

## 4.2 Datenspeicherung, -kuration und Langzeitverfügbarkeit

**Abstract:** Die langfristige Speicherung von Daten und deren nachhaltige Verfügbarkeit sind schon seit langem wichtige Desiderate in der Forschung. Experimente, Messungen, Simulationen oder Auswertungen liefern Daten, für die optimale Aufbewahrungsmöglichkeiten gefunden werden müssen. Längst stellt die Forschung Anforderungen, die über die „reine“ Speicherung der Daten hinausgehen. Dabei gibt es keine universelle Methode, sondern für jedes Forschungsvorhaben muss die geeignete Vorgehensweise gefunden werden. Je nach Bewertung der Forschungsdaten können verschiedene Erhaltungsstrategien angewandt werden, die wiederum unterschiedliche Anforderungen an die Art des physischen Speichers und den Zugriff haben. Zur Bewertung der Forschungsdaten spielen Selektion und Kuration daher eine immer wichtigere Rolle. Einerseits geht es dabei um die Auswahl und Klassifizierung der Daten, die langfristig aufgehoben werden sollen. Andererseits rückt im Kontext der Nachnutzbarkeit auch die geeignete Beschreibung der Daten und des Entstehungskontextes in den Fokus. Das Ziel einer langfristigen Verfügbarkeit und Interpretierbarkeit ist ohne die ausreichende Beschreibung der Rohdaten nicht erreichbar. Damit stellt die Langfristarchivierung auch die IT-Systeme und deren Architektur vor Herausforderungen. Ein möglicher Lösungsansatz ist das „Open Archival Information System“-Modell (OAIS-Modell) als Referenz zur Implementierung der Langzeitarchivierung digitaler Objekte. Der vorliegende Artikel beschreibt die unterschiedlichen Möglichkeiten der Speicherung von Forschungsdaten, erklärt was im Kontext von Langfristverfügbarkeit Datenkuration bedeutet und beschreibt das OAIS-Referenzmodell. Schließlich werden im Abschnitt Praxistransfer praktische Hilfestellungen zu den Themen des Artikels gegeben.

### Einleitung

Die ältesten bislang entdeckten Höhlenmalereien finden sich in der Höhle Cueva de El Castillo bei Puente Viesgo in Spanien.<sup>1</sup> Über den Zweck der Abbildungen gibt es verschiedene Theorien, z. B. dass es sich um die Darstellung von Jagderfahrungen handelt. Stellen diese Bilder eine frühe Form von Forschungsdaten dar, in denen experimentell entwickelte Jagdmethoden dokumentiert werden? Die Zuschreibung

---

<sup>1</sup> Vgl. Cabrera Valdes 1989, 577–584.

„Forschungsdaten“ ist wahrscheinlich eine etwas gewagte These. Jedoch stellen die Abbildungen im Rahmen der Erforschung und Entdeckung des eigenen Lebensraums ein gutes Beispiel für die langfristige Bewahrung von Menschen geschaffener Informationen dar. Schaut man weiter nach Beispielen langfristig erhaltenen Wissens, findet man in der Antike die in hieroglyphischer Schrift überlieferten Informationen vergangener Kulturen am Nil. Gleichzeitig lässt sich eine wichtige Voraussetzung langfristiger Verfügbarkeit daran gut verdeutlichen. Die Entdecker der Hieroglyphen konnten die einzelnen Zeichen erkennen, deren Entzifferung gelang aber erst viele Jahre später durch das Sprachgenie Jean-François Champollion.<sup>2</sup> Den Schlüssel zum Erfolg bildete der Stein von Rosetta,<sup>3</sup> in den ein Text über den König Ptolemaios in drei Sprachen, darunter auch in Hieroglyphen, eingemeißelt ist. Aufgrund der Mehrsprachigkeit der dargestellten Ereignisse wurden eine Interpretation und damit ein sprachliches Verständnis der hieroglyphischen Zeichen möglich. Damit konnte dann durch mühsame Vergleichsarbeit der Texte die Interpretation der Hieroglyphen abgeleitet werden. Man kann erkennen: Der alleinige Erhalt von Daten ist für eine spätere Interpretation oder Wiederverwendung nicht ausreichend. Es muss auch die Information erhalten werden, wie die Daten zu interpretieren sind. Das unterscheidet die eigentliche Datenspeicherung von der Langfristspeicherung von Daten. Insbesondere in der heutigen Zeit der digitalen Daten muss dieses Problem bei der Langzeitarchivierung mitgedacht werden, denn sowohl die Lesbarkeit als auch die Interpretation digitaler Daten hängt von speziellen Anwendungen ab. Dies betrifft sowohl den Hardware- als auch den Anwendungskontext digitaler Daten. Die meisten Daten werden heute in einer Form gespeichert, die einen Zugriff auf die Information nur über technische Hilfsmittel erlaubt, die im Falle einer langen Aufbewahrungsperiode möglicherweise veraltet, nicht mehr nutzbar oder gar zerstört sein könnten.

