Digitale Sammlungen

Eine Handreichung der Arbeitsgruppe »Digitale Sammlungen« (AG 3) der Allianz der deutschen Wissenschaftsorganisationen (Februar 2022)

unter Mitwirkung von

Mark Azzam, Philipp Cimiano, Alexander Geyken, Hans-Peter Hahn, Gerhard Heyer, Jana Hoffmann, Christian Langenbach, Reiner Mauer, Margit Palzenberger, Torsten Roeder, Patrick Sahle, Thomas Stäcker, Christian Thomas, Peer Trilcke, Nina Leonie Weisweiler, Andrea Wuchner, Thomas Zastrow

Die Onlineversion dieser Publikation finden Sie unter:
https://doi.org/10.48440/allianzoa.043.

Inhaltsverzeichnis

1 Einleitung: ›Digitale Sammlungen‹

Bereits seit Mitte der 1990er Jahre werden bedeutende digitale Sammlungen aufgebaut und online bereitgestellt. Die Vielfalt der laufend hinzukommenden Sammlungen sowie das hohe Tempo der digitalen Transformation aller Forschungsprozesse stellt die bestehenden Informations- und Wissensinfrastrukturen vor große Herausforderungen. In den letzten Jahren erschienen viele Publikationen von Daten, Objekten und Dokumenten ausschließlich in digitaler Form. Deren Erschließung und die Sicherstellung ihrer langfristigen Zugänglichkeit stoßen mit den ›traditionellen‹ Verfahren an Grenzen, sodass neue Lösungen mittels automatisierter Verfahren oder auch der Einsatz von künstlicher Intelligenz geboten scheinen.

Die interdisziplinär zusammengesetzte Arbeitsgruppe ›Digitale Sammlungen‹ hat im Zeitraum 2018 bis 2021 die Bedeutung des Konzepts ›Digitale Sammlungen‹ untersucht und neben einem Kurzpapier »Zur Bedeutung des Konzepts ›Digitale Sammlungen‹: Ein Diskussionspapier der Arbeitsgruppe ›Digitale Sammlungen‹ (AG 3)«[1] die vorliegende Handreichung erarbeitet, die einen Überblick geben soll über Art, Funktion und Nutzen digitaler Sammlungen.

Die Handreichung soll Orientierung in einem zunehmend unübersichtlichen Feld geben, indem sie über Konzepte und exemplarische Formen, Metadaten für digitale Sammlungen, Publikationswege, Zugriffsmöglichkeiten und Schnittstellen sowie über Verfahren der Auswertung von Daten- und Dokumentensammlungen, z. B. mittels Text- und Datamining, informiert, aber sich auch Fragen der Qualitätssicherung, der rechtlichen Rahmenbedingungen und der Organisation und dem Management digitaler Sammlungen zuwendet.

Mit abschließenden Empfehlungen, die sich aus der Behandlung dieser Themenkomplexe ergeben, richtet sich die Handreichung einerseits an Einrichtungen und Personen, die digitale Sammlungen aufbauen, verwalten und kuratieren. Sie wendet sich aber auch andererseits an alle, die digitale Sammlungen nutzen, begutachten, beforschen oder deren Aufbau und deren Kuratierung finanzieren und sich dazu über Grundbegriffe sowie über die gängige Praxis verständigen und informieren wollen.

Die Resultate der vorliegenden Handreichung ›Digitale Sammlungen‹ können sowohl in ihren deskriptiven Aussagen (über die Faktenlage, den Stand der Forschung etc.) als auch in ihren präskriptiven Bestandteilen (Empfehlungen) nur ein Ausschnitt sein. Sie spiegeln nicht sämtliche Erkenntnisse aller Fächer und Disziplinen wieder, sind relational und kontextabhängig und können auch nur einen kleinen Ausschnitt der zu diesem Thema erschienen Literatur berücksichtigen.[2] Wenngleich sie in ihrer Form nur relative Begründungen und Rechtfertigungen erlauben, hoffen die Beiträgerinnen und Beiträger für die wichtigsten Fragen, die mit ›Digitalen Sammlungen‹ zu tun haben, einen Beitrag zu deren Verständnis und Nutzen zu liefern.

Digitale Sammlungen gehören zu einem dynamischen, zwischen Forschung und Forschungsinfrastrukturen angesiedelten Feld, das für die Wissenschaft in nahezu allen Bereichen inter- und transdisziplinär an Bedeutung gewinnt und auch im Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI) eine zentrale Rolle spielt. Digitale Sammlungen ergänzen nicht nur klassische Sammlungen in Archiven, Bibliotheken, Museen und bei anderen Informationsdienstleistern, sondern treten in manchen Fällen im Sinne einer primären Informationsressource für die Wissenschaft auch an die Stelle der traditionellen analogen Sammlung. Dem liegt die Einsicht zugrunde, dass Forschung in Zukunft weit stärker mit digitalen Methoden arbeiten und auf die Auffindbarkeit von Daten, Dokumenten und Objekten über digitale Zugänge angewiesen sein wird. Der Begriff der ›digitalen Sammlung‹ kann hierbei eine Steuerungsfunktion übernehmen, um die spezifischen Bedarfe der Wissenschaft an die Nachnutzbarkeit und Nachhaltigkeit von Forschungsdaten gemäß der FAIR-Prinzipien[3] zu adressieren. Dabei wurde folgende Definition zugrunde gelegt:

> *Digitale Sammlungen (digital collections)*[4] *sind organisierte Daten- und Dokumentbestände, die im Forschungsprozess entstehen und Grundlage für anschließende weitere, intendierte bzw. im Entstehungsprozess noch nicht avisierte Forschung (Nachnutzung) sind bzw. sein können. Sie unterscheiden sich von Forschungsdaten im Allgemeinen durch mehrere Aspekte: Sie sind regelgeleitete, intentionale Zusammenstellungen, die als Produkte bzw. Ergebnisse einer Sammlungstätigkeit mit einem dokumentierten Qua-*

1 Vgl. Schwerpunktinitiative ›Digitale Information‹ der Allianz der deutschen Wissenschaftsorganisationen (2020): Zur Bedeutung des Konzepts ›Digitale Sammlung‹: Ein Diskussionspapier der Arbeitsgruppe ›Digitale Sammlungen‹ (AG 3). Allianz der deutschen Wissenschaftsorganisationen. https://doi.org/10.2312/allianzoa.040.

2 Vgl. z. B. den konzisen Überblick von Rice, R., & Southall, J. (2016). *The Data Librarian's Handbook*. Facet Publishing.

3 Vgl. *FAIR principles* (2021, May 27). GO FAIR. https://www.go-fair.org/fair-principles/.

4 Ein ähnliches Konzept findet sich im NSF Report »Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century« mit dem Begriff der »long-lived digital data collections«. Vgl. National Science Board (Pre-Publication Draft) (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (NSB-05-40), https://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf, S. 4.

litätsanspruch auf eine Nachnutzung ausgerichtet sind. Sie haben eine eigene Identität (im Objekttyp ›Sammlung‹), werden durch Metadaten beschrieben, von vertrauenswürdigen Infrastrukturen gesichert und zugänglich gemacht, können fortlaufend verändert (annotiert, angereichert), dauerhaft (auch in Versionen) referenziert und auf vielfältige Weise nachgenutzt sowie hinsichtlich ihrer Qualitäten evaluiert werden.

Digitale Sammlungen bestehen aus Daten und Metadaten. Der Datenbegriff ist ein Relationsbegriff und darin verwandt mit dem Informationsbegriff. Man kann argumentieren, dass allgemein durch jede Form von kontextualisierender Nutzung oder speziell durch Auswertung mittels eines Algorithmus ein Datum zu einer Information wird.[5] Daher wird das Suchen in digitalen Sammlungen passend im Englischen als *information retrieval* bezeichnet. Der Begriff ›Daten‹, der auch dieser Definition zugrunde liegt, ist als solcher allerdings schwer bestimmbar. Am ehesten kann man mit Floridi Daten als *lack of uniformity* beschreiben.[6] In Gebrauch sind eine Vielfalt verwandter, oft nicht trennscharfer Begriffe wie ›Dokument‹, ›Text‹, ›Objekt‹, ›Medium‹, ›Information‹, ›Software‹, ›Workflows‹ oder ›Hardwarekonfigurationen‹,[7] die nach ihren jeweiligen Verwendungskontexten bestimmt sind. All diese Begriffe werden hier der Einfachheit halber mit dem Begriff ›Datum‹ gleichgesetzt. In diesem allgemeinen Sinne sind digitale Sammlungen also Sammlungen von Daten; Sammlungen von Daten sind wiederum eine besondere Form von Forschungsdaten, die technisch gesehen als Datensets vorliegen (vgl. z. B. die Definition von VoID[8]). Auch wenn Sammlungen im Prinzip gegenstandsneutral sind und allein durch Genese, Provenienz oder eine Sammlungsintention bestimmt sein können – z. B. eine Sammlung von Gegenständen, die einer berühmten Persönlichkeit gehört haben –, liegt doch der Fokus der vorliegenden Handreichung auf Sammlungen, deren Gegenstände eine gewisse Gleichartigkeit aufweisen, also Sammlungen, die aufgrund von Eigenschaften ihrer Teile aggregiert worden sind.

Digitale Sammlungen sind ortsungebunden, volatil, verlustfrei kopierbar und lassen sich durch Aggregation oder Aufteilung nach den jeweiligen Forschungsfragen und Objekttypen neu bilden. Daraus ergeben sich eine Reihe von Fragestellungen, etwa was die Qualität, Provenienz bzw. Zuschreibung, Referenzier- und Zitierbarkeit, die Versionierung, Vereinheitlichung und Normierung oder die Möglichkeit der Zusammensetzung, des Aufbaus, der langfristigen Zugänglichkeit und Archivierung anlangt.

2 Arten und Formen

2.1 ›Digitale Sammlungen‹ – Konzepte und exemplarische Formen

Digitale Sammlungen können ihrer Form und ihrer Zusammensetzung nach unterschieden werden. So lassen sich Sammlungen z. B. nach ihrer Medienform differenzieren als Sammlungen von Texten, Bildern, Tönen, Objekten oder allgemein von numerischen, beschreibenden, gemessenen oder erhobenen Daten. Sammlungen erlauben aber auch andere Einteilungen: etwa in frei oder beschränkt zugängliche Sammlungen, nach ihrem Zweck als Referenzsammlungen oder nach ihrem Strukturierungsgrad (strukturiert, unstrukturiert, semistrukturiert). Sammlungen können spezifische technische Datentypen (z. B. differenziert nach *MIME types*) enthalten, fachlich (Natur-, Technik- oder Geisteswissenschaften) oder spartenspezifisch (Bibliothek, Archiv, Museum, Forschungseinrichtung) differenziert sein. Unterscheiden lassen sie sich auch nach der Persistenz (für einen *ad-hoc*-Zweck hergestellt oder auf Dauer angelegt) und verschiedenen Lebenszyklen der Sammlung (im

5 Vgl. Floridi, L. (2013). *The philosophy of information*. OUP. S. 83 f.

6 Vgl. Floridi, L. (2013). *The philosophy of information*. OUP. S. 85. Vgl. a. die offene Charakterisierung der NSF (2005): »The term data is used in this report to refer to any information that can be stored in digital form, including text, numbers, images, video or movies, audio, software, algorithms, equations, animations, models, simulations, etc.« National Science Board (Pre-Publication Draft) (2005). *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century* (NSB-05-40). https://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf.

7 Vgl. die Formulierung in den FAIR Prinzipien *FAIR principles* (2021, May 27). GO FAIR. https://www.go-fair.org/fair-principles/ oder EOSC IF https://op.europa.eu/s/uD3F.

8 Vgl. Klyne, G., & Carroll, J. (2004). *Resource Description Framework (RDF): Concepts and Abstract Syntax, W3C Recommendation*. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210. Vgl. Alexander, K., Cyganiak, R., Hausenblas, M., & Zhao, J. (2011). *Describing Linked Datasets with the VoID Vocabulary*. W3C. http://www.w3.org/TR/2011/NOTE-void-20110303/: »The fundamental concept of VoID is the ***dataset***. A dataset is a set of RDF triples that are published, maintained or aggregated by a single provider. Unlike *RDF graphs*, which are purely mathematical constructs, the term *dataset* has a social dimension: we think of a dataset as a *meaningful* collection of triples, that deal with a certain topic, originate from a certain source or process, are hosted on a certain server, or are aggregated by a certain custodian. Also, typically a dataset is accessible on the Web, for example through resolvable HTTP URIs or through a SPARQL endpoint, and it contains sufficiently many triples that there is benefit in providing a concise summary.«

Aufbau, abgeschlossen etc.), mit jeweiliger Referenzierbarkeit oder Versionierung. Von besonderem Interesse sind digitale Sammlungen vor allem dann, wenn sie sich zu größeren Einheiten bzw. neuen umfangreicheren Datensammlungen zusammenfassen oder sich interessante Teilsammlungen extrahieren lassen. Das kann durch Aggregieren der Daten, aber auch durch Vernetzen passieren.

Auch wenn es keine allgemeingültige oder abgeschlossene Liste von Merkmalen, die sammlungskonstitutiv sind, gibt und auch nicht geben kann, existieren doch einige Sammlungsformen und -typen, die eine größere Verbreitung aufweisen. Von diesen seien die Textsammlung, auch unter dem Begriff des ›Korpus‹ bekannt, die Bildsammlung (stille und bewegte Bilder) oder die Objektsammlung (mit der Repräsentation von Objekten durch Beschreibungsdaten) genannt. Im Übrigen finden sich zahllose Sonderformen, die in mehr oder weniger stark strukturierter Form vorliegen, z. B. Sammlungen von Vermessungsdaten, von Messdaten aus Instrumenten und Sensoren, Satellitendaten, Daten von Spektogrammen, sozialwissenschaftliche Daten, Verwaltungsdaten usw.

Nachstehend findet sich eine beliebig erweiterbare Tabelle von exemplarischen digitalen Sammlungen.

