschaften und erlaubt es gleichzeitig, neue Aspekte wie die Publikation digitaler Forschungsdaten zielgerichtet zu adressieren: »Archivalien« bezeichnen demnach weiterhin eine altbekannte Quellenart, es erfolgt keine semantische Neucodierung des bei Historiker*innen etablierten Begriffs. Dieser Begriff macht auch bei retrodigitalisierten wie auch originär digitalen Archivalien weiterhin Sinn, sofern diese nun am Bildschirm, statt auf Papier gelesen werden. Erst wenn die Archivalien als maschinenlesbare Texte vorliegen und im Rahmen historischer Forschung mithilfe digitaler, computergestützter Methoden und Verfahren wie Skripten und Algorithmen analysiert und bearbeitet werden können, bietet es sich an, von Daten zu sprechen. Davon zu trennen ist dann die Erstellung von Forschungsdaten als Ergebnis der Analyse analoger und digitaler Daten.

Durch diese Aufteilung ist es auch einfacher, die nötigen Kompetenzen aus dem Instrumentarium des Forschungsdatenmanagements an Forschende zu vermitteln (Backups erstellen, richtige Dateiformate wählen, etc.) und diese dabei zu begleiten, die Ergebnisse ihrer Forschung nicht nur als Aufsatz oder Monografie, sondern auch als digitalen Forschungsdatensatz zu veröffentlichen. »Klassisch« arbeitenden Historiker*innen kann so besser vermittelt werden, welche Vorteile sich jenseits des Lesens am Bildschirm ergeben, wenn Quellen als digitale, maschinenlesbare Daten vorliegen. Digital arbeitenden Historiker*innen, die bereits Daten computergestützt analysieren, können dann zielgerichtete Hinweise zu Softwaretools oder passenden Repositorien für die entstehenden Forschungsdaten gegeben werden. Ein weiteres Beispiel zu vermittelnder Kompetenzen ist, dass die Suche nach digitalisierten Archivquellen andere Fähigkeiten erfordert als die Erstellung neuer Forschungsdaten. Es handelt sich zwar beide Male um Aspekte von Datenkompetenz, aber mit unterschiedlichen Schwerpunkten: Einerseits die Nutzung von Discovery-Systemen, Kenntnisse von GLAM-Einrichtungen und deren Digitalisierungsstrategien, etc., andererseits Kenntnisse sinnvoller Softwarelösungen für die Auswertung und Nutzung gebräuchlicher Dateiformate und so weiter. Heutige Tools in der Forschungsdatenberatung wie Datenmanagementpläne (DMP) oder der Datenlebenszyklus gehen bereits davon aus, dass nicht alle Daten im Rahmen eines Forschungsprojektes auch als Forschungsdaten veröffentlicht werden. Nicht umsonst handelt es sich auch nicht

um Forschungsdatenmanagementpläne und -zyklen, sondern beide operieren mit dem weiteren Begriff der Daten.

Den Infrastruktureinrichtungen aus dem Kulturerbebereich kommt im Kontext dieser Datenlandschaftsbeschreibung die Aufgabe zu, Ihre Bestände und Sammlungen so zu digitalisieren, dass maschinenlesbare Daten für die Geschichtsforschung bereitstehen und über Schnittstellen frei abrufbar sind. Dies wird von den GLAM-Institutionen auch bereits gemacht, aufgrund finanzieller und rechtlicher Beschränkungen vielleicht nicht in dem Tempo, wie es sich Forschende wünschen würden, aber die Menge der so entstehenden Daten wächst. Das NFDI-Konsortium 4Memory plant beispielsweise mit seinem Data Space²⁴ ein Interface, über welches diese Daten zentral abrufbar sein sollen, sodass Forschende diese dann analysieren und bearbeiten können. Entsprechend der in diesem Beitrag vertretenen Position sind die in diesem Prozess entstehenden Forschungsdaten dann in einem geeigneten Repositorium zu veröffentlichen, sei es bei 4Memory selbst oder in institutionellen, fachspezifischen oder generischen Forschungsdatenrepositorien, mit Unterstützung von FDM-Stellen und Data Stewards. Diese publizierten Daten sollten selbstverständlich auch von 4Memory im Data Space nachgewiesen werden und würden so automatisch zu Daten für andere Forschende. Ein solcher Data Space und entsprechend konfigurierte zu Grunde liegende Datenquellen – Forschungsdatenrepositorien, digitale Sammlungen, digitale Archivalienrepositorien – würde auch verhindern, dass mit neuen Infrastrukturen für Forschungsdaten zusätzliche Datensilos entstehen. Ein solcher Data Space sollte zudem perspektivisch eine weitere Quelle von Daten beinhalten, welche für die historische Forschung zukünftig ebenfalls von unschätzbarem Wert sein werden. Der Begriff Open (Governmental) Data bezeichnet Infrastrukturen, welche originär digital entstandene Daten aus der staatlichen Verwaltung als Daten frei zur Verfügung stellen, beispielsweise Geodaten, statistische Daten entsprechender Landesämter, Wahlergebnisse, Haushaltsdaten oder Baumkataster. Im Rahmen dieser Open Data werden so teilweise auch Informationen direkt zugänglich, die bislang noch über den Umweg der Archive ihren Weg zu Forschenden gefunden haben.