Für die Sicherung von bedeutendem Kulturgut geht man deshalb einen besonderen Weg. So werden z. B. als Bundesaufgabe im Rahmen des Zivilschutzes<sup>4</sup> seit 1961 wichtige Archivalien mikroverfilmt<sup>5</sup> und die Filme in Spezialbehältern im Barbarastollen in Oberried bei Freiburg im Breisgau eingelagert.<sup>6</sup> Der Zugriff auf die Information kann mit Hilfe einer Lupe und einer Lichtquelle ohne weitere technische Hilfsmittel gewährleistet werden, solange die Information über die Interpretation der abgelichteten Sprachen nicht verlorengegangen ist. Neben dem Mikrofilm wer-

---

<sup>2</sup> Vgl. Majonica 2007.

<sup>3</sup> Vgl. Depuydt 1999, 686–687.

<sup>4</sup> Vgl. BMI 1987, 284–292.

<sup>5</sup> S. [https://www.bbk.bund.de/DE/AufgabenundAusstattung/Kulturgutschutz/Sicherungsverfilmung/sicherungsverfilmung\\_node.html](https://www.bbk.bund.de/DE/AufgabenundAusstattung/Kulturgutschutz/Sicherungsverfilmung/sicherungsverfilmung_node.html). Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

<sup>6</sup> S. [http://www.bbk.bund.de/DE/AufgabenundAusstattung/Kulturgutschutz/ZentralerBergungs-ort/zentralerbergungsort\\_node.html](http://www.bbk.bund.de/DE/AufgabenundAusstattung/Kulturgutschutz/ZentralerBergungs-ort/zentralerbergungsort_node.html).

den andere persistente Speichermedien verwendet. Im Projekt Memory of Mankind (MoM)<sup>7</sup> werden die Daten z. B. auf Keramikfliesen gebrannt. Diese Spezialfliesen sind bis 1200 Grad Celsius temperatur-, säure-, laugen- und strahlungsbeständig und werden in einem Salzbergwerk in Hallstatt gelagert. Solche Verfahren sind sehr kostenintensiv, haben eine sehr geringe Speicherdichte und sind somit nur für besonders wichtige Informationen sinnvoll einsetzbar. Der Auswahl der zu speichernden Informationen im Rahmen der Datenkuration kommt hier eine besonders wichtige Rolle zu.

Die Forderung nach der Überprüfbarkeit von Forschungsergebnissen führt dazu, dass die langfristige Speicherung von Daten nicht nur für einzelne Forschende ein zentrales Thema geworden ist, sondern auch für Forschungseinrichtungen. Dabei spielt die Kuration von Daten im Forschungsprozess eine immer größer werdende Rolle. Der Rat für Informationsinfrastrukturen (RfII) hat in seinen Empfehlungen<sup>8</sup> die Schaffung neuer Berufsbilder, wie etwa von Datenkuratoren, sogar ange-regt.

## 1 Datenspeicherung

Grundsätzlich sind Speichermedien durch Zugriffszeit, Datenrate und Speicherkapazität charakterisiert. Die Anforderungen an die Speichermedien in der Forschung haben eine große Bandbreite und sind je Anwendungsfall verschieden. So ist es offensichtlich, dass die Speicherung von Daten aus komplexen Experimenten, z. B. Kernfusionsexperimenten, wo pro Plasmaeinschluss in weniger als einer Sekunde mehrere 100 Megabyte an Daten entstehen, sich von der Speicherung von Auswertungen von sozialwissenschaftlichen Umfragen deutlich unterscheidet.

Bei den derzeit vorhandenen Speichertechniken besteht ein prinzipieller Konflikt zwischen der Minimierung der Zugriffszeit und der Maximierung der Speicherkapazität. In großen Rechner- und Speichersystemen werden deshalb unterschiedliche Speichertechnologien zu Speicherhierarchien kombiniert, um einen Kompromiss aus schnellem Zugriff und großen Speicherkapazitäten bei angemessenen Kosten zu erreichen. Um Forschende bei der Planung ihrer Datenhaltungsstrategie beraten zu können, müssen die aktuellen technischen Möglichkeiten und die zukünftigen Entwicklungen bekannt sein. Nachfolgend werden diese deshalb kurz beschrieben und weitere Implikationen betrachtet.

---

<sup>7</sup> MoM: Wie bewahrt man Information dauerhaft für 1 Million Jahre auf? Vgl. <https://www.memory-of-mankind.com/de/how-is-information-kept-legible-for-1-million-years/>.

<sup>8</sup> S. RfII 2016, <http://www.rfii.de/?wpdmdl=1998>.

## 1.1 Speichermedien in der EDV

Grundsätzlich lassen sich die Speichermedien in Rechnersystemen in Primär-, Sekundär- und Tertiärspeicher<sup>9</sup> unterscheiden.<sup>10</sup> Als Primärspeicher werden alle Speicher mit wahlfreiem Zugriff (Random Access) bezeichnet, auf die der Prozessor direkt mit voller Geschwindigkeit zugreifen kann. Dazu zählen Register eines Prozessors, Caches und der Hauptspeicher (Main Memory bzw. Arbeitsspeicher). Primärspeicher bieten sehr schnellen Zugriff im Nanosekunden-Bereich, sind aber hinsichtlich ihrer Kapazität begrenzt.