Name	**Objekttyp**	**Institution**	**Kommentar/Besonderheit**
Collection des Rijksmuseums Amsterdam	Kulturobjekte, Bilder	Rijksmuseum Amsterdam	Digitalisierte Kulturobjekte
Deutsches Textarchiv (DTA)	Textdaten	Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)	Korpus deutschsprachiger Texte
Copernicus Data and Exploitation Platform – Deutschland (CODE-DE)	Messdaten	Deutsches Zentrum für Luft- und Raumfahrt (DLR)	Satellitendaten in der Erdbeobachtung
EOWEB® GeoPortal (EGP)	Messdaten	Deutsches Zentrum für Luft- und Raumfahrt (DLR)	Multimissions-Erdbeobachtungsdatenportal des DLR
Umfragedaten aus der DDR und den neuen Bundesländern	Statistische Daten	GESIS	Studien aus der empirischen Sozialforschung der DDR, der staatlichen Hörer- und Zuschauerforschung sowie Studien, die aus der wissenschaftlichen Begleitung des sozialen Wandels in den neuen Bundesländern entstanden sind
TweetsKB	Statistische Daten	GESIS	TweetsKB ist eine öffentlich zugängliches umfangreiches RDF Korpus anonymisierter Daten von annotierten Tweets.
Parliaments and governments database (ParlGov)	Statistische Daten	ParlGov Project	ParlGov ist eine Platform für politische Wissenschaften und enthält Informationen zu allen EU- und den meisten OECD-Demokratien (37 Länder). Die Datenbank verknüpft ca. 1700 Partien, 990 Wahlen (9200 Ergebnisse) und 1500 Kabinette (3800 Parteien)
Drama Corpora Plattform (DraCor)	Textdaten	Higher School of Economics, Moscow, Universität Potsdam	Aggregation, Anreicherung und Homogenisierung zahlreicher nationalsprachlicher Dramenkorpora mit Analyse-Plugins/-Apps, offener API, SPARQL-Endpoint, Python- und R-Wrapper (in Entwicklung), zugleich abgelegt auf GitHub (als Ganzes ›kopierbar‹, aber auch einzelne Korpora)

SABIO-RK	Publizierte Daten / Messdaten	HITS gGmbH	Kuratierte Datenbank, die strukturierte Informationen zu biochemischen Reaktionen und deren Kinetik enthält. Dies umfasst alle verfügbaren kinetischen Parameter zusammen mit den entsprechenden Geschwindigkeitsgleichungen sowie das Kinetikgesetz und die Parametertypen sowie die experimentellen und Umgebungsbedingungen, unter denen die kinetischen Daten bestimmt wurden
GOVDATA	Daten der öffentlichen Verwaltung	Geschäfts- und Koordinierungsstelle GovData bei der Freien und Hansestadt Hamburg	Open Government Portal
ImageNet	Bilddaten	Stanford University, Princeton University	ImageNet ist eine Image-Datenbank, die nach der Hierarchie organisiert ist, in der jeder hierarchische Knoten durch zahlreiche Images abgebildet wird.
Korpora des DWDS	Textdaten	Berlin-Brandenburgische Akademie der Wissenschaften	Der deutsche Wortschatz von 1600 bis heute
Koordinierungsstelle für wissenschaftliche Universitätssammlungen in Deutschland	Bild- und Objektdaten	Hermann-von-Helmholtz-Zentrum für Kulturtechnik der Humboldt-Universität zu Berlin	Nachweis von Universitätssammlungen
Biodiversity Heritage Library	Textdaten	BHL Consortium	Die Biodiversity Heritage Library verbessert als Teil einer globalen Biodiversitätscommunity die methodische Forschung durch die kollaborative und freie Zugänglichmachung von Literatur zur Biodiversität.

Die Liste gibt einen exemplarischen Ausschnitt aus dem bestehenden Angebot an digitalen Sammlungen und soll deren Vielfalt verdeutlichen. Einen guten Überblick über bestehende Forschungsdatenrepositorien mit Zugriff auf digitale Sammlungen bietet re3data.[9] Auch Forschungsinstitutionen und -projekte wie die Forschungsrepositorien mit Beteiligung der Helmholtz-Gemeinschaft oder auch die Fachinformationsdienste der Deutschen Forschungsgemeinschaft widmen sich in zunehmendem Maße dem Aufbau und Nachweis von digitalen Sammlungen.

2.2 Metadaten für digitale Sammlungen

Metadaten sind, wörtlich genommen, Daten über Daten. Genauer genommen handelt es sich um deskriptive Daten über (digitale) Objekte. Der Metadaten-Begriff ist relational bzw. zyklisch, da Metadaten auch selbst Daten sein und mit Metadaten beschrieben werden können. Sie beschreiben im Allgemeinen Aspekte der Daten, die entweder eine Eigenschaft des Datenobjektes bezeichnen oder aber weitere Informationen liefern. Sie können insofern das Objekt eindeutig identifizieren und auffindbar machen, aber auch Rahmenbedingungen wie Zugänglichkeit oder Nachnutzbarkeit beschreiben. Eine typische Unterscheidung ist die Aufteilung in deskriptive, strukturelle, administrative und technische Metadaten.[10] Terminologisch werden Metadaten gelegentlich auch

9 Vgl. Karlsruhe Institute of Technology (n. d.). *re3data.org registry of research data repositories*. https://doi.org/10.17616/R3D.

10 So etwa im Modell von METS: Digital Library Federation (2010). <*METS*> *Metadata encoding and transmission standard: primer and reference manual*. https://www.loc.gov/standards/mets/METSPrimer.pdf

als Annotationen bezeichnet oder nähern sich bei lockerer Handhabung auch dem Begriff der Dokumentation[11] oder der Mikro- oder Nanopublikation[12] an.

Im Kontext digitaler Sammlungen beschreiben Metadaten einerseits auf Ebene der gesamten digitalen Sammlung die Art und Herkunft der Sammlung oder wie die Sammlung zugänglich ist, andererseits auf der Ebene der enthaltenen Datenobjekte deren Beschaffenheit und Struktur. Die Metadaten werden idealerweise so formuliert, dass sie in einer strukturierten Form ›maschinenlesbar‹ sind, d. h. von einer Maschine automatisiert interpretiert bzw. verarbeitet werden können, etwa um Sammlungen differenziert suchen und durchsuchen zu können. Metadatenformate sind insofern unter zwei Gesichtspunkten zu betrachten: Einerseits auf der konzeptionellen Ebene des Inhaltes bzw. Begriffs, andererseits auf der technischen Ebene ihrer ›Serialisierung‹, also wie sie maschinenlesbar gemacht werden. So kann der abstrakte Begriff des Autors oder Autorin eines Werkes mittels XML, JSON oder Turtle formuliert (›serialisiert‹) werden.

Metadaten spielen eine wichtige Rolle bei der Umsetzung der FAIR-Prinzipien. Hinweise zu Metadaten sind in allen vier Bereichen der FAIR-Prinzipien enthalten. So fordert z. B. das Prinzip der ›Findability‹, dass Daten durch ›reichhaltige‹ Metadaten beschrieben werden[13] oder dass zur Nachnutzung (*reuse*) Metadaten zu Lizenzbestimmungen, Copyright und Herkunft beigefügt werden, um Daten durch eine Menge von korrekten und relevanten Attributen (»accurate and relevant attributes«) zu beschreiben.[14] Das bezieht sich insbesondere auf rechtliche Aspekte der Datennutzung,[15] auf die Beschreibung der Herkunft der Daten (*provenance*)[16] sowie auf die Nutzung relevanter Community-Standards bei der Beschreibung der Metadaten.[17]

Neben den verschiedenen community- und disziplinspezifischen Metadatenformaten, die hier nicht im Einzelnen behandelt werden können, sind disziplinenunabhängige und sammlungsspezifische Metadatenformate bisher vergleichsweise wenig in Gebrauch, werden aber mit der zunehmenden Bedeutung digitaler Sammlungen für die Forschung immer wichtiger. Zu nennen ist der schon ältere Standard des *Dublin Core Collection Description AP* (DCCD)[18], der zwar in die *DFG-Praxisregeln »Digitalisierung«* [19] als Empfehlung Eingang fand, sich aber bislang nicht hat durchsetzen können. Neuer ist der vom W3C verabschiedete *DCAT*-Standard, der mittlerweile in einer erweiterten 3. Version vorliegt.[20] Er basiert auf dem Resource Description Framework (RDF) und kommt z. B. im Portal für offene Verwaltungsdaten zum Einsatz.[21] Daneben sind auch *schema.org*[22] und der *DataCite*-Standard zu nennen, der sich als Begleitstandard zur Vergabe einer DOI zur eindeutigen Identifikation eines Objektes etablieren konnte und von allen Einrichtungen, die DOIs für Datasets vergeben, genutzt werden muss. Allen diesen Standards gemein ist eine zumindest grobe Orientierung an *Dublin Core*. Daneben existieren aber noch zahlreiche disziplin- und spartenspezifische Metadatenstandards, die ebenfalls zur Beschreibung von Sammlungen verwendet werden können. Dazu gehört etwa *METS*[23], ein Wrapperformat, das vor allem in Bibliotheken zur Anwendung kommt, oder *EAD*[24], das im Archivkontext gebräuchlich ist. Mit dem aus dem Museumsumfeld stammenden und mit *CIDOC-CRM* kompatiblen *LIDO*[25] können ebenfalls Sammlungen beschrieben werden.

---

11 Vgl. Rice, R., & Southall, J. (2016). *The data librarian's handbook*. Facet Publishing. »Metadata and documentation become synonymous«, S. 24.

12 Groth, P., Gibson, A., & Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use*, *30* (1–2), 51–56. https://doi.org/10.3233/ISU-2010-0613.

13 Vgl. Findability-Regel F2: »Data are described with rich metadata«, *FAIR principles* (2021, May 27). GO FAIR. https://www.go-fair.org/fair-principles/.

14 Vgl. Reusability-Regel R1, ebda.

15 Vgl. Reusability-Regel R1.1., ebda.

16 Vgl. Reusability-Regel R1.2., ebda.

17 Vgl. Empfehlung in Reusability-Regel R1.3., ebda.

18 Vgl. hierzu die Dokumentation der »DCMI Collection Description Community«, Dublin Core Metadata Initiative, 2001ff.; Dublin Core Collection Description Task Group (2007, September 3). *Dublin Core™ collection description application profile*. https://www.dublincore.org/specifications/dublin-core/collection-description/collection-application-profile/.

19 Vgl. Deutsche Forschungsgemeinschaft (2016). *DFG Praxisregeln »Digitalisierung«*. https://www.dfg.de/formulare/12_151/12_151_de.pdf.

20 W3C (Working Draft) (2021). *Data Catalog Vocabulary (DCAT) - Version 3*. https://www.w3.org/TR/2021/WD-vocab-dcat-3-20210504/.

21 Vgl. Geschäfts- und Koordinierungsstelle GovData (n. d.). *DCAT-AP.de*. https://www.dcat-ap.de/.

22 Vgl. Schema.org (n. d.). *Documentation*. https://schema.org/docs/documents.html.

23 Vgl. METS Board & Network Development and MARC Standards Office (2021). *Metadata encoding and transmission standard (METS)*. https://www.loc.gov/standards/mets/.

24 Vgl. Technical Subcommittee for Encoded Archival Standards of the Society of American Archivists, & Library of Congress (2021). *EAD: Encoded archival description official site*. https://www.loc.gov/ead/.

25 Vgl. ICOM-CIDOC LIDO working group (2010). *Specification*. https://cidoc.mini.icom.museum/working-groups/lido/lido-technical/specification/.

Unabhängig von den jeweiligen Vor- und Nachteilen der diversen Formate und Implementierungen scheint eine Grundverständigung auf ein spartenübergreifendes und disziplinunabhängiges Format sinnvoll, um digitale Metadaten zu Sammlungen mit Blick auf ihre Spezifika (Inhalt, Umfang, Veränderung, Zuwachsraten, Verfügbarkeit, Rechte, etc.) auch über Fächer- und Spartengrenzen hinweg austauschen oder einheitlich durchsuchen sowie nachnutzen zu können. Neben *DCCD* ist hier *DCAT* gerade durch seine Empfehlung durch das W3C ein besonders aussichtsreicher Kandidat, zumal es wesentliche Konzepte von *DCCD* integriert und bereits im Bereich governmental data intensiv genutzt und auch im Kontext der European Open Science Cloud (EOSC) bzw. der Initiative GO FAIR[26] aufgegriffen wurde.

3 Angebot und Nutzung

3.1 Publikationswege, Zugriffsmöglichkeiten und Schnittstellen gemäß den FAIR-Prinzipien

3.1.1 Publikation digitaler Sammlungen

Digitale Sammlungen werden auf vielfältige Weise öffentlich bereitgestellt und damit publiziert. Für die Form dieser Publikation haben sich bisher noch keine breit akzeptierten Konventionen entwickelt.

Oft werden Sammlungen von der herstellenden Institution selbst gehostet. A priori unterliegen diese selbst gehosteten digitalen Sammlungen damit unter Umständen keinen Regularien in Hinblick auf Qualität und Nutzbarkeit von Daten und Metadaten. Ihre Auffindbarkeit ist durch die Möglichkeiten generischer Suchmaschinen begrenzt. Die Nachnutzung setzt ggf. sehr spezielle Kenntnisse über die Datensammlung voraus und das Ermitteln relevanter Charakteristika kann vielfach sehr zeitaufwändig sein.

Ein Schritt zur strukturierten Publikation einer solchen Sammlung ist die Indexierung in spezialisierten Suchmaschinen und Verzeichnissen, die für eine Publikation von Sammlungen relevante Metadatenfelder erfassen. Dazu gehören fachübergreifend B2FIND oder Google Dataset Search sowie zahlreiche fachspezifische Repositorien und Datenportale. Verzeichnisse wie das *Registry of Research Data Repositories* (re3data) dienen als Nachweisinstrumente für Forschungsdatenrepositorien und -portale und erleichtern das Monitoring der Infrastrukturlandschaft sowie die Recherche nach geeigneten Orten zur Datenpublikation bzw. -suche. Hersteller und Administratoren von Sammlungen sollten möglichst umfassende Informationen für solche Aggregationsdienste bereitstellen und regelmäßig aktualisieren. Der Aufbau und die Nutzung von übergeordneten Verzeichnissen wird auch von der EOSC »FAIR Working Group« als Best Practice empfohlen.[27]

Einige Anbieter von digitalen Sammlungen publizieren darüber hinaus in Zeitschriftenartikeln eine Beschreibung ihrer Sammlung und erreichen damit hohe Auffindbarkeit und Zitierbarkeit im bestehenden wissenschaftlichen Kommunikationsgefüge. So erscheint z. B. jährlich ein Artikel über die »NCBI GenBank« in der Zeitschrift *Nucleic Acid Research*.[28]

Seit etwa 10 Jahren etablieren sich mehr und mehr dezidierte Data Journals, die sich der Veröffentlichung von Datenpublikationen (*Data Papers*) im Rahmen etablierter wissenschaftlicher Prozesse widmen. Damit kuratieren Verlage und Herausgeber den Aufbau von Sammlungen, die auf Datensätze eines bestimmten Typs oder Fachgebietes verweisen und deren Inhalte im Format wissenschaftlicher Artikel beschrieben werden. Diese Artikel und die dazugehörigen Daten müssen im Normalfall formale und inhaltliche Kriterien erfüllen und werden einem Peer Review unterzogen. Einige dieser Zeitschriften sind bereits sehr erfolgreich, werden im Web of Science gelistet und erzielen hohe Zitationsraten (z. B. *Earth System Science Data*, *GigaScience*, *Scientific Data* oder *Biodiversity Data Journal*).