Durch die Verwendung des Kategorienpaares Daten/Forschungsdaten wird auch die Gefahr gebannt, dass durch einen sehr breit gewählten Forschungsdatenbegriff sämtliche Schritte der Digitalisierung von Kulturerbe – im Sinne vom Online-Stellen eines Katalogs mit im Volltext durchsuchbaren, maschinenlesbaren Materialien – als Forschungsdatenprojekte etikettiert werden und die für FDM-Maßnahmen im engeren Sinne vorhandenen Mittel hierfür verwendet werden. Vielmehr handelt es sich bei den genannten Schritten um die grundlegenden Voraussetzungen dafür, überhaupt sinnvoll digital arbeiten zu können und auf dieser Basis Forschungsdaten zu erstellen.

Fazit

Von Forschenden digitalisierte Archivdokumente können als Forschungsdaten in Betracht gezogen und veröffentlicht werden, wenn sie die folgenden Anforderungen erfüllen:

- Die digitalen Daten sind maschinenlesbar;
- Metadaten liefern Kontextinformationen;
- Die Bildqualität ist ausreichend;
- Rechte sind geklärt;
- Datensouveränität und -ethik sind berücksichtigt;
- Die Daten sind aufgrund einer wissenschaftlichen Fragestellung zusammengestellt worden.

Für die praktische Umsetzung braucht es noch Infrastruktureinrichtungen, wobei es hier mit NFDI und fachspezifischen Repositorien durchaus potenzielle Adressaten gibt.

Generell jedoch erscheint es sinnvoller, digitalisierte Archivalien als Daten zu betrachten und als Forschungsdaten nur datenförmige Ergebnisse wissenschaftlicher Forschung zu einem bestimmten Thema zu begreifen. Auf diesem Wege wird die Kommunikation mit Historiker*innen deutlicher, Informationsangebote und Infrastrukturen können passgenauer angeboten werden, und Forschungsdatenprojekte werden nicht mit umfassenden Digitalisierungsaufgaben überfrachtet.

Der Verfasser dankt Florian Grumbach und Gregor Horstkemper für die anregenden Gespräche und hilfreichen Kommentare beim Verfassen des Beitrags.

Anmerkungen

- 1 Innerhalb derer die Osteuropäische Geschichte die größte Fachdisziplin ist.
- 2 Der Beitrag spiegelt die persönliche Meinung des Verfassers wider, basierend auf seiner langjährigen Arbeit im Bereich Forschungsdatenmanagement bei FID Ost-, Ostmittel- und Südosteuropa, OstData und NFDI4Memory.
- 3 Der Beitrag befasst sich nur mit digital vorliegenden Daten, sei es born digitals oder retrodigitalisierten Ursprungs. Damit soll nicht verneint werden, dass es auch analoge Daten gibt (sehr viele sogar, z. B. Prä-digitale Wetterdaten), allerdings basiert der Beitrag auf den Erfahrungen mit der Arbeit digital vorliegender (Forschungs-)Daten und deren Publikationen in entsprechenden technischen Infrastrukturen und zieht daraus seine Schlüsse.
- 4 Es steht außer Frage, dass Archivalien mittlerweile auch in rein digitaler Form vorliegen, Stichwort E-Akte zum Beispiel. Bei digitalen Archivalien stellen sich aber andere Fragen des Zugangs, da etwa Reise- und Zugangsbeschränkungen von Staaten weniger eine Rolle spielen, dafür aber technische Barrieren, wenn etwa bestimmte IP-Ranges oder Anfragen aus bestimmten Ländern komplett blockiert werden. So sind beispielsweise eine Reihe staatlicher Bibliotheken in Russland aus dem Netz der BSB nicht mehr erreichbar, was sich freilich durch die Nutzung von VPNs oder dem TOR-Browser verhältnismäßig leicht umgehen lässt.