Hintergrundspeicher mit index-sequentiell (quasi-wahlfreiem) Zugriff, wie magnetische Festplatten oder RAID-Systeme (Redundant Arrays of Independent Disks) werden als Sekundärspeicher bezeichnet. Diese Speicher verfügen im Vergleich zu Primärspeichern über große Kapazitäten, weisen allerdings einen um den Faktor  $10^6$  langsameren Zugriff auf.<sup>11</sup> Dieser gravierende Unterschied in der Zugriffszeit wird auch als Zugriffslücke bezeichnet.

Speichertechnologien mit Speichermedien, auf die nicht direkt zugegriffen werden kann, gliedern sich in die Kategorie der Tertiärspeicher ein. Nicht direkt zugreifbar bedeutet, dass Medien manuell bedient werden müssen oder in robotergesteuerten Bibliotheken organisiert sind. Erst beim Zugriff auf die Daten werden diese in die entsprechenden Schreib-/Lesegeräte bewegt. Die Zugriffslücke zwischen Sekundär- und Tertiärspeicher erreicht ebenfalls einen Faktor von bis zu  $10^6$ . Dabei haben sich heute drei Tertiärspeichertechnologien etabliert: magneto-optische Speicher, optische Speicher und Magnetbänder. Magneto-optische und optische Speicher werden überwiegend bei kleinen bis mittleren Datenmengen (GByte bis TByte) eingesetzt und wenn schneller Zugriff erforderlich ist. Für die Speicherung sehr großer Datenmengen (TByte bis PByte) werden vor allem Magnetbänder verwendet.

Seit Beginn der Speicherung von Daten in Computersystemen hat es eine positive Entwicklung sowohl bei den Zugriffszeiten als auch im Bereich der Speicherdichte gegeben. Bei der Kapazität konnte in den letzten 40 Jahren eine Steigerung der Speicherdichte um den Faktor 25 Milliarden erreicht werden,<sup>12</sup> wenn man die ersten Lochkarten mit einer heutigen SD-Karte vergleicht. Die Zugriffszeiten wurden ebenfalls gesteigert, wobei hier oft der Vorteil durch die Verwendung von schnellem Zwischenspeicher (Caches) entsteht, die die Zugriffslücken überbrücken können. Die Entwicklung von Speichermedien wird aber auch von der Verbesserung der Robustheit der Speichermedien gegenüber Datenverlust, Ermüdungserscheinungen

<sup>9</sup> In der Literatur wird manchmal der Tertiärspeicher nicht explizit aufgeführt, sondern zu den Sekundärspeichern gezählt.

<sup>10</sup> Vgl. TG 2001.

<sup>11</sup> Vgl. Hennessy 2007, 359.

<sup>12</sup> S. <https://wkdiscpress.de/ratgeber/chronik-der-speichermedien/>.

des Materials und der Kosten für die Speichermedien geprägt. In der Geschichte gab es dabei auch durchaus kuriose Entwicklungen. Im Jahr 1998 wurde z. B. die Möglichkeit der Speicherung von Daten auf handelsüblichem Tesafilm vorgestellt,<sup>13</sup> bei der bis zu zehn GByte auf einer Rolle Tesafilm abgelegt werden konnten. Das Speichermedium wäre in diesem Fall als günstiges Massenprodukt zu kaufen gewesen. Jedoch hat sich als sogenannter WORM-Speicher<sup>14</sup> dann aber die DVD durchgesetzt. Die Chancen für weitere Leistungssprünge im Speicherbereich stehen gut, denn neue Technologien sind bereits in den Startlöchern oder befinden sich bereits im Einsatz. Ein paar aktuelle Entwicklungen werden nachfolgend kurz vorgestellt.

Phase-Change-Memory-Chips (PCM-Chips) basieren auf chemischen Verbindungen, die mit ihrer Struktur ihre elektrische Leitfähigkeit verändern können.<sup>15</sup> Die durch einen starken Stromimpuls verursachte starke Erhitzung verändert die Ordnung der Moleküle und damit den Widerstand. Erneute geringe Stromzufuhr führt wieder zum Originalzustand in der Leitfähigkeit. Somit können binäre Informationen gespeichert werden. Diese Technik ist vor allem in Smartphones bereits im Einsatz. Die Vorteile der PCMs gegenüber Flash-Speichern liegen in der günstigeren Herstellung und in der etwa fünfmal höheren Speicherdichte. Der größte Vorteil liegt aber darin, dass PCMs mehr als zehn Millionen Mal beschrieben werden können, wohingegen die Garantie herkömmlicher Flash-Speicher nach einigen 1000 Schreibvorgängen erlischt.