3.1.2 Zugriffswege und Schnittstellen

Digitale Sammlungen können wie eigenständige digitale Objekte betrachtet werden. Hieraus folgt eine Weise des automatisierten Zugriffs, die sich grob in vier unterschiedliche Operationen einteilen lässt (CRUD):

- Create: Anlegen einer neuen digitalen Sammlung, Festlegen von deren Basiseigenschaften

26 Hier als »example framework« in der Findability-Regel F2 »Data are described with rich metadata«, Vgl. *FAIR principles* (2021, May 27). GO FAIR. https://www.go-fair.org/fair-principles.

27 Vgl. Genova, F., Aronsen, J. M., Beyan, O., Harrower, N., Holl, A., Principe, P., Slavec, A., & Jones, S. (2021) *Recommendations on certifying services required to enable FAIR within EOSC*. EOSC Executive Board Working Group (WG) FAIR Task Force (TF). https://doi.org/10.2777/127253.

28 Vgl. Sayers, E. W., Karsch-Mizrachi, I., Cavanaugh, M., Clark, K., Ostell, J., & Pruitt, K. D. (2018). GenBank. *Nucleic Acids Research*, *47*(1), S. 94–99. https://doi.org/10.1093/nar/gkz956.

- Read: Lesender Zugriff auf Elemente und Eigenschaften einer digitalen Sammlung
- Update: Hinzufügen neuer und bearbeiten bestehender Elemente und Eigenschaften
- Delete: Entfernen bestehender Elemente aus der Sammlung

Je nach konkretem Anwendungsfall müssen diese Zugriffsmöglichkeiten weiter ausdifferenziert werden. Hinzukommen im Vorfeld festzulegende Eigenschaften, wie zum Beispiel Rekursivität: Darf eine digitale Sammlung weitere Sammlungen als Elemente enthalten oder nicht? Alle Zugriffe, sowohl lesender als auch schreibender Natur, müssen durch ein entsprechendes Rechtemanagement vorab geprüft und gegebenenfalls abgelehnt werden.

Im Rahmen der RDA Working Group »Research Data Collections«[29] wurde eine generische, dem CRUD-Prinzip folgende REST-API entwickelt. Diese ist unabhängig von konkreten Softwareprodukten und wurde bereits in der Praxis in mehreren Daten-Projekten bzw. Repositorien implementiert.[30]

Wichtige Schnittstellen, die aber nicht unbedingt sammlungsspezifisch sind, wären neben OAI-PMH und REST auch SPARQL oder im einfachsten Fall Datendumps, wenn man darin eine Schnittstelle sehen will. Bei umfangreicheren Sammlungen ist auch das Angebot kleinerer Samples sinnvoll, um sich vor einem langwierigen Download einen Eindruck von Art und Zusammensetzung der Daten verschaffen zu können. Wenigstens kursorisch hinzuweisen ist am Ende auf die eher philosophische bzw. medientheoretische Dimension des Begriffs der Schnittstelle, der das Verständnis der Nutzung digitaler Sammlungen wesentlich beeinflusst, hier aber nicht weiter behandelt werden kann.[31]

3.2 Auswertung von Daten- und Dokumentensammlungen: KI, Text- und Datamining[32]

Qualitätsgeprüfte digitale Sammlungen bilden als Referenzdaten für die Generierung von Referenzmodellen und die Kalibrierung von Text- und Data Mining-Algorithmen eine wichtige Grundlage für KI- und Sprachtechnologie-Anwendungen. Erste Erfahrungen zeigen, dass nur qualitativ hochwertige Referenzdaten auch qualitativ hochwertige Referenzmodelle erzeugen (eigentlich ein Grundprinzip guter wissenschaftlicher Praxis). Bei den so erzeugten Daten handelt es sich um Basisdaten, welche in der Regel anderen Einrichtungen für ihre weitere Forschung zur Verfügung gestellt werden. In solchen Fällen tragen die Erzeuger der Daten ein hohes Maß an Verantwortung, da fehlerhafte Basisdaten, die z. B. Vorurteile codieren, zu (auch gesellschaftsrelevanten) Systemfehlern führen können. Zum besseren Verständnis dieser zunehmend wichtigen Nutzung von Daten- und Dokumentensammlungen sollen nachfolgend kurz einige zentrale Konzepte erläutert werden.

›Text und Data Mining‹ bezeichnet im Sinne der EU-Richtlinie 2019/790 vom 17. April 2019 »eine Technik für die automatisierte Analyse von Texten und Daten in digitaler Form, mit deren Hilfe Informationen unter anderem – aber nicht ausschließlich – über Muster, Trends und Korrelationen gewonnen werden können«.[33]

Im engeren Sinne werden unter dem Terminus ›Text Mining‹ computergestützte Verfahren für die semantische Analyse von Texten bezeichnet, welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großer Mengen von Texten, unterstützen.[34] Unter dem Begriff des ›Data Mining‹

29 Vgl. Research Data Collections WG (2021, August 4). *Research data collections WG*. RDA. https://www.rd-alliance.org/groups/research-data-collections-wg.html.

30 Vgl. Weigel, T., Almas, B., Baumgardt, F., Zastrow, T., Schwardmann, U., Hellström, M., Quinteros, J., & Fleischer, D. (2017, December 18). *Recommendation on Research Data Collections*. Research Data Collections Working Group. DOI: 10.15497/RDA00022 sowie die »Recommendation on Research Data Collections«. https://github.com/RDACollectionsWG/specification/blob/master/Recommendation%20package/rda-collections-recommendation.pdf. Vgl. auch Data Foundation and Terminology Interest Group (DFT IG) of the Research Data Alliance (RDA. (2018). *Term definitions version 2.0 »Berlin«*. https://smw-rda.esc.rzg.mpg.de/dft-2.0.html.

31 Vgl. z. B. Wirth, S. (2016). *Between Interactivity, Control, and ›Everydayness‹ – Towards a Theory of User Interfaces*. In: *Interface Critique*. Hrsg. v. Hadler, F. & Haupt, J. S. 33: »Building on the idea of the interface as a process, the term ›interface‹ opens up the possibility to think the relation of human user and technology in its operability. The operability not only enables ›interactivity‹ but first and foremost generates the possibility of it by providing a ›ready-to-hand‹ space of possible activity.«

32 Vgl. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, S. 993–1022.Vgl. auch Heyer, G., Quasthoff, U., & Wittig, T. (2011). *Text Mining – Wissensrohstoff Text*. W3L. 3. Nachdruck. Überarbeitete und aktualisierte Neuauflage: Biemann, C., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text: Eine Einführung in das Text Mining*. Springer Campus. Vgl. auch Witten, I., Eibe, F., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

33 EU-Richtlinie 2019/790 vom 17. April 2019. Art. 2, Satz 2. https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32019L0790.

34 Vgl. Heyer, G., Quasthoff, U., & Wittig, T. (2011). *Text Mining – Wissensrohstoff Text*. W3L. 3. Nachdruck. Überarbeitete und aktualisierte Neuauflage: Biemann, C., Heyer, G., & Quasthoff, U. (2022). *Wissensrohstoff Text: Eine Einführung in das Text Mining*.

werden Verfahren aus der Statistik und künstlichen Intelligenz zusammengefasst, welche in großen (Big Data) oder komplex strukturierten Datenbeständen Muster und statistische Zusammenhänge erkennen.[35] Text und Data Mining unterscheiden sich darin, dass beim Text Mining die linguistische Struktur des Textes berücksichtigt wird, beispielsweise die grammatische Kategorie von Wörtern oder die syntaktische Struktur von Sätzen, während beim Data Mining eher Techniken zum Einsatz kommen, die auf strukturellen bzw. tabellarisch geordneten oder anderen nicht-textlichen Formaten aufbauen, wie Images (vgl. ImageNet, COCO dataset für die Objekterkennung in Bildern) oder Audio (vgl. Tierstimmenarchiv, Xeno-Canto für die Vogelerkennung in Audioaufnahmen).

Zu den methodischen Ansätzen des Text und Data Mining gehören regelbasierte, statistische und neuronale Verfahren unter Verwendung von Methoden des maschinellen Lernens.

Im Folgenden soll der Prozess der intelligenten bzw. maschinellen Nutzung von digitalen Sammlungen zum Zwecke des Information Retrieval und mit Hilfe der KI[36] kurz am Beispiele des Text Mining erläutert werden.

Mit dem rasanten Wachstum der textuellen Quellen des Internets und der sozialen Medien sowie der zunehmenden Transformation gedruckter (und sogar handschriftlicher) in maschinenlesbare Texte mit Hilfe der Optical Character Recognition (OCR) hat das Text Mining in den letzten Jahren stark an Bedeutung gewonnen. Erfolgreiche Anwendungen erfordern allerdings nicht nur inhaltsanalytische Verfahren, sondern den Aufbau einer vollständigen Prozesskette, welche insbesondere die folgenden Arbeitsschritte umfasst:

A. Auswahl und Bereitstellung von Text (z. B. durch Crawlen, OCR oder Nachnutzung bestehender digitaler Sammlungen)
B. Vorverarbeitung von Text (z. B. Entfernen von Steuerzeichen, Tokenisierung, Dublettenbeseitigung)
C. Einbindung von Wissensquellen für die Inhaltsanalyse (z. B. Lexika oder Enzyklopädien wie Wikipedia)
D. Nutzung oder Entwicklung von den Aufgaben angemessener Algorithmen für die Inhaltsanalyse
E. Einbindung der Inhaltsanalyse in einen aufgabenspezifischen Workflow
F. interaktive Visualisierung der Analyseergebnisse
G. Speicherung bzw. Sicherung der Analyseergebnisse und zugrundeliegenden Daten

Grundlage von Text Mining-Anwendungen bilden natürlichsprachliche digitale Sammlungen von Dokumenten wie Presseartikel, technische Dokumente, Bücher und Korrespondenzen, Internetforenbeiträge, digitale Nachlässe oder Meldungen in sozialen Medien. Welche Texte für eine Anwendung verwendet werden, hängt davon ab, welche Fragestellung beantwortet werden soll. Dabei ist zu beachten, dass die inhaltliche Auswahl der Textquellen einen entscheidenden Einfluss auf die Qualität der Analyseergebnisse hat.

Für die Kalibrierung und Evaluation insbesondere von musterbasierten und statistischen Verfahren für die Textanalyse sind umfangreiche Trainingsdaten in Form von digitalen Sammlungen unverzichtbar. Diese digitalen Sammlungen umfassen allgemeinsprachliche und fachspezifische Referenzdaten für die Berechnung von Sprachmodellen, und zwar unabhängig davon, mit welchen Verfahren die Sprachmodelle trainiert werden. Dabei sind vor allem die Frequenzdaten von Wörtern sowie sog. N-Gramme (Abfolgen) von Wörtern und Buchstaben in den Referenzkorpora von Interesse. Neben der Anpassung von Text Mining-Verfahren an bestimmte Sprachen, Genres und Fachsprachen ist eine hochaktuelle Anwendung derartiger Sprachmodelle die Erschließung von nur in Papierform oder als digitales Bild vorliegenden Dokumenten mittels Software für die Schrifterkennung (Optical Character Recognition (OCR) bzw. Handwritten Text Recognition (HTR)).

Große, annotierte Sammlungen werden seit Anfang der 1990er Jahre für Evaluationszwecke im *Information Retrieval* (IR) bereitgestellt. Die Sammlungen enthalten Dokumente, Topics und Relevanzurteile, die im Rahmen von sog. *Text-Retrieval-Konferenzen* (TRECs) für die Evaluation und den Vergleich von Retrieval-Systemen genutzt werden.[37] Unter Bereitstellung einer von menschlichen Nutzern vorgegebenen *Ground Truth* für die Relevanzbewertung von Retrieval-Ergebnissen wird die Effektivität von Retrieval-Maschinen auf der Basis statistischer *Recall-Precision*-Werte bewertet. Neben Texten aus verschiedenen Wissenschaftsbereichen wie Medizin,

Springer Campus.

35 Vgl. Witten, I., Eibe, F., & Hall, M. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Morgan Kaufmann.

36 Vgl. Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., & Wiegand, F. (2018). 10. Das deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. *Digitale Infrastrukturen für die germanistische Forschung*, *6*, S. 219–248. https://doi.org/10.1515/9783110538663-011. Vgl. auch Goldhahn, D., Eckardt, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the 8th International Language Ressources and Evaluation (LREC'12)*, S. 759–765. Vgl. auch Harman, D. (1998). The Text Retrieval Conferences (TRECs) and the Cross-Language Track. *Proceedings of the First International Conference on Language Resources & Evaluation*, S. 517–522.

37 Vgl. Harman, D. (1998). The Text Retrieval Conferences (TRECs) and the Cross-Language Track. *Proceedings of the First International Conference on Language Resources & Evaluation*, S. 517–522.

Recht, Chemie oder Bioinformatik werden auch Texte aus verschiedenen Medien, bspw. Blogs, Nachrichtenportalen, Frage-Antwort-Systemen oder natürlichsprachlichen Dialogsystemen, angeboten.[38]

Das Format *TREC* dient der Herstellung der Vergleichbarkeit von IR-Verfahren durch die Verwendung gleicher Sammlungen und Evaluierungsverfahren. Neben der Bereitstellung realistisch großer Textkorpora umfasst die Plattform auch eine Umgebung für die Testdurchführung und -auswertung.