- 5 DFG. *Leitlinien zum Umgang mit Forschungsdaten* [Zugriff am: 24. Oktober 2024]. Verfügbar unter: www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf
- 6 RfI – Rat für Informationsinfrastrukturen. *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen: RfI, 2016, S. A-13.
- 7 OstData – Forschungsdatendienst für die Ost-, Ostmittel- und Südosteuropaforschung [Zugriff am: 24. Oktober 2024]. Verfügbar unter: <https://www.osmikon.de/servicemenue/ueberuns/ueber-ostdata>
- 8 FRANK, Ingo; KÖNIG, Sandra; KÖRFER, Anna-Lena; KROLL, Benedikt; KURZWEIL, Moritz; SKOWRONEK, Thomas; ŠTANZEL, Arnošt und Peter VALENA. *OstData Forschungsdatendefinition*. 2022 [Zugriff am: 30. Oktober 2024]. Verfügbar unter: <https://doi.org/10.5281/zenodo.6088524>
- 9 Wikipedia. *Daten* [Zugriff am: 18. Oktober 2024]. Verfügbar unter: <https://de.wikipedia.org/wiki/Daten>
- 10 RfI – Rat für Informationsinfrastrukturen. *Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen: RfI, 2016, S. A-6.
- 11 Vgl. HJØRLAND, Birger: (2020). Data: With big data and database semantics. In: Birger Hjørland und Claudio Gnoli, Hrsg. *ISKO Encyclopedia of Knowledge Organization*. Version 1.4. 2022 [Zugriff am: 29. Oktober 2024]. Verfügbar unter: <http://www.isko.org/cyclo/data>
- 12 Wikipedia. *Daten* [Zugriff am: 18. Oktober 2024]. Verfügbar unter: <https://de.wikipedia.org/wiki/Daten>
- 13 Auch der Begriff der Informationen ist alles andere als eindeutig definiert, zur Diskussion siehe beispielsweise KUHLEN, Rainer und Wolfgang SEMAR. *A 1 Information – ein Konstrukt mit Folgen*. In: Rainer Kuhlen; Dirk Lewandowski; Wolfgang Semar und Christa Womser-Hacker. *Grundlagen der Informationswissenschaft*. Berlin: De Gruyter, 2023, S. 1-26.
- 14 Beispielsweise bei Handschriften kann natürlich neben dem Text selbst auch die grafische Gestaltung der Seite von Forschungsinteresse sein.
- 15 Je nach Inhalt der Archivalien wäre auch zu nennen, dass etwa schematische Abbildungen zu Vektografiken oder die einzelnen Datenpunkte von Tabellen korrekt erfasst werden.
- 16 Ein ähnliches Ziel verfolgt beispielsweise auch das US-amerikanische Projekt »A Search for Research. Mapping Archives in a Time of War«, vgl. *A Search for Research: Mapping Archives in a Time of War* [Zugriff am: 22. Oktober 2024]. Verfügbar unter: <https://networks.h-net.org/search-research>
- 17 In Anlehnung an das Archivportal-D könnten dies beispielsweise die folgenden Angaben sein: Name des Ursprungsarchivs, dortige Signatur und Bestand, Entstehungskontext, Laufzeit, zusätzliche Angaben zur Provenienz sofern vorhanden, Vermerke zu den enthaltenen Dokumenten in einer Archiveinheit, etc., vgl. Archivportal-D [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://www.archivportal-d.de/>
- 18 Theoretisch ist auch der Fall denkbar, dass Forschende auf Archivalien stoßen, die nicht zu ihrem Forschungsprojekt passen, die sie aber für so interessant erachten, dass Sie davon gerne eine digitale Version als Forschungsdatensatz veröffentlichen möchten. In solchen Einzelfällen kann eine Veröffentlichung auch im Sinne eines vorausschauenden Bestandsaufbaus bejaht werden, auch wenn keine Fragestellung vorliegt. Generell aber erscheint dieses Kriterium sinnvoll zu sein, um Forschungsdaten von anderen Daten wie staatlichen statistischen Erhebungen zu unterscheiden, die wissenschaftlich-methodisch vorgehen, aber keine genuine Fragestellung zu Grunde liegen haben.
- 19 Die Staatlichen Archive Bayern sprechen beispielsweise davon, dass von 47 Millionen Originalen knapp 166.000 Archivalien mit über 5,4 Millionen Seiten digitalisiert worden sind, vgl. Staatliche Archive Bayern. *Digitalisierung* [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://www.gda.bayern.de/fachinformationen/digitalisierung>
- 20 LaMBDa [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://lambda.ios-regensburg.de/>
- 21 osmikon.search [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://www.osmikon.de/suchen>
- 22 Im Archivportal-D ist digitales Archivgut aus ganz Deutschland nachgewiesen, vgl. Archivportal-D [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://www.archivportal-d.de/>
- 23 Discuss Data. *Archival Materials* [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://discuss-data.net/categories/archival-materials/>
- 24 NFDI for memory. *Task Area 3: Data Services* [Zugriff am: 23. Oktober 2024]. Verfügbar unter: <https://4memory.de/task-areas/task-area-3-data-services/#collapse3>



Verfasser

Dr. Arnošt Štanzel, Fachinformationsdienst Ost-, Ostmittel- und Südosteuropa – Koordination Fachportal und Forschungsdaten, Osteuropaabteilung, Bayerische Staatsbibliothek, Ludwigstraße 16, 80539 München, Telefon +49 89 28638-2832, arnost.stanzel@bsb-muenchen.de
Foto: privat