Bei dem von IBM entwickelten „Racetrack“-Speicher werden, ähnlich wie bei Magnetbändern, digitale Daten in einer Reihe von magnetischen Domänenwänden (DWs) gespeichert. Im Unterschied zu Magnetbändern werden diese jedoch in Nanodrähten gespeichert, die in einem 3D-Array angeordnet sind.<sup>16</sup> Der Betrieb eines „Rennstreckenspeichers“ beruht darauf, dass die DWs entlang der Nanodrähte mit bis zu 2000 Metern pro Sekunde bewegt werden können, indem ein Strom durch den Draht geleitet wird. Da dabei nur Elektronen bewegt werden, können die Daten etwa 100 000 Mal schneller gelesen werden als von heutigen Festplatten und es gibt auch fast keine mechanische Abnutzung. Die Drähte haben nur einen Durchmesser im Nanometerbereich, so dass sich etwa 180 000 Drähte auf der Breite eines Menschenhaares unterbringen lassen. Somit könnten auf mobilen Endgeräten mehrere tausend Filme gespeichert werden. Wegen des geringen Energiebedarfs können die Speicher wochenlang mit einer Akkuladung laufen und hätten eine quasi unendlich lange Lebensdauer. Es gibt derzeit noch keine Umsetzung, die eine Massenproduk-

<sup>13</sup> S. <https://www.spektrum.de/news/tesafilm-als-datenspeicher/341007>.

<sup>14</sup> WORM Speicher sind Speichermedien, die nur einmalig schreibenden (Engl. write once) aber mehrfach lesenden Zugriff (Engl. read many) erlauben.

<sup>15</sup> Vgl. Ovshinsky 1968.

<sup>16</sup> Vgl. Parkin 2008.

tion erlaubt. Grundsätzliche Fragen, wie z. B. die Genauigkeit der Positionierung von DWs sind Gegenstand aktueller Forschung.<sup>17</sup>

Im Bereich der Langzeitdatenspeicherung von sehr großen Datenbeständen werden derzeit ebenfalls interessante neue Technologien entwickelt. Ein Beispiel ist die Erforschung der Möglichkeit, DNA-Material als Speicherbaustein zu verwenden.<sup>18</sup> Damit wären sehr hohe Speicherdichten von etwa ein Exabyte/mm<sup>3</sup> (10<sup>9</sup> GB/mm<sup>3</sup>) zu erreichen. Zudem ist die Speicherung mit DNA sehr langlebig (Halbwertszeit etwa 500 Jahre im Vergleich zu 30 Jahren für Magnetbänder). Die bislang erreichten Verfahren sind noch langsam, skalieren nicht und sind zudem sehr teuer. Aber neuere Entwicklungen lassen marktreife Verfahren in den nächsten Jahren erwarten.<sup>19</sup> Mit dieser Technologie könnten künftig die Informationsinhalte ganzer Rechenzentren in etwa eine Handfläche passen.

Es besteht also die Aussicht, dass es auch zukünftig geeignete Medien geben wird, um die immer größer werdende Masse an Informationen adäquat zu speichern. Trotz steigender Speicherkapazitäten und neuer Speichertechnologien wird es aber auch in Zukunft eine Herausforderung sein, Daten strukturiert und veränderungssicher zu speichern.

## 1.2 Verwaltung von Daten auf den Speichermedien

Grundsätzlich muss ein Speicher, von dem Informationen gelesen oder auf den Bits und Bytes geschrieben werden sollen, in irgendeiner Form organisiert werden. Dazu werden die Speicherbereiche auf den Medien mit Hilfe verschiedener Methoden in Einheiten aufgeteilt. Bei einem Blockspeicher werden die Speichereinheiten in Blöcken bestimmter Größe bereitgestellt, die durch die Anwendung angesprochen werden. Diese Zugriffsart wird z. B. von Datenbankanwendungen verwendet. Die derzeit gängigste Methode, Speicherplatz zu verwalten, sind Filesysteme. Dabei werden die Daten in Dateien organisiert, die in hierarchischen Dateisystemen abgelegt werden. Für den Zugriff auf die Informationen benötigt man den Pfad zu der Datei im hierarchischen Dateibaum. Die Organisation dieser Strukturen zeigt Limitierungen hinsichtlich der Erweiterungen und der möglichen Dateigrößen. Die Erweiterung über beliebig viele Speichermedien (einzelne Geräte) ist nicht ohne Probleme möglich, da z. B. die Adressierung über die Verzeichnishierarchie nur endliche Speicherkapazität zulässt.

---

<sup>17</sup> Vgl. Mohamed 2020.

<sup>18</sup> Vgl. Clelland et al. 1999.

<sup>19</sup> S. [https://www.wissenschaft-aktuell.de/artikel/Fehlerfreier\\_Datenspeicher\\_aus\\_DNA\\_Molekulen1771015590328.html](https://www.wissenschaft-aktuell.de/artikel/Fehlerfreier_Datenspeicher_aus_DNA_Molekulen1771015590328.html).