Das TREC-Prinzip hat sich für die Bewertung und Weiterentwicklung von IR-Verfahren, auch für die Weiterentwicklung der Evaluationssystematik selbst, als außerordentlich hilfreich erwiesen. Für die Evaluation von KI-Verfahren haben sich in den letzten Jahren daher ähnliche Plattformen etabliert, die auf der Grundlage von großen, annotierten Sammlungen und einer aufgabenspezifischen *Ground Truth* die Evaluation von KI-Verfahren ermöglichen. Beispielhaft genannt seien an dieser Stelle *SemEval*, eine Plattform der Association for Computational Linguistics (ACL) für die Evaluation von Verfahren der *Computational Semantics*,[39] und *SentEval*, eine auf GitHub bereitgestellte Bibliothek für die Evaluation der Qualität von sogenannten *Sentence Embeddings*, d. h. Repräsentationen von Wortformen in einem n-dimensionalen Vektorraum zur Darstellung semantisch ähnlicher oder semantisch verwandter Wörter.[40] Ähnlich wie bei TREC werden auch bei diesen Plattformen umfangreiche Sammlungen für verschiedene Themenbereiche und Genres für Evaluationszwecke zur Verfügung gestellt. Zusätzlich können diese Sammlungen auch zum Trainieren von Verfahren verwendet werden, die mit maschinellen Lernverfahren arbeiten.

Für das Trainieren von KI- und Text Mining-Verfahren im Deutschen, insbesondere für die Erstellung von Sprachmodellen, können beispielhaft die *Leipzig Corpora Collection* (LCC)[41] sowie das *Deutsche Textarchiv* (DTA)[42] genannt werden. Die LCC umfasst aktuelle Korpora in verschiedenen Sprachen und Größen unter Verwendung gleicher Formate und vergleichbarer Quellen. Alle Daten liegen als Plaintext vor und können in eine MySQL-Datenbank importiert werden. Sie sind sowohl für die wissenschaftliche Verwendung durch Korpuslinguisten als auch als Trainingsmaterial für Verfahren des Text Mining geeignet. Die Korpora enthalten zufällig ausgewählte Sätze der jeweiligen Korpussprache und sind in Größen von 10.000 bis 1.000.000 Sätzen verfügbar. Als Quelle werden typischerweise entweder Nachrichtentexte oder das Ergebnis allgemeinen Webcrawlings verwendet. Aus urheberrechtlichen Gründen werden die verwendeten Texte immer in einzelne Sätze zerlegt und diese zufällig sortiert, so dass eine Wiederherstellung des Ursprungstextes nicht möglich ist. Ungrammatische Sätze und fremdsprachliches Material sind bestmöglich entfernt worden. Für historische Texte des Deutschen stellt das DTA ein Referenzkorpus deutscher Werke aus dem Zeitraum von ca. 1600 bis 1900 im Volltext bereit. DTA trägt u. a. zum Referenzkorpus für die Erstellung von *Ground Truth*-Daten im Projektvorhaben OCR-D bei, in dem ein Softwarepaket zur OCR-Verarbeitung von Digitalisaten des gedruckten deutschen Kulturerbes des 16. bis 19. Jahrhunderts entwickelt werden soll.[43]

3.3 Qualitätssicherung

Daten*qualität* kann trotz der Bezeichnung verschiedene sowohl *qualitative* als auch *quantitative* Aspekte betreffen. Im Prinzip kann sie zwar nur im Hinblick auf die intendierte Funktion oder Nutzung der Daten für bestimmte Fragestellungen beurteilt werden. Es gibt aber trotzdem einige allgemeine Aspekte, die für alle Sammlungen zu betrachten sind, um deren Qualität zuverlässig bewerten zu können.

Qualitätssicherung betrifft damit zwei verschiedene Ebenen, die aber in engem Zusammenhang zueinander stehen. Zum einen gehören dazu allgemeine, formale Fragen, die für alle Ressourcen beantwortet werden können. Zum anderen ist die Qualität der Daten selbst letztlich nur fachspezifisch hinsichtlich bestimmter Anforderungsprofile und Nutzungsszenarien zu prüfen und zu bewerten – dort aber dann durchaus auch durch sinnvolle Maßzahlen (z. B. Genauigkeit, Fehlerraten) zu quantifizieren.

38 Vgl. National Institute of Standards and Technology (2004, July 15). *Text retrieval conference (TREC) data*. Text REtrieval Conference (TREC). https://trec.nist.gov/data.html.

39 Vgl. SemEval 2020 (2021). *SemEval-2020 International Workshop on Semantic Evaluation*. ALT Website – Arabic Language Technologies Group. https://alt.qcri.org/semeval2020/index.php?id=tasks.

40 Vgl. Conneau, A., & Douwe, K. (2020). *Facebookresearch/SentEval: A python tool for evaluating the quality of sentence embeddings*. GitHub. https://github.com/facebookresearch/SentEval.

41 Vgl. Goldhahn, D., Eckardt, T., & Quasthoff, U. (2012). Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the 8th International Language Ressources and Evaluation (LREC'12)*, S. 759–765.

42 Vgl. Geyken, A., Boenig, M., Haaf, S., Jurish, B., Thomas, C., & Wiegand, F. (2018). 10. Das deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. *Digitale Infrastrukturen für die germanistische Forschung*, *6*, S. 219–248. https://doi.org/10.1515/9783110538663-011.

43 OCR-D. Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR). http://www.ocr-d.de/.

Die Qualität von Sammlungen hängt eng mit der Qualität von Forschungsdaten insgesamt zusammen. Dazu liegt mit dem Papier »Herausforderung Datenqualität« des RfII seit Herbst 2019 eine umfassende Handreichung vor, die den allgemeinen Rahmen ebenso absteckt, wie sie konkrete Erläuterungen und Empfehlungen im Detail bietet.[44] Diese sind für die Zwecke der vorliegenden Handreichung auf die Spezifik von digitalen Sammlungen auszurichten. Dabei werden die Unterschiede in der Betrachtung dadurch verringert, dass Forschungsdaten in der Praxis fast nie als Einzeldaten angelegt und verstanden werden, sondern immer schon in einem weiteren Kontext und in systematischem Vorgehen – und damit zu Sammlungen führend – entstehen.

Der Unterschied besteht folglich zum einen in der Intention, nicht nur einzelne Datenmengen bzw. Datenbestände, sondern insgesamt kuratierte digitale Sammlungen verfügbar zu machen, und damit zum anderen in der besonderen wissenschaftlichen und institutionellen Bedeutung der Einheit ›Sammlung‹, die noch nachdrücklicher die Frage nach einer gelingenden Nachnutzung und nach verlässlichen Qualitätsniveaus aufwirft. Jenseits der Forderung nach Reproduzierbarkeit von Forschungsergebnissen führt zusätzlich die Ausrichtung von digitalen Sammlungen auf ihre Nachnutzbarkeit zu einem besonders hohen Qualitätsanspruch. Nicht zuletzt der zunehmende Einsatz sowohl hoch formalisierter (algorithmischer) als auch hoch komplexer (KI, *deep learning*) Verfahren zur breiten Verwendung umfangreicher Sammlungen verlangt nach Datenbasen, die einem *Gold Standard* oder anderen qualifizierten Qualitätsmarkern entsprechen. Solche Sammlungen können dann neben der weiteren analytischen Auswertung auch als Referenzdaten, als Trainingsdaten und als Evaluationsmittel für Algorithmen dienen.

Auf dem genannten Papier des RfII aufbauend wird hier für die spezifischen Herausforderungen der digitalen Sammlung eine erste Systematisierung und Konkretisierung von Fragen vorgenommen, die als grobe Checkliste und als Ausgangspunkt differenzierter Evaluationsprozesse dienen kann. Zu unterscheiden ist dabei zwischen (I.) den allgemeinen Bedingungen einer Qualitätsbetrachtung, den (II.) gegenstandsspezifischen Merkmalen und schließlich (III.) weiteren Aspekten einer Kultur der Datenqualitätssicherung.

I. Allgemeine Beobachtungen und Rahmenbedingungen

a) **Fassbarkeit und Ansprechbarkeit**: Identifikation – Bibliografische Referenzierbarkeit – Metadaten. Digitale Sammlungen müssen identifizierbar sein. Das setzt in der Regel einen formalen Identifier aus einem geeigneten Schema (z. B. DOI, URN, handle) voraus, der von einem sprechenden Titel begleitet sein kann. Identifikatoren können auch auf andere Granularitätsebenen hinabreichen und z. B. Teilsammlungen, Sammlungsobjekte oder Datensätze und schließlich Objektteile und Einzelinformationen betreffen. Die Versorgung mit feingranularen Identifikatoren hat Einfluss auf die Benutzbarkeit von Sammlungen und Sammlungsteilen. Grundlage für eine Beschreibung, Zitation und Auseinandersetzung mit einer Sammlung im wissenschaftlichen Diskurs ist darüber hinaus die Möglichkeit einer bibliografischen Beschreibung, für die verantwortliche Beteiligte, der institutionelle Kontext und die Entstehungszeit benannt oder zumindest identifizierbar sein müssen. Wünschenswert ist die Beschreibung von Sammlungen mit geeigneten Metadaten nach sammlungsorientierten Schemata und in übergreifenden Sammlungskatalogen. Eine fachübergreifende Praxis zur Nutzung standardisierter Beschreibungsmodelle hat sich allerdings noch nicht etabliert, mit DCAT (s. oben) steht aber ein vielversprechender Kandidat zur Verfügung.

b) **Dokumentation**. Die nachhaltige und belastbare Nutzung von Datensammlungen hängt stark von deren möglichst vollständiger und transparenter Dokumentation ab. Die Inhalte der Sammlung und ihr Umfang sind zusammenfassend zu beschreiben. Ebenso die Struktur der Sammlung und der Datensätze, z. B. indem die verwendeten Datenmodelle, Datenstandards und Schemata offengelegt werden. Zudem müssen Genese und Kontext der Sammlung beschrieben werden. Es muss eine Antwort auf die Frage möglich sein: Wer hat wann was unter welchen Umständen zu welchem Zweck getan? Hierzu gehört auch die Dokumentation der Methodik zur Erarbeitung der Datensammlung: zur Auswahl, zur digitalen Aufnahme, zur Verarbeitung und zur Anreicherung der Daten bzw. allgemein zu Änderungen, die durch Versionierung nachvollziehbar bleiben müssen. Im Idealfall werden auch Hinweise zu den Nutzungsmöglichkeiten der Daten gegeben. Schließlich sind die bereits getroffenen Maßnahmen und durchgeführten Aktivitäten zur Qualitätssicherung zu beschreiben.

c) **Institutionelle Verortung und Verantwortung**. Qualität, Zuverlässigkeit und Integrität einer Sammlung wird häufig dadurch gesichert, dass sie in einem stabilen institutionellen Kontext erarbeitet und zur Verfügung gestellt wird. In Betracht zu ziehen ist deshalb, ob mit einer Sammlung eine institutionelle

44 Vgl. RfII – Rat für Informationsinfrastrukturen. (2019). *Herausforderung Datenqualität: Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel*. https://rfii.de/?p=4043.

Zuständigkeit und Verantwortung verbunden ist. Gibt es ein institutionelles Bekenntnis (»Commitment«) zur Erstellung, zur weiteren Pflege und dauerhaften Bereitstellung?

d) **Umsetzung der FAIR-Prinzipien auf Sammlungsebene**. Die FAIRness einer digitalen Sammlung trägt zu ihrer Qualität bei. Die vier FAIR-Kriterien sind deshalb auch hier zu berücksichtigen. *Findability*: Ist die Datensammlung in fachspezifische oder in allgemeine Repositorien integriert und darüber gut sichtbar und leicht auffindbar? Gibt es andere Maßnahmen, die für eine Propagierung und gute Sichtbarkeit der Sammlung sorgen sollen? *Accessibility*: Ist die Datensammlung leicht zugänglich? Besteht die Möglichkeit zu einem vollständigen Download? Gibt es Schnittstellen und APIs zur automatisierten Interaktion mit den Daten? *Interoperability*: Welche Formate und Standards werden für die Sammlung genutzt? Sind dies die aktuellen und gängigen Standards im betreffenden Feld? Werden kodifizierte Standards strikt verwendet oder handelt es sich um lokale Anpassungen oder gar idiosynkratische Modelle von geringer Verbreitung? *Reusability*: Für welche Nutzungsszenarien ist die Datensammlung verwendbar? Ist die Nachnutzbarkeit durch entsprechende Rechteeinräumungen und Lizenzregelungen sichergestellt?

e) **Berücksichtigung der CARE- Prinzipien.** Jenseits von FAIR sind digitale Sammlungen auch nach den CARE-Prinzipien[45] zu beurteilen, d. h. unter den Gesichtspunkten von allgemeinem Nutzen, Zugang und Kontrolle über die eigenen Daten sowie Verantwortung und Beachtung ethischer Grundsätze.Im Zuge von Forschung und Wissenschaft werden auch personalisierte Daten mit Bezug zu Individuen und Gruppen erhoben, die zum Teil auch vulnerable Gruppen betreffen. Deshalb sollte bei jeder digitalen Sammlung geprüft werden, ob es einen gemeinsamen Nutzen für die beteiligten Gruppen und Individuen gibt, und wie sich inhärente Macht- und Kontrollmechanismen auswirken. Ethische Normen der Gerechtigkeit und Schadensminimierung durch die Nutzung sind explizit festzulegen; die Verantwortung im Umgang mit Daten sollte dokumentiert sein. Gerade deutsche Forschende und Institutionen bauen durch ihre Ansässigkeit im Globalen Norden auf einer globalen strukturellen Ungleichheit auf, die einen bewussten und ethischen Umgang mit Forschungsdaten und die genaue Abwägung von Verantwortlichkeiten, Reziprozität und Nutzen digitaler Sammlungen erfordert, im globalen Kontext auch jenseits indigener und vulnerabler Gruppen.

