Schwerpunktinitiative „Digitale Information“ der
Allianz der deutschen Wissenschaftsorganisationen

Zur Bedeutung des Konzepts ›Digitale Sammlung‹

Ein Diskussionspapier der Arbeitsgruppe “Digitale Sammlungen” (AG 3)[1]

[Oktober 2020]

Die Beschleunigung und Ausweitung der Bereitstellung digitaler oder in maschinenlesebare Form transformierter (*datafication*) Daten und Dokumente führt in den Wissenschaften heute zu einem exponentiellen Wachstum von Forschungsdaten, die für sehr unterschiedliche Anwendungsszenarien und mit unterschiedlichen Qualitätsansprüchen erzeugt, gespeichert, publiziert und genutzt werden. Um die damit einhergehenden Bedarfe zu adressieren, wird derzeit der Auf- und Ausbau von Forschungsdateninfrastrukturen auf regionaler, nationaler und internationaler Ebene ebenso intensiv wie extensiv betrieben. Eine Herausforderung in diesem Prozess ergibt sich aus der Tatsache, dass der Begriff Forschungsdaten hinsichtlich seiner Bedeutung und seines Umfanges wenig bestimmt ist und sich aus den vergangenen, gegenwärtigen und antizipierten Anwendungsszenarien der jeweiligen Daten ergibt. Zugleich wird zunehmend gefordert, Forschungsdaten nach Maßgabe der FAIR Prinzipien nicht nur besser zugänglich zu machen, sondern auch Infrastrukturen zu etablieren, um ihre Erzeugung, Aggregation Speicherung, Publikation und Pflege in professionelle Hände zu legen. Das Konzept der ›Digitalen Sammlung‹ kann hier insofern eine Steuerungsfunktion übernehmen, als es konzeptionell die Möglichkeit eröffnet, Forschungsdaten unter dem Gesichtspunkt von Einheitlichkeit, intentionalem Aufbau und institutioneller Pflege zu denken. In Anlehnung an das Konzept der ›Sammlung‹, das in Einrichtungen wie Bibliotheken, Archiven, Museen seit Jahrhunderten entwickelt wurde und wird, dient das Konzept der ›Digitalen Sammlung‹ in diesem Sinne zur Priorisierung spezifischer Arten von Forschungsdaten, die es „wert“ sind, in eine Sammlung aufgenommen zu werden. Zugleich markiert die Digitalität einer solchen Sammlung aber auch Unterschiede zu herkömmlichen Sammlungen. Die Nutzung einer digitalen Sammlung ist nicht an Zeit und Raum gebunden, digitale Objekte sind leicht kopier- und transferierbar. Auszeichnende Merkmale wie sammlungstypische Unikalität entfallen. Vor allem bieten sich erweiterte analytische Möglichkeiten und eine verbesserte Anschlussfähigkeit an Forschung, die mit digitalen Methoden arbeitet.

---

[1] Autorinnen und Autoren: Mitglieder der AG 03„Digitale Sammlungen“

Die Onlineversion dieser Publikation finden Sie unter: https://doi.org/10.2312/allianzoa.040

›Forschungsdaten‹ und ›Digitale Sammlungen‹

Bereits heute hat sich die Einsicht durchgesetzt, dass angesichts der Unübersichtlichkeit vorhandener digitaler Datenressourcen, der Unklarheit über Zugriffs- und Nutzungsrechte und der heterogenen Datenqualität eine systematische Erstellung, Zusammenführung, Aufbereitung und Qualitätssicherung leicht zugänglicher und fachlich passgenauer Ressourcen unverzichtbar ist. In diesem Sinne bilden Datensammlungen als qualitätsgeprüfte kuratierte Forschungsdaten die Voraussetzung für die Evaluation, Entwicklung und Anwendung von Analysemethoden mit großem wissenschaftlichem Potential, z.B. Text- und Datamining (TDM) oder Maschinelles Lernen (ML) bzw. Künstliche Intelligenz (KI). Auf der Grundlage digitaler Sammlungen können somit innovative Forschungsfelder und Methoden entstehen, aber auch neue Wirtschaftsfelder erschlossen werden.