Als Lösung dieser Problematik ist die neue Organisationsform des Objektspeichers<sup>20</sup> von den großen Datenanbietern im Internet (Cloudspeicher) eingeführt worden. Objektspeicher organisieren Daten weder in hierarchisch angeordneten Verzeichnisbäumen mit Ordnern und Unterordnern noch in Form des Zugriffs auf die kleinsten Speichereinheiten (den Blöcken) bereitgestellt. Stattdessen fassen sie Daten inklusive ihrer externen Dateiattribute, inhaltsbezogenen Metadaten und applikationsspezifischen Parametern zu einem dezidierten Objekt zusammen. Das Objekt wird mit einer eindeutigen Objekt-ID versehen, die aus dem Datei-Inhalt und den Metadaten berechnet wird. Über diese ID ist das Objekt unabhängig vom eigentlichen Speicherort erreichbar.

Der Vorteil dieser Speicherorganisation ist die einfache und beliebige Erweiterung des Speicherplatzes. Der Zugriff auf die Daten erfolgt über ein Application Programming Interface (API) und über URL. Leider gibt es bisher dazu noch keine einheitliche Normierung. Die Ansprache der gespeicherten Informationen über einen weltweit nutzbaren Identifier stellt jedoch für die Zukunft in Aussicht, dass Daten einfach über den Aufruf einer URL genutzt werden können, ohne das darunterliegende Speichersystem lokal vorhalten zu müssen.

Die Möglichkeit der Speicherung von Metadaten als direkte Annotation zu den Daten bietet aber auch die Chance, die deskriptiven Metadaten schon in die Dateiablage zu integrieren. Damit wäre die Beschreibung von Daten mit Metainformationen unabhängig von zusätzlichen externen Systemen zur separaten Speicherung dieser Metainformationen denkbar. Das könnte eine wesentliche Vereinfachung bei der Beschreibung von Forschungsdaten darstellen.

Die fehlende Standardisierung dieser Speicherform stellt derzeit noch eine Hürde für den praktischen Einsatz im Bereich des Forschungsdatenmanagements (FDM) dar. Jedoch kristallisiert sich die von Amazon entwickelt „Simple Storage Service“-Schnittstelle (S3-Schnittstelle) als potentieller Kandidat dafür heraus.<sup>21</sup> In naher Zukunft sind hier wegweisende Entwicklungen zu erwarten. Der Objektspeicher könnte eine Lösung für die Problematik bei der Beschreibung von Forschungsdaten werden.

### 1.3 Datensicherheit

Auch wenn die Ausfallsicherheit von Speichersystemen durch redundanten Aufbau, z. B. bei RAID-Systemen, immer höher geworden ist, gibt es immer die Möglichkeit des Versagens technischer Geräte. Auch wenn die Verwendung von Cloud-Speichern, bei denen die redundante Speicherung Grundlage der Architektur ist, einen

<sup>20</sup> Vgl. Factor 2005.

<sup>21</sup> Vgl. [https://www.theregister.co.uk/2016/07/15/the\\_history\\_boys\\_cas\\_and\\_object\\_storage\\_map/](https://www.theregister.co.uk/2016/07/15/the_history_boys_cas_and_object_storage_map/).

Verlust der Daten immer unwahrscheinlicher macht, gibt es mittlerweile andere Gefahren für die Daten. In vernetzten Systemen ist es denkbar, dass Daten von anderen verändert werden. Diesen Manipulationen oder auch Fehlern des Speichermediums (z. B. Verlust der Remanenz bei magnetischen Medien oder Materialzersetzung bei optischen Speichermedien) kann man durch Prüfsummenmethoden entgegen treten. Die Bitstream Preservation<sup>22</sup>, also die Kontrolle der Beibehaltung der ursprünglichen Bitfolgen, ist deshalb als grundlegende Erhaltungsstrategie<sup>23</sup> ein Bestandteil aller Systeme, die für die langfristige Speicherung von Daten im Einsatz sind.

Aber nicht nur technische Aspekte sind bei der Speicherung von Daten zu beachten. Insbesondere bei der Verarbeitung von schützenswerten Daten, z. B. personenbezogenen Daten, stehen weitere Sicherheitsaspekte im Vordergrund. So ist die Speicherung personenbezogener Daten in Cloud-Speichern an besondere Anforderungen des Speichers gebunden.<sup>24</sup> Somit bestehen im Bereich der Forschung mit personenbezogenen Daten weiterhin Risiken bei der Inanspruchnahme von IT-Dienstleistungen und Cloud-Diensten. Diese Problematik ist bei der Veröffentlichung von Daten ein besonderes Problem, da alleine die Anonymisierung der Daten nicht ausreicht, z. B. bei soziokulturellen oder ethnischen Forschungen.