II. Inhaltsbezogene Qualitätsevaluation

Eine Qualitätsmessung im engeren Sinne fragt danach, ob die Sammlung ›gute‹ oder ›schlechte‹ Daten enthält und in welchem Maße die Daten für die weitere Forschung ›nutzbar‹ sind. Dies kann aber nicht absolut, sondern nur in Bezug auf die individuellen Nutzungsabsichten und die jeweils gültigen Rahmenbedingungen beantwortet werden und erfordert sach- und fachspezifische Kriterien. Verallgemeinernd lassen sich folgende Startpunkte für evaluierende Fragen identifizieren:

a) **Erschließungstiefe und Annotation**. Sind die Daten flach oder tief erschlossen? Handelt es sich um rein mechanisch gewonnene oder angereicherte oder annotierte Daten? Ist das angewandte Datenmodell oder Beschreibungsraster auf Einfachheit oder Reichhaltigkeit ausgerichtet?

b) **Datenerhebung**. Entspricht das angewandte Verfahren zur Erzeugung der Daten der aktuellen *Best Practice*? Welchen Ansprüchen kann es genügen? Ist das Verfahren konsequent und fehlerfrei angewandt worden? Hat es zu konsistenten, gleichmäßigen Daten geführt?

c) **Datengenese**. Sind die Bedingungen der Datengenese hinreichend transparent? Ist sicht- und nachvollziehbar, wie die beobachtete Welt bzw. die digitalisierten Objekte in Daten transformiert worden sind? Ist erkennbar, welche Filter, Auswahlentscheidungen und ggf. Schieflagen in die Geräte und Sensoren der Datenerfassung eingeschrieben gewesen sind? Ist dokumentiert oder rekonstruierbar, welche Konzepte und Modelle der Datenerfassung zugrunde gelegen haben und welche Erfassungspraktiken im Prozess eine Rolle gespielt haben?

d) **Datenaufbereitung**. Entspricht das angewandte Verfahren zur Vorverarbeitung und ggf. Anreicherung der Daten mit weiteren Informationen den gegenwärtigen Erwartungen und Anforderungen? Sind die entsprechenden Verfahren konsequent, gleichmäßig und fehlerfrei angewandt worden?

e) **Maßnahmen zur Qualitätssicherung**. Ist die Sammlung einem strukturierten, dokumentierten und nachvollziehbaren Prozess zur Qualitätssicherung unterworfen worden? Hat eine Konsistenzprüfung oder Fehlerbereinigung stattgefunden?

45 Vgl. Global Indigenous Data Alliance (GIDA). (2018). *CARE principles of Indigenous data governance – Global Indigenous data alliance*. https://www.gida-global.org/care.

f) **Kennzahlen**. Gibt es bereits Untersuchungen oder Besprechungen zur Qualität der Daten? Gibt es Maßzahlen zur Qualität der Daten? (z. B. Fehlerraten, Vollständigkeit der Erschließung und Annotation) Lassen sich Angaben zur Vollständigkeit machen?

Aus den genannten Kriterien sollten sich zusammenfassende Qualitätsbewertungen ableiten lassen, die aber streng genommen immer nur in Relation zu bestimmten Verwendungsszenarien stehen können: ›Für Szenario X ist diese Datensammlung von ausreichender | bedingt ausreichender | nicht ausreichender Qualität‹.

III. Weitere Aspekte der Qualitätssicherung bei digitalen Sammlungen. Zu einer Kultur der Evaluation.
Der Umgang mit Forschungsdaten im Allgemeinen und digitalen Sammlungen im Besonderen muss in eine allgemeine Kultur der Datenerstellung und Datennutzung eingebettet sein. Diese schafft die nötige Sensibilität und sorgt für die erforderlichen Kompetenzen beim Aufbau und der Verwendung von Sammlungen für die Forschung. Zu den Bereichen einer sich allmählich entwickelnden Datenkultur gehören u. a.:

a) **Reviewing und Rezensionswesen**. Digitale Sammlungen sollten möglichst breit einem (*Peer*) *Review*-Verfahren unterworfen werden. Allmählich etablieren sich Datenjournale (s. o.) und Rezensionszeitschriften (z. B. RIDE), in denen die Qualität von Forschungsdaten besprochen und bewertet wird. Dies trägt nicht nur zu einer besseren und fundierten Nutzung bei, sondern hilft auch bei der Verbreitung und Durchsetzung allgemeiner Qualitätsstandards.
b) **Daten- und Sammlungs-Autorschaft und Kreditierung**. Gute Sammlungen beruhen häufig auf hohem Ressourceneinsatz. Ihre Erstellung und Kuratierung ist sehr aufwändig und verlangt u. U. besondere wissenschaftliche Kompetenzen. Der Ausweis von Autorschaften und mitarbeitenden Beteiligungen ist nicht nur eine Frage der angemessenen Kreditierung von Leistungen im akademischen System der Reputation und Berufsentwicklung. Sie hilft auch dabei, die Qualität von Daten durch die Verbindung mit verantwortlichen und kompetenten Erstellerinnen und Erstellern besser einschätzbar zu machen.
c) ***Data Literacy* für Datenkuration und Datennutzung**. Die Verwendung von Forschungsdaten und der daraus hervorgehenden Datensammlungen erfordert besondere Lesefähigkeiten, die im Aus- und Fortbildungssystem der Wissenschaft noch weiter ausgebaut werden müssen. Der kritische Umgang mit Daten ist nicht nur die Grundlage für ihre sachgemäße Bewertung, sondern auch für eine anhaltende Kuratierung und Informationsanreicherung. Datenqualität entsteht auf der Basis von Kompetenzen bei Erstellenden und Nutzenden und aus dem beständigen Abgleich der Genese und der Verwendung von Daten.

Forschende als Akteure und Nutzende der Qualitätssicherung.
Die Evaluation von digitalen Sammlungen richtet sich an insgesamt drei Gruppen: Nutzer, Gutachter, Ersteller. Erstens werden für die (Nach-)Nutzung von Daten zuverlässige Bewertungen gebraucht, auf deren Grundlage die weitere Forschung vollzogen werden kann. Für die Begutachtung müssen dazu zweitens geeignete übergreifende Kriterien und Untersuchungsmodelle geschaffen werden, die zu Sammlungs-Rezensionen in Journalen und auf anderen Plattformen führen können. Drittens hat sich bereits in der Vergangenheit gezeigt, dass Kriterienkataloge zur Bewertung von Forschungsressourcen nicht nur für die Rezeption von Bedeutung sind, sondern bereits bei der Erstellung von Datensammlungen eine wertvolle Hilfestellung bieten können. Denn schließlich sind die Kriterien zur Qualitätsuntersuchung zugleich eine Handreichung zum Aufbau möglichst qualitätvoller Ressourcen. Aus beiden Perspektiven sind die gleichen Punkte zu berücksichtigen.

3.4 Rechtliche Rahmenbedingungen

Die vorliegenden Handreichungen können die komplexen rechtlichen Rahmenbedingungen, die sich mit dem Aufbau und der Nutzung digitaler Sammlungen verbinden, nicht ausführlich behandeln. Sie verweisen daher insbesondere auf die Aktivitäten der AG 7 »Recht für Wissenschaft im digitalen Zeitalter« der Allianzinitiative. Für rechtssichere Auskünfte zu konkreten Sachlagen muss im Zweifelsfall auf juristisch einschlägig geschultes Personal zurückgegriffen werden.

Dennoch lassen sich zumindest einige Anforderungen formulieren, die unter rechtlichen Gesichtspunkten beim Aufbau, Angebot und bei der Nutzung digitaler Sammlungen beachtet werden sollten. Dazu gehört vor allem die Wahl einer möglichst freie Lizenz für die Daten. Ideal sind hier z. B. *Creative Commons Zero* (CC0)[46]

46 Vgl. *Creative Commons – CC0 1.0 Universell* (n. d.). Creative Commons. https://creativecommons.org/publicdomain/zero/1.0/deed.de.

oder *Public Domain Dedication and Licens*e (PDDL)[47], da es mit diesen Lizenzen keine Probleme beim Aggregieren oder Kombinieren von Daten gibt. Umgekehrt kann schon die Lizenzierung mit der Lizenz CC BY-SA in dem Fall problematisch sein, wenn Daten unter dieser Lizenz mit Daten unter der Lizenz CC BY-NC kombiniert werden, da für das gesamte Datenset die Einschränkungen beider Lizenzen wirksam werden. Eine rechtlich differenzierte Behandlung einzelner Datensätze in einem aus unterschiedlich lizenzierten Datenquellen zusammengesetzten Datenset ist in der Praxis nahezu unmöglich.

Zu beachten sind daneben vor allem Fragen des Datenschutzes bei personenbezogenen Daten bzw. die Einhaltung der DSGVO.[48] Hier sollten beim Aufbau der Sammlung Einwilligungen von betroffenen Personen eingeholt werden oder anderweitig geeignete Maßnahmen zum Datenschutz ergriffen werden. Im Sinne der weitestgehenden Nutzbarmachung von Daten empfiehlt es sich, im Zweifelsfall Daten nicht einfach zu sperren, sondern zumindest die Daten freizugeben, die rechtlich unbedenklich sind. Anbieter können Daten z. B. anonymisieren, indem sie Identifikatoren und Daten getrennt speichern oder nur Teile von Datensets zugänglich machen. Weiter sind bei den Nutzungsszenarien Grundsätze der Eignung, der Erforderlichkeit und der Verhältnismäßigkeit zu beachten und Zugriffsregelungen für eine datenschutzkonforme Verwendung festzulegen (z. B. in Nutzungsverträgen mit Forschern).

Im Weiteren empfiehlt es sich, unabhängig von direkten Rechtsfragen auch auf die allgemeinen Regeln der guten wissenschaftlichen Praxis[49], auf Richtlinien internationaler Initiativen[50] oder auf besondere problematische Aspekte in den Daten hinzuweisen. Ein gutes Beispiel dazu ist die Diskussion um Sammlungen, die koloniale Kontexte berühren, die Unterdrückung und Verfolgung von Personengruppen betreffen oder kulturell bedingtes Unwohlsein auslösen (Abbildung von Ahnen etc.).[51]

4 Organisation und Management

4.1 Aufbau Digitaler Sammlungen

4.1.1 Präambel

Sammlungen entstehen aus den unterschiedlichsten Anlässen. Sowohl repräsentative als auch museale, merkantile als auch wissenschaftliche Ausrichtungen zählen bis in die Gegenwart zu den Facetten einer lebendigen und vielfältigen Sammlungspraxis. Wenngleich die Digitalisierung dies einerseits fortsetzt, schafft sie andererseits für die Entstehung und den Aufbau von Sammlungen neue Rahmenbedingungen und neue Möglichkeiten, die bislang nur ansatzweise ausgelotet wurden.[52]

Die Standortbindung von physischen Sammlungen geht oft mit einer konzentrierten Zweckgebundenheit einher, die sich im Bereich der digitalen Sammlung aufzulösen scheint. Mit dem Wegfall der physisch notwendigen Verankerung und ferner mit der Trennung der digitalen Objekte von der Präsentationsform derselben sind neue, unterschiedlichste Entstehungskontexte für digitale Sammlungen denkbar. Physisch ungebunden sind zudem nicht nur die Sammlungen, sondern in zunehmender Anzahl auch gesammelte Objekte (*born digital*), woraus sich neue Sammlungs- und Archivierungspraktiken ergeben (*Web Scraping*, etwa durch die *Wayback Machine*). Da die maschinenlesbaren Formen und Formate digitaler Sammlungen der Sichtbarkeit weitgehend entzogen sind, werden sie nur selten in der wissenschaftlichen Öffentlichkeit thematisiert. Dort spielen aber fachlich anerkannte, offene Vokabulare und Normdaten eine entscheidende Rolle beim Sammlungsaufbau, um die grundsätzliche Anschlussfähigkeit an ein größeres Wissensnetz überhaupt zu gewährleisten (*Giant Global Graph*).[53]

47 Vgl. Open Knowledge Foundation (n. d.). *Open data Commons public domain dedication and license (PDDL) – Open data Commons: Legal tools for open data*. https://opendatacommons.org/licenses/pddl/.

48 Eine (exemplarisch für einen bestimmten Forschungsbereich) gute Übersicht mit praktischen Hinweisen bietet die »Handreichung Datenschutz«, Bäcker, M., & Golla, S. J. (2020). *Handreichung Datenschutz 2. vollständig überarbeitete Auflage*. Rat für Sozial- und Wirtschaftsdaten (RatSWD). https://www.forschungsdaten-bildung.de/files/RatSWD_Output8.6_HandreichungDatenschutz_2.pdf.

49 Deutsche Forschungsgemeinschaft (DFG) (n. d.). *Gute wissenschaftliche Praxis. Kodex ›Leitlinien zur Sicherung guter wissenschaftlicher Praxis‹*. https://www.dfg.de/foerderung/grundlagen_rahmenbedingungen/gwp/.

50 Vgl. Europeana (n. d.). *Reusability FAQ*. https://www.europeana.eu/de/help/reusability.

51 Vgl. Crowley, B. (2021, October 13). Acknowledging harm, rethinking collections. *Biodiversity Heritage Library*. https://blog.biodiversitylibrary.org/2021/10/acknowledging-harm-rethinking-collections.html.

52 Vgl. Stäcker, T. (2019). Die Sammlung ist tot, es lebe die Sammlung! *Bibliothek Forschung und Praxis*, *43*(2), S. 304–310. https://doi.org/10.1515/bfp-2019-2066.

53 Vgl. Blogbeitrag von Berners-Lee (2007): https://web.archive.org/web/20160713021037/http://dig.csail.mit.edu/breadcrumbs/node/215.

Im Lichte der neuen Möglichkeiten steht aber auch die Frage nach der Motivation und der Autorisierung von Sammlungen erneut zur Debatte. Grundsätzliches Interesse haben zunächst Institutionen, die ihre physischen Sammlungen digital ›ins Netz‹ bringen möchten. Sofern es nicht um Massendigitalisierungsprojekte geht, können damit aber auch individuelle Forschungsziele oder auch museale Zwecke verbunden werden; digitale Sammlungen können letztlich sogar als Nebenprodukt oder alternative Nachnutzungsszenarien von Forschungsaktivitäten entstehen. Zunehmend spielt dabei auch das allgemeine Interesse der Öffentlichkeit hinein, was sowohl Sammlungsprozesse als auch spezifische Sammlungsobjekte betreffen kann.

Noch zu weit im Hintergrund steht dabei der öffentliche Diskurs um eine offene, transparentere Wissenschafts- und Forschungspraxis (*Open Science*). Während Digitalisierungsprozesse im engeren Sinne davon seltener berührt werden, führt dies bereits im Vorfeld der Sammlungserstellung zu notwendigen Überlegungen auf dem Gebiet der Erschließungs-, Bereitstellungs- und Nachnutzungspraxis: Welcher Grad an Nachnutzbarkeit wird durch die angestrebten Lizenzen erreicht, welche Erschließungstiefe und -qualität wird den spezifischen Ansprüchen eines Sammlungsvorhabens und gleichzeitig den allgemeinen Richtlinien der Institutionen und Fördereinrichtungen gerecht? Welche Impulse kann eine digitale Sammlung für die Anschlussforschung potenziell geben, und welche Maßnahmen begünstigen diese? Kann die Öffentlichkeit möglicherweise bereits bei der Erschließung einbezogen werden (*Citizen Science*, *Crowd Sourcing*)? Und inwieweit müssen bereits in der Aufbau- und Erschließungsphase einer digitalen Sammlung die möglichen Präsentations- und Bereitstellungsformate antizipiert werden?