Während Forschungsdaten im Allgemeinen auch niederschwelligen Anforderungen genügen können, etwa weil sie projektgebunden für einen spezifischen Forschungszusammenhang von Bedeutung sind, besteht im Fall von ›Digitalen Sammlungen‹ eine durch Referenzcharakter oder Repräsentativität begründete übergreifende Relevanz, der durch höherschwellige Qualitätsstandards und institutionelle Fürsorgepflicht Rechnung zu tragen ist. In den meisten Fällen sind ›Digitale Sammlungen‹ daher nicht lediglich opportunistisch zusammengestellt, sondern vollständig, repräsentativ oder balanciert. Im Unterschied zum Begriff ›Forschungsdaten‹ beschreibt der enger gefasste Begriff der ›Digitalen Sammlung‹ eine nach einheitlichen disziplinären, thematischen oder formalen Kriterien intendierte Zusammenführung von Datenressourcen sowie die damit verbundene Absicht, den Zugang langfristig, auch unabhängig vom ursprünglichen Entstehungs- und Nutzungszusammenhang in professionellen, möglichst zentralen Organisationsstrukturen zu garantieren, um deren Auffindbarkeit und Nachnutzbarkeit sicherzustellen.

Ob Forschungsdaten zu einer Sammlung zusammengeführt werden, ist das Ergebnis von Aushandlungsprozessen zwischen Forschungs-*Communities* und datenhaltenden Infrastruktureinrichtungen. Diese disziplinär oder thematisch zu organisierenden Aushandlungsprozesse können von dem tradierten und bewährten Konzept der ›Sammlung‹ und der allgemeinen Diskussion über Forschungsdaten profitieren, indem aus ihnen Anforderungen für den Aufbau und Pflege ›Digitaler Sammlungen‹ abgeleitet werden.

›Digitale Sammlungen‹: Relevanz und Qualität

›Digitale Sammlungen‹ zeichnen sich in Hinblick auf ihre Relevanz für die jeweilige Forschungscommunity vor allem auf Grund dreier Aspekte aus: a) ihres *Provenienz*zusammenhangs, b) ihres *Referenz*charakters und/oder c) ihres Erkenntnis*potenzials.* D.h. sie sind durch ihre Entstehung in besonderen Zusammenhängen gekennzeichnet, durch die besonders intensive und/oder extensive Referenz der Forschung auf eine ›Digitale Sammlung‹ und/oder durch die Möglichkeit, besonders innovative Methoden oder Forschungsfragen an ihr zu entwickeln bzw. zu diskutieren.

Die Relevanz einer ›Digitalen Sammlung‹ geht mit den hohen Qualitätsanforderungen einher, die sie erfüllen muss. Neben einer an den allgemeinen FAIR-Kriterien sowie an disziplin- oder

domänenspezifischen Standards orientierten Datenstruktur und -qualität zeichnen sich ›Digitale Sammlungen‹ insgesamt durch eine hohe Konsistenz und Homogenität aus. Sie sind umfänglich durch sammlungsspezifische Metadaten beschrieben, die Auskunft über Art und Umfang der Sammlung geben.

Auch wenn aus verschiedenen Gründen der Zugang zu digitalen Sammlungen beschränkt sein kann, sollte dort, wo es datenschutz- und urheberrechtlich möglich ist, ›Digitale Sammlungen‹ für die Forschung dem Open-Science-Gedanken verpflichtet sein, sodass sie ungehindert referenziert, aggregiert, selektiert, rearrangiert oder homogenisiert werden können. Offenheit bedeutet dabei auch, dass entsprechende Schnittstellen für den Austausch oder Downloadmöglichkeiten in nachnutzbaren Formaten zur Verfügung stehen. Zugleich ist sicherzustellen, dass zum Zwecke der wissenschaftlichen Nachvollziehbarkeit jederzeit die Genese einer ›Digitalen Sammlung‹ (ihre Provenienz) eindeutig ist. Dabei ist der „Ort“ (Hosting) einer Sammlung nur noch als Indiz eines Provenienzzusammenhangs relevant und kann zumindest vom Prinzip her als physischer Ort weitgehend vernachlässigt werden. Ein Qualitätsmerkmal für Datensammlungen ist nicht mehr nur ihre freie Zugänglichkeit, sondern, wo datenschutz- und urheberrechtlich möglich, ihre bequeme Nutzbarkeit unter Lizenzen, die Aggregationsprozesse fördern (z.B. CC0).