## 2 Datenkuration

Wie die bisherigen Beispiele langfristiger Aufbewahrung zeigen, ist eine vollumfängliche, langfristige Aufbewahrung von Informationen eine kostenintensive Kulturaufgabe. Daher ist die Kuration von Forschungsdaten im Kontext langfristiger Speicherung und Verfügbarkeit unabdingbar. Sie beinhaltet im Wesentlichen vier Aufgabenbereiche hinsichtlich der aufzubewahrenden Daten:

- Selektion,
- Standardisierung/Normalisierung,
- Annotation archivierungsrelevanter Informationen durch Metadaten,
- Lizenzvergabe.

---

<sup>22</sup> S. [http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor\\_handbuch\\_artikel\\_163.pdf](http://nestor.sub.uni-goettingen.de/handbuch/artikel/nestor_handbuch_artikel_163.pdf).

<sup>23</sup> S. Abschnitt 3.1.

<sup>24</sup> Vgl. Borges 2016.

## 2.1 Auswahl archivierungswürdiger Forschungsdaten

Aufgrund der stetig wachsenden Menge an digitalen (Forschungs-)Daten ist es derzeit nicht möglich und wird auch in Zukunft nicht möglich sein, alle erzeugten Daten langfristig aufzubewahren. Außerhalb der Aufbewahrungspflicht für Forschungsdaten seitens Dritter, ist daher eine Selektion der aufzubewahrenden Daten sinnvoll. Wie kann man jedoch die Bedeutung von Forschungsdaten bestimmen? Gibt es dafür quantifizierbare Kriterien? Auf diese Fragen gibt es sicherlich keine eindeutige und objektive Antwort. Jedoch kann man anhand folgender Kriterien eine gute Einschätzung über die Archivwürdigkeit und die zukünftige Bedeutung der Daten treffen. Es gilt wie überall im Leben, mit Augenmaß zu entscheiden und im aktuellen Kontext eine möglichst realistische Abschätzung zukünftiger Entwicklungen und Bedürfnisse zu machen.

Grundsätzlich sollte die erste Frage sein: Wer oder was ist das Ziel der Langzeitarchivierung der jeweiligen Forschungsdaten, wie sehen also die Anforderungen der Nachnutzenden aus bzw. gibt es überhaupt potentielle Nachnutzende? Hat diese Frage eine positive Antwort, muss überprüft werden, inwieweit die Daten in Zukunft neu generiert oder reproduziert werden können. Handelt es sich um unikale, nicht reproduzierbare Daten (z. B. Wetterbeobachtungen, Interviews mit Zeitzeugen, kontextabhängige Messungen/Beobachtungen oder historisch einmalige Aufnahmen), sollten diese eine hohe Priorität für die Langzeitarchivierung erhalten. Ebenso müssen Daten, die noch nicht vollständig wissenschaftlich untersucht sind, langfristig aufbewahrt werden. Empirische Studien (z.B. in den Sozial- und Verhaltenswissenschaften), die hinsichtlich eines bestimmten Kriteriums erhoben und ausgewertet wurden, können zusätzlich einen großen historischen Wert haben, insbesondere wenn es sich um ausgedehnte Längsschnittstudien handelt. Insofern sollte bei der Bewertung der Daten die Bedeutung für die zukünftige Forschung andere Wissenschaftsgebiete berücksichtigt werden. Im Gegensatz dazu ist es nicht unbedingt notwendig, Daten aus Standardverfahren oder Messergebnisse aufzubewahren, die immer wieder und teilweise mit besseren Verfahren in der Zukunft neu generiert werden können. Man sollte daher bei der Beurteilung der Daten auch eine Vorhersage der technologischen Weiterentwicklung versuchen und diese in die Überlegungen einbeziehen. Insbesondere in der naturwissenschaftlichen Forschung ist diese Fragestellung wichtig bei der Selektion archivwürdiger Forschungsdaten.

Unabhängig von jedem verfahrenstechnischen Einfluss muss die Qualität der Daten in die Auswahlentscheidung einbezogen werden.<sup>25</sup> Dazu sollte bei der Auswahlentscheidung eine technische und inhaltliche Qualitätsprüfung stattfinden. Die technische Qualitätsprüfung kann z. B. eine Validierung des Datenformats be-

---

<sup>25</sup> S. a. Beitrag von Kiraly und Brase, Kap. 4.3 in diesem Praxishandbuch.

inhalten oder die Prüfung zur Einhaltung vorhandener Standards im Umgang mit Dateien (Strukturierung, Dateibenennung, Auflösung bei Bildformaten).

Neben den genannten Kriterien zur Überprüfung der eigenen Daten gibt es auch noch externe, für die Forschungsbereiche spezifische Indizien, die die Bewertung der Archivwürdigkeit von Daten beeinflussen. Gibt es z. B. im wissenschaftlichen Kontext schon eine hohe Abdeckung mit publizierten, korrekten und gut dokumentierten Daten, ist es fraglich, ob zusätzliche Daten ergänzend sinnvoll oder eher redundant sind.