4.1.2 Methodische Aspekte des Aufbaus

Der Aufbau einer digitalen Sammlung unterscheidet sich prinzipiell von dem einer physischen Sammlung, da sie ortsunabhängig ist sowie dynamisch verändert und modifiziert werden kann (vgl. Einleitung).[54] Eine Sammlung besteht aus Sammlungsteilen, die unter einem oder mehreren gemeinsamen Gesichtspunkten zusammengeführt wurden. Ziel der hier im Vordergrund stehenden wissenschaftlichen Sammlung[55] ist es, distinkte, voneinander verschiedene Einheiten nach für den Sammlungsaufbau relevanten Kriterien zu versammeln, d. h. die Sammlung soll möglichst keine Dubletten bzw. identische Kopien enthalten, unbeschadet dessen, dass in der naturwissenschaftlichen Praxis Redundanz eine wichtige Rolle spielt. Was dabei die Einheit bzw. Granularität der Sammlung ist bzw. ausmacht, ist kontingent. Als kleinste sammlungsbildende Einheit kann das ›Datum‹ angesehen werden, das als solches einen Unterschied in einem Kontinuum macht (s. oben).[56] Die Art der Unterschiede der Daten wiederum ergeben sich aus der Sammlungscharakteristik. So werden einzelne Personen in einer Personennormdatei durch individualisierende Kriterien (Geburts- und Sterbejahr, Geburts- und Sterbeort etc.) voneinander unterschieden und werden damit zu Daten einer Sammlung. Die Qualität einer Sammlung hängt ganz wesentlich von der Klarheit dieser Kriterien zu ihrem Aufbau ab. Je weniger präzise das Kriterium der Sammlung ist, desto weniger aussagekräftig wird sie sein. Erfolgreiche Sammlungsbildung ist daher immer abhängig von einem klaren Begriff und sorgfältigen Beschreibung der Art des gesammelten Gegenstandes.

Eine (wissenschaftliche) Sammlung ist mit Blick auf ihre Kriterien in der Regel auf Vollständigkeit angelegt. Eine Sammlung von Biografien von französischen Schriftstellern soll alle verfügbaren distinkten Einheiten bzw. Daten enthalten. Sammlungen, die diese Vollständigkeit realiter nicht erfüllen (können), versuchen Vollständigkeit durch Repräsentativität oder andere (statistische) Modelle zu ersetzen. Selbst eine Sammlung nach Opportunität[57] ist eine pragmatische Hilfskonstruktion für eine durch äußere Gegebenheiten unerreichbare, aber zumindest erhoffte Vollständigkeit, mit anderen Worten: man unterstellt, dass eine opportunistische Sammlung sich ›irgendwie‹ ähnlich verhält wie eine vollständige Sammlung.

54 S. a. Rat für Informationsinfrastrukturen (2021): Bestandsbezogene Forschung gestalten: zukunftsfähige Verschränkungen von »digital« und »analog«. Ein Diskussionsimpuls zur wissenschaftlichen, wissenschaftsnahen und kulturellen Nutzbarkeit von Sammlungen, Göttingen, S. 1 f.: »Sammlungsobjekte verändern mit ihrer digitalen Darstellbarkeit und der dadurch ermöglichten neuartigen Wahrnehmung ihre ›Identität‹. Sie lassen sich auf ganz neue Weise verknüpfen und sind keine klar lokalisierbaren, physisch raumgreifenden oder auch konzeptionell begrenzten Entitäten mehr.«

55 Vgl. Sommer, M. (1999). *Sammeln: Ein philosophischer Versuch.* suhrkamp taschenbuch wissenschaft. Manfred Sommer unterscheidet die ästhetische von der ökonomischen Sammlung. Die wissenschaftliche Sammlung ist in diesem Sinne ein Sonderfall der ›ästhetischen‹ Sammlung.

56 Vgl. Floridi, L. (2013). *The philosophy of information*. OUP. S. 85 »data as a lack of uniformity« bzw. »datum = def x being distinct from y.«

57 Vgl. Schöch, C. (2017). Aufbau von Datensammlungen. In: F. Jannidis, H. Kohle, & M. Rehbein (Hrsg.), *Digital humanities: Eine Einführung* (S. 223–233). Springer-Verlag. Mit dem Konzept der opportunistischen Sammlung charakterisiert Schöch die Situation, dass in einigen Fällen mangels Verfügbarkeit nur mit ungesicherten Datengrundlagen gearbeitet werden kann, dass aber dennoch sinnvolle Aussagen möglich sind.

Digitale Sammlungen von Daten lassen sich in solche unterscheiden, die *born digital* sind, und solche, die durch eine Transformation von einem analogen in ein digitales Medium, entstehen bzw. hergestellt werden. Neben dem primären Aufbau einer Sammlung *de novo* durch Erfassung von Daten entstehen digitale Sammlungen vor allem durch Aggregation, Filterung, Umwandlung,[58] Harmonisierung,[59] Rearrangieren oder aber Referenzieren von Sammlungsteilen.

Am relevantesten scheint hier die Aggregation oder Integration verschiedener Sammlungen in eine neue Sammlung. Hier entstehen oft Probleme, da unterschiedliche Datenformate und -strukturen oder auch Granularitäten zu beachten sind, die nicht selten aufwändige Schritte der Datenbereinigung, -anreicherung und/oder -homogenisierung erfordern. Nicht weniger herausfordernd sind die Probleme, die rechtlicherseits oder organisatorisch aus dem Zusammenspielen oder kollaborativen Erstellen von Sammlungen erwachsen.[60] Um die Aggregation oder Integration von Sammlungen oder Sammlungsteilen zu erleichtern, sind vor allem standardisierte Dokument- und Datenstrukturen wichtig,[61] die auf generischen Datenstrukturen wie CSV, JSON, XML oder auch RDF-Serialisierungen (RDF/XML, N3, turtle, JSON-LD) aufsetzen.

Unterscheiden lassen sich dabei strukturierte von semistrukturierten Datenformen. Während CSV eher für stark strukturierte Daten in Frage kommt, eignet sich XML als Markup-Sprache vor allem für semistrukturierte Daten wie Textdokumente. Eine andere Unterscheidung liegt im Unterschied von binären und textlichen Daten. Zu beachten ist hier allerdings, dass der Begriff des Datums nicht mit dem technischen Begriff der Datei gleichgesetzt werden kann, auch wenn das Datum durch die Datei repräsentiert wird.[62] Ein Datum ist nur durch seine Differenz zur Konstituierung der Einheit einer digitalen Sammlung bestimmt und kann aus mehreren, auch nach ihrem *MIME type* verschiedenen Dateien zusammengesetzt sein, wie z. B. eine Webseite.

Für den Sammlungsaufbau spielen Dateien (*Files*) eine wichtige Rolle, da nur durch Zugriff auf die Dateien Daten gelesen werden und Sammlungen ausgewertet werden können. Die Dateien müssen zu diesem Zweck an einem Ort zusammengeführt oder aber über Schnittstellen bzw. APIs (*Application Programming Interfaces*) ausgelesen und in ein gemeinsames Ergebnis integriert werden, etwa über *Federated Search*. Ob man Dateien physisch oder über verteilte Abfragen zusammenführt, hängt von der Stabilität und Performance der Verbindung, der Nachhaltigkeit des Zugangs, der Integrierbarkeit oder dem Umfang der digitalen Sammlung ab (z. B. erlauben Dateien im Kontext des HPC, *High Performance Computing*, aufgrund ihrer enormen Größe keine Verlagerung auf andere Systeme). Liegen Daten in einem verteilten System vor, sind Fragen der Synchronisation und dafür erforderliche Techniken und standardisierte Protokolle zu beachten, um die Aktualität der Daten sicherzustellen. Typische Protokolle zur Synchronisierung von Daten sind z. B. OAI-PMH oder Resourcesync.[63]

4.1.3 Retrodigitalisierung und Sammlung von Born Digitals

Die technischen Entwicklungen im Bereich der Digitalisierung führen dazu, dass analoge Bestände in Bibliotheken, Archiven und Museen digitalisiert werden, um neue Potenziale der Nachnutzung und Beforschung zu eröffnen. z. B. sind mit den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16., 17. und 18. Jahrhunderts einzigartige, verteilte und kollaborativ organisierte wissenschaftliche Datenbanken entstanden. Eine objektive Beurteilung des gesamten Fortschritts der Digitalisierung in Deutschland ist allerdings schwierig, da es keine Übersichtskennzahlen gibt.

Obwohl einzelne Programme von Bund und Ländern gefördert werden und die Deutsche Forschungsgemeinschaft (DFG) im Bereich der Wissenschaftlichen Literaturversorgungs- und Informationssysteme (LIS) Projekte an wissenschaftlichen Einrichtungen, insbesondere Service- und Informationseinrichtungen in Deutschland

58 Vgl. Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., & Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_006.

59 In der Wissenschaft werden durch verschiedene Methoden – bspw. durch Messung durch Sensoren – digitale Daten erhoben. Diese liegen zunächst als unstrukturierte digitale Daten vor. Um diese für eine Sammlung aufzubereiten, müssen sie erschlossen und harmonisiert werden.

60 Zu dieser und weiteren Aspekten des Zusammenspiels von Daten, vgl. Engelhardt, C., & Kusch, H. (2021). 5.3 Kollaboratives Arbeiten mit Daten. In: M. Putnings, H. Neuroth, & J. Neumann (Hrsg.), *Praxishandbuch Forschungsdatenmanagement*, S. 451–476. Walter de Gruyter. https://doi.org/10.1515/9783110657807-025.

61 Vgl. Corcho, O., Eriksson, M., Kurowski, K., Ojstersek, M., Choirat, C., Sanden, M., & Coppens, F. (2021). *EOSC interoperability framework: Report from the EOSC executive board working groups FAIR and architecture*. EOSC Executive Board. https://doi.org/10.2777/620649.

62 Vgl. Berg-Cross, G., Ritz, R., & Wittenburg, P. (2015). *RDA Data Foundation and Terminology DFT: Results RFC*. https://www.rd-alliance.org/system/files/DFT%20Core%20Terms-and%20model-v1-6.pdf: »a Digital Object is represented by a bitstream, is referenced and identified by a persistent identifier and has properties that are described by metadata«.

63 Vgl. Open Archives Initiative. (2017, February 22). *ResourceSync framework specification*. https://www.openarchives.org/rs/toc.

im Bereich der Digitalisierung finanziell unterstützt, fehlt es an einer systematischer Förderung und Planung der Digitalisierung durch eine übergeordnete Stelle in Deutschland.

Um Antragstellenden die Planung von Digitalisierungsprojekten zu erleichtern und die Begutachtung von Anträgen vergleichbar zu gestalten, wurden die *Praxisregeln »Digitalisierung«* herausgegeben[64]. Sie enthalten Regeln zu verschiedenen Aspekten der Digitalisierung wie Metadaten, Langzeitarchivierung oder Volltextgenerierung. Daneben stellen sie auch Regeln zum Zitieren und zur Bereitstellung der Digitalisate im Open Access auf.

Digitalisierung ist als stetige Aufgabe der Gedächtnisinstitutionen zu begreifen, diese sind mit den erforderlichen zusätzlichen Mitteln auszustatten, auch sind entsprechende Karrierewege zu etablieren. Vordringlich ist, dass der Aufbau von Sammlungen auch kollaborativ zwischen den verschiedenen Institutionen und spartenübergreifend erfolgt, um eine möglichst vollständige Abdeckung des jeweiligen digitalen Sammlungsgegenstandes zu erreichen.

Eine bislang noch nicht vollständig gelöste Problematik stellt die Tiefe einer generischen Digitalisierung mit Blick auf mögliche differenzierte fachspezifische Anforderungen dar. Das Papier *Stand der Kulturdigitalisierung in Deutschland*[65] identifiziert folgende medienunabhängige, aufeinander aufbauenden Stufen der Digitalisierung:

- Erschließung der Objekte mittels Metadaten,
- Erstellung von digitalen Repräsentationen,
- Erstellung von prozessierbaren Daten,
- Anreicherung der prozessierbaren Daten mittels Normdaten, strukturierten Klassifikationen und Annotationen.

Darüber hinaus sollten in Anlehnung an die FAIR-Prinzipien insbesondere folgende Anforderungen für digitalisiertes (Kultur-)Gut gelten:

- persistente und eindeutige Referenzierbarkeit mittels PIDs,
- Maschinenlesbarkeit und Prozessierbarkeit über standardisierte Schnittstellen,
- Versionierung von neuen Fassungen und Zusammenstellungen,
- Angabe von eindeutigen Lizenz- und Nutzungsangaben – sofern rechtlich möglich im Open Access,
- Dokumentation.

Im Weiteren können je nach Fach- und Gegenstandsgebiet weitere Anforderungen hinzukommen, z. B. spezifische Annotationen in der Computerlinguistik. Insofern ist die Digitalisierung von (Kultur-)Gut kontextabhängig und muss ggf. spezifischen Anforderungen gehorchen, damit die (wissenschaftliche) *Nachnutzung* sichergestellt werden kann. Hier ist die Entwicklung und die Verwendung interoperabler einheitlicher Standards auf disziplinärer, aber auch transdisziplinärer bzw. generischer Ebene notwendig.

Eine weitere offene Frage ist die Digitalisierung von Material, das derzeit noch urheberrechtlich geschützt ist oder bei dem sich Fragen des Daten- oder des Persönlichkeitsschutzes ergeben. Hier können anonymisierte Teilpublikationen oder urheberrechtlich unbedenkliche ›abgeleitete‹ Datenformate zum Einsatz kommen.[66]

Voraussetzung zur Nachnutzung von Digitalisaten ist deren *Auffindbarkeit* und *Zugänglichkeit*. Hierfür gibt es schon verschiedene Nachweisinstrumente für Digitalisate, wie die Deutsche Digitale Bibliothek, das Archivportal-D, die Europeana und die Verbundsysteme der wissenschaftlichen Bibliotheken, die Bestände verschiedener Institutionen zusammenführen. Unabhängig von konkreten zentralen Portalen ist die Auffindbarkeit der Digitalisate durch ausreichende, für den jeweiligen Zweck angemessene Metadaten sicherzustellen. Denn nur so können Forschende die für sie relevanten digitalen Objekte mittels Suchmaschinen im Netz ermitteln.