Dynamik und Volatilität ›Digitaler Sammlungen‹

Auch wenn ›Digitale Sammlungen‹ in der Regel aus spezifischen, disziplinären Anforderungen oder Sammeltätigkeiten heraus entstehen, gewinnt ihre inter- und transdisziplinäre Nutzbarkeit etwa in den Data Sciences zunehmend an Bedeutung. In diesem Sinne muss auch das Konzept der ›Digitalen Sammlung‹ inter- und transdisziplinär ausgerichtet sein. Zu entwickeln ist eine fachliche und formale Klassifikation, anhand derer die typologische Breite von ›Digitalen Sammlungen‹ präziser beschrieben werden kann und die es zugleich ermöglicht, einzelne Sammlungen zu kontextualisieren sowie mehrere Sammlungen zu aggregieren.

Als besondere Herausforderung erweist sich die dynamische Natur von ›Digitalen Sammlungen‹. Anders als analoge Sammlungen können die Sammlungsgegenstände fortlaufend verändert und in ihrer Zusammensetzung neu kombiniert und geordnet werden. Da leicht zu kopieren, bestehen bei entsprechenden offenen Lizenzregelungen anders als bei analogen Sammlungen keine Einschränkungen hinsichtlich der Möglichkeit, Daten zu verteilen und redundant an verschiedenen physischen Orten zugänglich zu machen, sofern ihre Integrität über die verschiedenen Kopien hinweg sichergestellt ist. Daraus ergeben sich jedoch hohe Anforderungen an Versionierung und Synchronisierung von ›Digitalen Sammlungen‹, die letztlich nur von professionellen Einrichtungen oder Organisationen, die sich der Kuratierung solcher Sammlungen widmen, geleistet werden können.

Sowohl bei ihrer Erzeugung als auch bei ihrer Speicherung und Nutzung sollten Anbieter ›Digitaler Sammlungen‹ für ein hohes Maß an Kooperation offen sein: Forschung und Datenkuration müssen Hand in Hand gehen. Die Kuratierung von ›Digitalen Sammlungen‹ erfordert ein eigenes, weiter zu professionalisierendes Set von Kompetenzen im Schnittfeld zwischen allgemeiner Datenkuratierung

und domänenspezifischer Expertise. Zugleich sollten bestehende Infrastruktureinrichtungen wie Bibliotheken, Archive, Museen und Datenzentren ertüchtigt und mit Blick auf die zunehmende Entstehung und Nutzung digitaler Sammlungen insbesondere hinsichtlich der Etablierung geeigneter Infrastrukturen für eine Nationale Forschungsdateninfrastruktur (NFDI) weiterentwickelt werden.

Aktivitäten der AG ›Digitale Sammlungen‹ in der Allianz-Initiative

Auf der Grundlage der skizzierten Bedarfe und Herausforderungen möchte die AG ›Digitale Sammlungen‹ in der Schwerpunktinitiative ›Digitale Information‹ der Allianz der deutschen Wissenschaftsorganisationen über dieses Diskussionspapier hinaus eine umfangreichere Darstellung verfassen, um Orientierung durch Zusammenfassung, Systematisierung und Reflektion der aktuellen Diskussion zum Thema zu bieten. Sie richtet sich einerseits an Einrichtungen und Personen, die ›Digitale Sammlungen‹ aufbauen, verwalten und kuratieren, andererseits aber auch an alle, die Digitale Sammlungen nutzen, begutachten, beforschen oder ihren Aufbau und ihre Kuratierung finanzieren und sich dazu über Grundbegriffe und gängige Praxis verständigen und informieren wollen. Neben der Einführung ins jeweilige Thema enthält sie Hinweise zu wichtigen Akteuren, Best-Practice-Beispiele, Evaluationskriterien, Hinweise auf Datenstandards und ausgewählte Literatur.