## 2.2 Transformation/Normalisierung der Forschungsdaten

Ist die Entscheidung zur langfristigen Aufbewahrung der Forschungsdaten getroffen, muss man sich Gedanken darübermachen, in welcher Form die Daten am sinnvollsten aufbewahrt und in Zukunft wieder genutzt werden können. Die Frage ist also: Welche Zielgruppen könnte es geben und welche Anforderungen an die Authentizität der Daten, die Struktur und den Inhalt ergeben sich daraus? Liegt das Hauptaugenmerk auf der Konservierung des Wissens respektive der enthaltenen Information oder müssen zusätzlich dazu auch Struktur und Kontext erhalten bleiben? Also muss z. B. ein proprietäres Format einer Herstellersoftware auch in Zukunft bedient werden können? Reicht die menschenlesbare Interpretierbarkeit aus (z. B. Text), um Informationen zu erhalten, oder müssen Struktur und Layout bewahrt werden (z. B. bei Präsentationen oder Designvorlagen)? Aus der Beantwortung vorgenannter Fragen ergibt sich die Notwendigkeit der Beibehaltung des Originalformats oder die Freiheit, die Dateien in sinnvolle Standardformate zu migrieren. Wichtig bei der Umwandlung in ein anderes Format ist dabei die Beibehaltung der Bedeutung der Daten. Sollte es möglich sein, die Daten vor der Archivierung in ein anderes Format zu transformieren, ohne dass notwendige Informationen verloren gehen, dann gilt: Je einfacher die Darstellung, umso besser. Konkret bedeutet dies, dass man versuchen sollte, Standarddatenformate<sup>26</sup> zu nutzen und die Daten möglichst in eine menschenlesbare und -interpretierbare Form zu bringen. Je weniger Medienwechsel zur Darstellung der Informationen notwendig sind, umso geeigneter sind die Daten für die Archivierung und eine zukünftige Nachnutzung.

Weiterhin muss man für die aufzubewahrenden Daten entscheiden, ob sie in vorliegender Fassung überhaupt gespeichert werden dürfen oder ob Vorkehrungen getroffen werden müssen, die Daten vor der Archivierung zu anonymisieren. Insbe-

---

<sup>26</sup> Eine Aufstellung archivfähiger Formate für unterschiedliche Objekttypen befindet sich im Abschnitt *Praxistransfer* (s. Abschnitt 4: Datenkuration – Normalisierung/Standardisierung).

sondere sind hier die Einhaltung des Datenschutzgesetzes und die Richtlinien zum Umgang mit sensiblen Daten zu berücksichtigen.<sup>27</sup>

## 2.3 Begleitdokumentation der Forschungsdaten

Die Grundlage der späteren Interpretierbarkeit von Archivgut ist die Beschreibung des Entstehungs- und Darstellungskontextes. Diese erfolgt im Sinne der Langzeitarchivierung durch sogenannte Erhaltungsmetadaten. Grundsätzlich sind Erhaltungsmetadaten eine Kombination oder besser gesagt ein Subset schon vorhandener Informationen aus den Metadaten zum digitalen Objekt. Sie entstammen bestenfalls den deskriptiven, strukturellen, administrativen und technischen Metainformationen.<sup>28</sup> Zu den Informationen, die zur Erhaltung notwendig sind, gehören Referenz-, Provenienz-, Kontext- und Persistenz Informationen, sowie Angaben zu Zugriffsrechten. Wichtige Standards für Erhaltungsmetadaten sind LMER<sup>29</sup> und PREMIS.<sup>30</sup>

## 2.4 Lizenzvergabe

Das Urheberrecht gilt für alle Werke mit ausreichender Schöpfungshöhe, wodurch in der Regel die Person, die die Forschungsdaten als Urheberin geschöpft hat, das gesetzliche Recht an den Daten hat. Dies beinhaltet auch die Festlegung darüber, wie die eigenen Werke (Daten) durch andere genutzt werden dürfen. Man kann sein Urheberpersönlichkeitsrecht nicht abtreten, sehr wohl aber die Nutzungsrechte an den eigenen Werken, die sogenannten Urheberverwertungsrechte. Um eine zukünftige Nachnutzung der eigenen Forschungsdaten rechtlich abgesichert zu ermöglichen, kann man daher eine Standardlizenz nutzen oder eigene Nachnutzungsbedingungen außerhalb einer solchen Standardlizenz festlegen.<sup>31</sup> Im Bereich der Forschungsdaten sind die Creative Commons Lizenzen<sup>32</sup> weitverbreitet.

<sup>27</sup> S. a. Beiträge von Lauber-Rönsberg, Kap. 1.4, sowie Rösch, Kap. 1.5, in diesem Praxishandbuch.

<sup>28</sup> Vgl. Verheul 2006, 46 ff.

<sup>29</sup> S. a. LMER, Version 1.2, Referenzbeschreibung deutsch, 2005 (urn:nbn:de:1111-2005041102); weitere Informationen in Kapitel 6.4 „LMER“ von Tobias Steinke in Neuroth et al. 2016, Kap. 6.14–Kap. 6.16 (urn:nbn:de:0008-20090811294).