Neben der Verwendung von 2D-Digitalisaten kommen zunehmend auch 3D-Digitalisate in der Forschung zum Einsatz, die man in bestimmten Forschungszusammenhängen als Äquivalente oder Stellvertreter für physische Objekten ansehen kann (vgl. a. das Konzept des ›digitalen Zwillings‹). Folglich kommen sie für Simulationen in Betracht, die am physischen Objekt durchgeführte Tests ersetzen können.

64 Vgl. Deutsche Forschungsgemeinschaft (2017). *DFG-Praxisregeln »Digitalisierung«*. http://www.dfg.de/formulare/12_151/12_151_de.pdf.

65 Vgl. Klaffki, L., Schmunck, S., & Stäcker, T. (2018). *Stand der Kulturdigitalisierung in Deutschland - Eine Analyse und Handlungsvorschläge des DARIAH-DE Stakeholdergremiums »Wissenschaftliche Sammlungen«*. GEODOC, Dokumenten- und Publikationsserver der Georg-August-Universität. http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2018-1-3.

66 Vgl. Schöch, C., Döhl, F., Rettinger, A., Gius, E., Trilcke, P., Leinen, P., Jannidis, F., Hinzmann, M., & Röpke, J. (2020). Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. *Zeitschrift für digitale Geisteswissenschaften*. https://doi.org/10.17175/2020_006.

Neben Sammlungen, die durch mediale Transformation (*Datafication*) entstanden sind, gibt es Sammlungen, die *born digital* sind. Sie unterscheiden sich von digitalisierten Sammlungen lediglich darin, dass sie keine Rückbindung an ein analoges Original aufweisen. Das stellt insofern eine Herausforderung dar, als vor allem die Provenienz der Daten, die die Sammlung konstituieren, sorgfältig beschrieben werden muss. Als generischer Standard bietet sich hier z. B. *PROV* an.[67] Eine weitere Besonderheit liegt darin, dass ohne eine analoge Grundlage kein implizites Qualitätskriterium in der Genauigkeit der Übertragung des Informationsgehalts der analogen Quelle gegeben ist. *Born digital*-Sammlungen müssen daher auf ihren Referenzcharakter geprüft bzw. abgeleitete Datensammlungen an diesen gemessen werden. In praktischer Hinsicht gelten die in 4.1 formulierten methodischen Grundsätze, denen zufolge Datenlieferanten und Sammlungsanbieter vor allem Vorsorge treffen sollten, dass Sammlungen den FAIR-Prinzipien entsprechen. Sofern Sammlungen insgesamt frei sind, sollten sie über geeignete Mechanismen zum Download oder Schnittstellen zur Verfügung gestellt werden. Beispiele für solche Bereitstellungen sind die Angebote von Datenbank-Dumps bei Wikidata[68], der Deutschen Nationalbibliothek[69] oder auch des Deutschen Textarchivs[70]; Schnittstellen bieten z. B. der Wikidata Query Service[71] oder Zenodo[72]. Für die Verlagswirtschaft, die vor allem im Bereich Textdaten tätig ist, sollten Texte zum Zwecke der Sammlungsaggregation möglichst in strukturierten Datenformaten wie JATS über entsprechende APIs bereitgestellt werden.[73] Gleiches gilt für Anbieter aus dem öffentlichen Bereich wie Bibliotheken, die Texte anbieten. Gerade für mit öffentlichen Mitteln finanzierte Einrichtungen sollten Sammlungen sowohl in technischer als auch rechtlicher Sicht ungehindert und kostenfrei nutzbar sein.

Die digitale Sammlung im Sinne des Alleinbesitzes eignet sich nicht mehr wie die analoge Sammlung als gewissermaßen ›USP‹ einer Kultureinrichtung. Im Gegenteil, wenn die leichte Zugänglichkeit und Beschaffbarkeit (Kopierbarkeit) ein wesentliches Kriterium von Qualität einer *born digital*-Sammlung ist, dann ist die Unikalität einer Sammlung an einem Ort[74] eher als disqualifizierendes Merkmal zu werten.[75] Umgekehrt gewinnen die Qualitätssicherung, die Fähigkeit zur Aufbereitung und Bereitstellung bzw. Kuratierung von Texten und Daten bis hin zum Forschungsdatenmanagement für Institutionen an Bedeutung.

4.1.4 Infrastrukturen

Sowohl zum Aufbau als auch zum Management digitaler Sammlungen braucht es verschiedene Akteure. Diese sind ausführlich in Kapitel 4.2 beschrieben. Zusätzlich spielen insbesondere bei der Digitalisierung von Sammlungen auch private Dienstleister und Firmen eine Rolle, die Bibliotheken, Archive und Museen in ihren Digitalisierungsaktivitäten unterstützen.

Zur Kuratierung von digitalen Sammlungen gehört eine angemessene technische Ausstattung: Dazu zählen u. a. Hochleistungsscanner, Software für Transformationsprozesse und Datenverwaltung oder Netzwerk- und Servertechnologie. Die Digitalisate werden auf Rechnern bearbeitet und müssen nach Maßgabe der Erfordernisse der Langzeitarchivierung in geeigneten Systemen (Rosetta, Archivematica, Dimag, DANRW u. a.) gespeichert werden. Für die Erschließung, die Speicherung und die Annotation von Digitalisaten bedarf es Standards, die sowohl von Maschinen als auch von Menschen genutzt werden können.

Um die Digitalisate als Sammlung zugänglich zu machen, sind über das Internet zugängliche zentrale Nachweissysteme nötig, wie die bereits erwähnte Deutsche Digitale Bibliothek, Archivportal-D, Europeana oder instituts- bzw. sammlungsspezifische Portale.

Insgesamt erfordert der Aufbau und die Kuratierung digitaler Sammlungen einen hohen Aufwand, der langfristig sowohl von einzelnen Einrichtungen, aber auch von Fachgemeinschaften erbracht werden muss. Für die Aufbauprozesse der erforderlichen Infrastrukturen ist die systematische Unterstützung durch Drittmittelgeber und ein die technische Transformation begleitendes Digital Architecture Management (DAM) unver-

67 Vgl. W3C Working Group (2013, March). *PROV-Overview An Overview of the PROV Family of Documents*. https://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

68 Vgl. *Wikidata:datenbank-download* (2021, July 9). Wikidata. https://www.wikidata.org/wiki/Wikidata:Database_download/de.

69 Vgl. data.dnb.de (n. d.). https://data.dnb.de/opendata/.

70 Vgl. Deutsches Textarchiv (n. d.). https://www.deutschestextarchiv.de/.

71 Vgl. Wikidata (n. d.). *Wikidata Query Service*. https://query.wikidata.org/ (SPARQL Schnittstelle).

72 Vgl. Zenodo (n. d.). *Zenodo Developers*. https://developers.zenodo.org/#changes (OAI-PMH Schnittstelle).

73 Vgl. Springer Nature (n. d.). *Springer Nature API Portal*. https://dev.springernature.com/.

74 Natürlich ist der Ort auch für Kopien von Sammlungen neben der engeren technischen Bereitstellung weiter von Bedeutung. Allerdings kann auf die Frage der Kontextualisierung einer Sammlung, wo es also einen Unterschied macht, ob sie sich in einem demokratischen oder nicht-demokratischen Land oder an einem kulturell diversen Standort befindet, hier aus Umfangsgründen nicht eingegangen werden.

75 Vgl. Stäcker, T. (2019). Die Sammlung ist tot, es lebe die Sammlung! *Bibliothek Forschung und Praxis*, *43*(2), S. 304–310. https://doi.org/10.1515/bfp-2019-2066.

zichtbar. Historisch gewachsene und dadurch oftmals heterogene IT-Systemlandschaften schränken die Einrichtungen bei der Realisierung des Zugriffs auf Datensammlungen ein. Die inkrementelle Weiterentwicklung der IT-Systemlandschaft und oftmals als Leuchtturmprojekte vermarktete Digitalisierungsaktivitäten der Einrichtungen sind meist aus einzelnen Fachbereichen heraus motiviert – und befördern die Heterogenisierung. Ein ganzheitliches Modell im Sinne eines übergreifenden DAM könnte Einrichtungen in die Lage versetzen, ihre IT-Infrastruktur kontinuierlich aus verschiedenen Perspektiven der Digitalen Sammlungen zu betrachten. So wird die IT-Infrastruktur der Einrichtungen unter Berücksichtigung strategischer, kultureller und technischer Aspekte systematisch weiterentwickelt und vereinheitlicht. Es entstehen (Quasi-)Standards, innerhalb derer über die Einführung und Anwendung digitaler Technologien, Prozesse und Systeme strategisch entschieden werden kann.

4.1.5 Publikation von Digitalen Sammlungen

Sowohl die Objekte in einer Sammlung als auch die gegenstandsspezifischen Klassifikationssysteme und Erschließungsinformationen (Vokabulare und Metadaten) sind Ergebnisse von Forschungsarbeit. Insofern ist auch das Konzipieren und Anlegen einer wissenschaftlichen digitalen Sammlung als wissenschaftliche Leistung zu betrachten – ebenso wie deren Veröffentlichung und langfristige Kuratierung.

Dies bringt mehrere Anforderungen und Konsequenzen mit sich. Wenn mit digitalen Sammlungen geforscht wird, bedarf es einer transparenten und gut nachvollziehbaren Referenzierbarkeit dieser Sammlungen. Dabei kann der Umfang der referenzierten Sammlungen oder Sammlungsteile beliebig groß sein.

Für die Zitierbarkeit sind digitale Sammlungen und deren Teile bzw. Einheiten grundsätzlich mit einem persistenten digitalen Identifier (PI) zu versehen. Dies kann unabhängig von der Verfügbarkeit der digitalen Sammlungsobjekte geschehen,die aus rechtlichen oder ethischen Gründen eingeschränkt sein kann. Falls die Zugriffsmöglichkeiten eingeschränkt sind, empfiehlt sich eine Angabe, unter welchen Voraussetzungen der Zugriff für wissenschaftliche Zwecke ermöglicht werden kann. Im Hinblick auf Verwertungsrechte sollte bei digitalen Sammlungen, die nicht unmittelbar im *Open Access* erscheinen und durch Verlage veröffentlicht werden, möglichst kurzfristige *Moving Walls* für die Publikation im *Open Access* vereinbart werden, um den Zugriff nicht dauerhaft zu beschränken. Auch hier sind abgestufte Modelle denkbar.

Grundsätzlich scheint es ratsam, digitale Sammlungen sowohl im Rohformat (die möglicherweise proprietär oder projektspezifisch sind) als auch in einem offenen Standardformat anzubieten. Dies sollte auch gelten, wenn Dritte an der Datenerzeugung beteiligt werden (bspw. Verlage, OCR-Dienstleister oder Forschungs-/ industrielle Partner).

Die dargelegten Maßnahmen steigern die Nachnutzbarkeit von Sammlungen im Hinblick auf die Reproduzierbarkeit von Forschungsergebnissen und erleichtern darauf aufbauende Forschung sowie die Publikation von aggregierten (und nicht selbst erstellten) Sammlungen für bestimmte Forschungszwecke. In jedem Fall sollte eine Dokumentation der Sammlung und des verwendeten Datenformats mitgeliefert werden.[76]

Parallel zur Nutzung durch Forscherinnen und Forscher ist gleichzeitig – im Sinne der FAIR-Prinzipien – für die maschinelle Zugänglichkeit und Interoperabilität Sorge zu tragen. Die Ziele unterscheiden sich dabei möglicherweise von dem unmittelbar wissenschaftlichen Zweck, wenn beispielsweise Forschungsdaten zunächst aus mehreren Quellen aggregiert oder deren Metainformationen abgeglichen werden sollen. Im Idealfall sind sowohl Sammlungs- als auch Objektdaten jedoch offen verfügbar und qualitativ durch offene Vokabulare miteinander verlinkt, so dass mithilfe von RDF-basierten Technologien ein größeres Wissensnetz geharvestet und abgefragt werden kann. Dadurch könnte an oder mit den Daten geforscht werden, ohne dass eine neue digitale Sammlung erzeugt werden müsste. Voraussetzung dafür ist die Anwendung standardisierter Protokolle für Datenschnittstellen (vgl. Kap. 3.1.2) sowie eine hinreichende Performance der Infrastrukturen (andernfalls kann dies zu Einschränkungen des Angebots führen). Ferner sollte auch die Schnittstelle den Anforderungen an wissenschaftliche Referenzierbarkeit genügen.

Die Veröffentlichung einer digitalen Sammlung kann auf unterschiedlichen Plattformen vollzogen werden (vgl. a. Kap. 3.1.2). Aus der Perspektive von Wissenschaftlerinnen und Wissenschaftler bieten sich dafür zunächst die Repositorien der eigenen Institution an. Daneben besteht eine Vielzahl fachspezifischer oder überinstitutioneller Repositorien, die grundsätzlich für Sammlungen geeignet sind.[77] An generischen (institutionell, thematisch

76 Vgl. International Organization of Standardization (2019). *ISO/IEC 11179-7:2019 Information technology – Metadata registries (MDR) – Part 7: Metamodel for data set registration* (1st ed.). Möglich ist z. B. auch die Platzierung in einem Data Journal, siehe dazu Forschungsdaten.org (n. d.). *Data journals – Forschungsdaten.org*. https://www.forschungsdaten.org/index.php/Data_Journals.

77 Vgl. Karlsruher Institute of Technology (n. d.). *re3data.org - Registry of Research Data Repositories*. https://doi.org/10.17616/R3D. Vgl. a. COAR e. V. (2021). *Building a sustainable, global knowledge commons*. COAR. https://www.coar-repositories.org/.

und fachlich ungebundenen) Angeboten sind beispielsweise Zenodo, GitHub und Figshare zu nennen. Mit der Wahl einer Plattform geht auch eine Entscheidung für das jeweilige Commitment der Anbieter und die ggf. verfügbaren Schnittstellen einher. Dies kann insbesondere für Sammlungen, deren Erstellung durch Drittmittel finanziert wurde, von strategischer Bedeutung sein.

So unabsehbar die Zielgruppe einer publizierten Sammlung erscheint, so ist sie zumindest durch ihr Thema oder ihr Fachgebiet eingrenzbar auf die jeweilige wissenschaftliche Community. Eine Nachnutzung kann jedoch auch interdisziplinär erfolgen, sofern die Fürsorge für eine fundierte Menschen- und Maschinenlesbarkeit der Datensammlung gleichermaßen gegeben ist (in Veranstaltungen wie Data-Hackathons lässt sich anschaulich das interdisziplinäre Potenzial von digitalen Sammlungen zeigen). Aus der Sicht der rein maschinellen Verarbeitung ist sicher die Einbringung in die Linked Open Data Cloud zielführend, die perspektivisch großes Potential für eine dezentral angelegte Datennutzung hat.

4.2 Management Digitaler Sammlungen

Aus den vorherigen Abschnitten haben sich eine Reihe von neuen Funktionen und Aufgaben ergeben, die sich im Management digitaler Sammlungen niederschlagen. Im Folgenden soll versucht werden einerseits diese Funktionen überblicksartig in ihrem Ineinandergreifen zu illustrieren, andererseits die neuen Aufgaben und Berufsfelder mit Blick auf die jeweiligen Akteure darzustellen, die die sich aus dem Aufbau, der Pflege und Zugänglichmachung digitaler Sammlungen ergeben.

4.2.1 Aufgaben beim Management digitaler Sammlungen

Beim Management digitaler Sammlungen fallen verschiedene Aufgaben an:

Diese einzelnen Aktivitäten sind von verschiedenen Beteiligten in unterschiedlicher Funktion in enger Kooperation durchzuführen.

4.2.2 Menschliche Akteure beim Management digitaler Sammlungen

Am Aufbau digitaler Sammlungen sind im Wesentlichen folgende menschliche Akteure mit unterschiedlichen Funktionen beteiligt:

a) **Wissenschaftlerinnen und Wissenschaftler an den Universitäten und Forschungsorganisationen:** Ein Wissenschaftler oder eine Wissenschaftlerin bzw. ein Konsortium aus solchen kann den Impuls zum Aufbau einer wissenschaftlichen Sammlung geben. Entweder soll eine bestimmte wissenschaftliche Fragestellung mit der Sammlung beantwortet werden oder ein Forschungsprojekt zielt explizit darauf ab, eine Sammlung mit einem gewissen Scope aufzubauen. Wissenschaftlerinnen und Wissenschaftler vertreten beim Aufbau einer Datensammlung den Blickwinkel der Wissenschaft. Sie wirken an der Zusammenstellung der Anforderungen an eine digitale Sammlung mit und können den Aufbauprozess mit begleiten. Ebenso können sie an der Kuration der Sammlung – bspw. an der wissenschaftlichen Beschreibung der Objekte – beteiligt sein. Oft nutzen Wissenschaftlerinnen und Wissenschaftler Datensammlungen, indem sie auf Objekte zugreifen, um wissenschaftliche Fragestellungen zu bearbeiten.

b) **Software-Entwicklerinnen und -Entwickler**: Um die technischen Voraussetzungen für die Nutzung einer digitalen Sammlung zu schaffen, bedarf es Software-Entwicklern. Diese kümmern sich in Zusammenarbeit mit einem Administrator oder Administratorin um die technische Implementierung der Datenbank, modifizieren diese entsprechend und entwickeln gemeinsam mit Wissenschaftlerinnen und Wissenschaftlern und Mitarbeiterinnen und Mitarbeitern der jeweiligen Dateninfrastruktur die angestrebte Lösung. Später können Änderungen, Anpassungen und Verbesserungen durch diese notwendig sein.

c) **Administratorin und Administrator:** Diese sind für den Betrieb der Datensammlung in einer sicheren IT-Umgebung zuständig. Sie sorgen dafür, dass die IT-Sicherheit und der Zugang gewährleistet sind.

Neben den beschrieben Funktionen tauchen vermehrt Begriffe wie ›**Datenarchivar‹, ›Datenmanager‹ ›Data Librarian‹** oder ›**Data Steward‹** auf. Diese neuen Begriffe bezeichnen Fachpersonal aus den Forschungsinfrastrukturen oder der Wissenschaft, das speziell für den Aufbau und den Betrieb datenbezogener Dienstleistungen qualifiziert ist. Die Zukunft wird zeigen, welches Profil diese Tätigkeit genau gewinnt und welche Begriffe sich langfristig durchsetzen. In der derzeitigen Praxis findet der Aufbau von z. B. wissenschaftlichen Datenbanken meist im Zusammenwirken von Wissenschaftlern und Software-Ingenieuren statt, wobei zunehmend deutlich wird, dass informationswissenschaftliche Kompetenzen (etwa im Bereich der Metadaten und Standardisierung) hinzutreten müssen und wesentlich für die Nachhaltigkeit eines Angebots sind.

4.2.3 Institutionelle Akteure beim Management digitaler Sammlungen

Auch Organisationen übernehmen als institutionelle Akteure zentrale Aufgaben beim Aufbau und Betrieb wissenschaftlicher Sammlungen.

- **Förderorganisationen**: Förderorganisationen finanzieren durch die Förderung wissenschaftlicher Projekte, den Aufbau von wissenschaftlichen Sammlungen.
- **Universitäten und außeruniversitäre Forschungsorganisationen/Forschungsinstitute**: An Universitäten und außeruniversitären Forschungsorganisationen werden wissenschaftliche Sammlungen aufgebaut. Hierzu arbeiten die Wissenschaft mit der Software-Entwicklung, Administration und Datenkuration eng zusammen. Sammlungen werden langfristig von Universitäten und Forschungsorganisationen betrieben. Weiterhin können Universitäten und Forschungsorganisationen Träger und Betreiber von Datenzentren und Datenarchiven sein. Dies sind Einheiten, die digitale Objekte sammeln und für die Nachnutzung zur Verfügung stellen.
- **Wissenschaftliche Bibliotheken**: Wissenschaftliche Bibliotheken (Landes- und Hochschulbibliotheken sowie Spezialbibliotheken) sind wichtige Akteure beim Aufbau und bei der nachhaltigen Bereitstellung digitaler Sammlungen. Sie sammeln Printbestände, die sie digitalisieren, oder born digital Daten und Texte, die sie zur Nachnutzung in digitalen Sammlungen aufbereiten, nach eingeführten Standards erschließen und anbieten. Sie hosten fachliche Sammlungen, die von Wissenschaftlerinnen und Wissenschaftlern in ihrer Einrichtung erstellt wurden und kümmern sich um die Langzeitarchivierung. Oftmals sind sie auch

erster Einstiegspunkt in eine digitale Sammlung, die über den Webauftritt der Bibliothek zugänglich gemacht wird.

- **IT-Einrichtungen der Universitäten und Forschungsorganisationen/Forschungsinstitute, IT-Einrichtungen der Bibliotheken, Archive und Museen:** Für den Aufbau und den Betrieb einer digitalen Sammlung sind verschiedene IT-Dienstleistungen notwendig. Dazu gehören die technische Administration und der Betrieb von Servern, die Bereitstellung von Speicherplatz die Langzeitarchivierung und zuverlässige Backup-Strategie sowie permanentes Monitoring (vgl. oben Kap. 4.1.5). All diese Dienstleistungen werden von IT-Abteilungen innerhalb der Universitäten, Forschungsorganisationen bzw. spartenspezifischen IT-Einrichtungen erbracht.
- **Datenzentren**: Datenzentren sind eigenständige Institutionen, die darauf ausgelegt sind, Daten disziplinübergreifend zu sammeln, zu erschließen und zur Nachnutzung langfristig zur Verfügung zu stellen. Sie zeichnen sich durch geeignete technische Ausstattung und entsprechendes Personal als Teil nationaler und internationaler Infrastrukturprojekte im Bereich Forschungsdatenmanagement aus. Um qualitative und zuverlässige Services anzubieten und dies nachzuweisen, unterziehen sie sich oft einer Akkreditierung.

Je größer die digitale Sammlung, desto mehr Personen sind an ihrem Aufbau und Erhalt beteiligt. Die oben beschriebenen Aufgabenteilung wird in den meisten Fällen an die aktuelle gegebene Situation angepasst werden müssen. Bei kleineren Projekten werden eher weniger Personen die Aufgaben übernehmen müssen, wohingegen bei größeren Projekten auch eine weitere Ausdifferenzierung als hier beschrieben denkbar ist, wie z. B. bei der Europeana.[78]

5 Empfehlungen

Wie schon in der Einleitung dargelegt, richtet sich die Handreichung einerseits an Einrichtungen und Personen, die digitale Sammlungen aufbauen, verwalten und kuratieren wollen, andererseits aber auch an alle, die digitale Sammlungen nutzen, begutachten, beforschen oder deren Aufbau und deren Kuratierung finanzieren und sich dazu über Grundbegriffe sowie über die gängige Praxis verständigen und informieren wollen. Die nachstehenden Empfehlungen greifen noch einmal einzelne Aspekte der Darstellungen heraus. Allgemeine Empfehlungen, wie die FAIR und CARE Prinzipien, aber auch solche der guten wissenschaftlichen Praxis, wie im Kodex der DFG gefordert, werden als selbstverständliche Grundlage vorausgesetzt:

- Denken Sie daran, dass digitale Sammlungen als eine besondere Form von Forschungsdaten (s. die Definition in der Einleitung) auf Nachhaltigkeit ausgelegt sind und bedenken Sie beim Aufbau einer Sammlung die relevanten zu beteiligenden Akteure.
- Machen Sie sich Synergien in der Zusammenarbeit von Forschungsabteilungen, Archiven und Bibliotheken im Bereich der persistenten Bereitstellung und Langzeitarchivierung digitaler Sammlungen zunutze.
- Sorgen Sie, ggf. über Kooperationen mit Infrastrukturanbietern und IT-Experten, für eine nachhaltige personelle und budgetäre Absicherung für die Einrichtung und Wartung moderner und hoch komplexer Infrastruktursysteme, um langfristige Verfügbarkeit und Verlässlichkeit sicherzustellen.
- Verwenden Sie, ggf. ergänzend zu proprietären Metadaten, möglichst spartenübergreifende und disziplinunabhängige Metadatenformate und Standards. Digitale Sammlungen sind ohne Metadaten wertlos, sowohl für die menschliche als auch die maschinelle (Nach-)Nutzung. Zugleich hängt der Mehrwert von digitalen Sammlungen an der Nutzung einheitlicher und möglichst generischer Metadaten und Standards.
- Betrachten Sie Ihre Sammlung als mögliche Komponente im *semantic web*. Entwickeln Sie ein weitreichendes Linked-Open-Data-Angebot für die jeweils eigenen digitalen Sammlungsobjekte und verlinken Sie sie mit dem Linked-Open-Data-Network.
- Bedenken sie Anforderungen an das Datenformat und an Schnittstellen, die für die Analyse Ihrer Daten notwendig sind (z. B. für Text- und Datamining)
- Nutzen Sie vorhandene offene Forschungsinfrastrukturen (z. B. FID, NFDI) und greifen auf bereits entwickelten Portale, Repositorien und Tools zurück
- Technische Lösungsansätze wie beispielsweise die Speicherung von Datensammlungen in verteilten Systemen für die gemeinsame Nutzung sollten bereits während der Entwicklung auf ihren Anwendernutzen überprüft werden. Dazu sind die Methoden der IT-Sicherheit, Schnittstellendefinitionen und der Kuratierung gemeinsam integrativ heranzuziehen

78 Vgl. Europeana Foundation (n. d.). *Europeana Foundation staff*. https://pro.europeana.eu/page/office-employees.

- Entwickeln Sie Kriterien für die Evaluation der Qualität Ihrer digitalen Sammlung und richten Sie das Qualitätsmanagement danach aus: Für Szenario X ist diese Datensammlung von ausreichender | bedingt ausreichender | nicht ausreichender Qualität‹.
- Entwickeln Sie standortgerechte Handreichungen zur praktischen Umsetzung der FAIR- und CARE-Prinzipien für digitale Sammlungen. Dies sollte an den konkreten Bedarfen und Möglichkeiten einer Einrichtung, einer Abteilung oder eines Projekts, ggf. auch eines Forschungsverbunds formuliert werden
- Sorgen Sie für »Data Literacy« des wissenschaftlichen Personals, erstens um eigene digitale Sammlungen inhaltlich kuratieren zu können, zweitens um die Nachnutzung oder Auswertung von eigenen oder anderen Sammlungen anzuregen, durch z. B. gezielte Fortbildungen oder durch Gewinnung qualifizierten, dauerhaften Personals. Dazu gehören auch die wissenschaftliche Akkreditierung einer Sammlungsautorenschaft sowie die Herausbildung eines Rezensionswesens
- Erhöhen Sie die Sichtbarkeit Ihrer digitalen Sammlungen durch Citizen-Science-Initiativen, beispielsweise durch digitale und hybride Ausstellungen sowie Engagement-Angebote wie z. B. Hackathons oder Transcribathons, insbesondere in Zusammenarbeit mit Galerien und Museen, für den Anschluss der Wissenschaft an ein breiteres Publikum.
- Nutzen Sie möglichst offene Lizenzen. Ein Erfolgsfaktor für die Nutzung digitaler Sammlungen ist ihre Fähigkeit, zu qualitativ und quantitativ neuen Sammlungen aggregiert, selektiert und modifiziert werden zu können. Dafür müssen ggf. komplette oder Teile von Sammlungen herausgezogen, zusammengeführt, strukturiert, überarbeitet und neu geordnet werden. Lizenzen, die diese Bearbeitungsschritte erschweren oder gar unmöglich machen, sollten daher vermieden werden. So kann z. B. eine gut gemeinte CC BY-NC Lizenz dazu führen, dass die Daten nicht mit einer anderen Sammlung vereinigt werden kann, die unter CC BY-SA lizenziert wurde. Empfohlen wird daher, bei allen digitalen Sammlungen, deren Einzelelemente selbst nicht urheberrechtlich geschützt sind, wie Artikel, Monographien, aber auch Datenbanken bzw. Datensammlungen, eine Lizenz zu nutzen, die keinerlei Einschränkungen in der Nachnutzung vorsieht. Die entstehende Sammlung selbst kann und sollte dessen ungeachtet gemäß guter wissenschaftlichen Praxis attribuiert werden. Eine Relizenzierung[79] ursprünglich freier Daten sollte vermieden werden, auch wenn diese eine eigene wissenschaftliche und ökonomische Leistung darstellt, um eine offene Weiternutzung der Daten, gleich auf welcher Aggregationsstufe, zu erleichtern.
- Machen Sie Daten, die unter dem Gesichtspunkt von Datenschutz oder Schutz der Persönlichkeit relevant sind, soweit es unter Beachtung der jeweiligen Rechte möglich ist, durch geeignete Maßnahmen (z. B. Anonymisierung) zugänglich.

79 Vgl. CORE (n. d.). *CORE: The world's largest collection of open access research papers*. https://core.ac.uk/.