<sup>30</sup> Weitere Informationen in Kapitel 6.3 „PREMIS“ von Olaf Brandt in Neuroth et al. 2016, Kap. 6.9–Kap. 6.13 (urn:nbn:de:0008-20090811281).

<sup>31</sup> Ausführliche Informationen zur Lizenzierung von Forschungsdaten finden sich im Beitrag von Lauber-Rönsberg, Kap. 1.4, sowie Friedrich und Recker, Kap. 5.1 in diesem Praxishandbuch.

<sup>32</sup> S. <https://creativecommons.org/use-remix/cc-licenses/>.

### 3 Langzeitverfügbarkeit

Es gibt verschiedene wissenschaftsinterne und externe Gründe für die langfristige Verfügbarkeit von Forschungsdaten. Zuerst kann man bei vielen Forschungsprojekten davon ausgehen, dass der Forschungsgegenstand zum Ende der Projektlaufzeit selten komplett erforscht wurde, oder aber, dass sich in der Zukunft weitere Fragestellungen zum gleichen Gegenstand ergeben. Im Falle der Weiterbearbeitung einer Forschungsfrage ist es praktisch, wenn die früheren Daten noch zur Verfügung stehen, wobei dies sowohl das Auffinden als auch die Nachnutzbarkeit umfasst.<sup>33</sup> Insofern liegt eine langfristig sichere Aufbewahrung der eigenen Forschungsdaten für eine zukünftige Weiternutzung schon im Interesse einer oder eines jeden Forschenden selbst. Darüber hinaus gibt es Datenerhebungen, die nur einmalig möglich sind und nicht repliziert werden können. Klassische Beispiele solcher Daten sind die Beobachtungsergebnisse aus der Klimaforschung (z. B. Wetterbeobachtung, Temperaturmessungen, Satellitenbilder von Wetterphänomenen) oder die Aufzeichnung historischer Ereignisse. Aber auch im Bereich der Sprach- und soziokulturellen Forschung gibt es nicht-replizierbare Forschungsdaten, denkt man z. B. an die Beschäftigung mit historischen Sprachen, Dialekten oder Völkern.<sup>34</sup> Um die Möglichkeit der wissenschaftlichen Auseinandersetzung auch mit diesen Daten zu erhalten, müssen diese langfristig verfügbar und ausführlich dokumentiert sein. Schließlich erheben die meisten Forschungsförderer einen Anspruch auf langfristige Verfügbarhaltung von Forschungsergebnissen aus geförderten Projekten. Einerseits soll dies eine Nachhaltigkeit aufgewandter Steuergelder sicherstellen, indem redundante Datenerhebungen vermieden werden und die Überprüfung von Forschungsergebnissen möglich wird. Andererseits garantieren bspw. die Anforderungen der Deutschen Forschungsgemeinschaft (DFG) im Umgang mit Daten die Einhaltung von Grundsätzen zur guten wissenschaftlichen Praxis.<sup>35</sup>

Im Sinne der Langzeitarchivierung (LZA) geht es also darum, einerseits Forschungsdaten [...] langfristig digital zur Verfügung zu stellen und damit verifizierbar, interpretierbar und nachnutzbar zu machen und andererseits Forschungsdaten auf der Basis von Forschungsinfrastrukturen miteinander zu vernetzen und so insbesondere die potentielle Nachnutzung auch interdisziplinär zu erhöhen.<sup>36</sup>

**33** Ausführliche Informationen zur Auffindbarkeit und Nachnutzung von Forschungsdaten finden sich im Beitrag von Friedrich und Recker, Kap. 5.1 in diesem Praxishandbuch.

**34** Einen guten Überblick über gefährdete Sprachen liefert der „UNESCO Atlas of the World’s Languages in Danger“ unter <http://www.unesco.org/languages-atlas/>.

**35** Hierzu: Standards guter wissenschaftlicher Praxis im Forschungsprozess in DFG, September 2019, Leitlinien zur Sicherung guter wissenschaftlicher Praxis, Bonn: [https://www.dfg.de/download/pdf/foerderung/rechtliche\\_rahmenbedingungen/gute\\_wissenschaftliche\\_praxis/kodex\\_gwp.pdf](https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf).

**36** Oßwald, Scheffel und Neuroth 2012, 15.

In Anbetracht der vielfältigen Ziele, die mit der langfristigen Aufbewahrung von Forschungsdaten verbunden sind, ergibt sich eine Vielzahl von Aufgaben, die durch die Langzeitarchivierung umgesetzt werden sollten:

- Langfristige, sichere Aufbewahrung der Daten
- Erhalt der Interpretierbarkeit der Daten
- Auffindbarkeit der Daten sicherstellen
- Nachvollziehbarkeit der Daten gewährleisten

Langzeitverfügbarkeit ist im Umfang und in der Art der langfristigen Aufbewahrung (abhängig vom zugrundeliegenden Objekttyp) grundsätzlich jedoch abzugrenzen vom Back-up von Daten. Wie das Beispiel der Hieroglyphen sehr gut aufzeigt, beinhaltet schon die langfristige Verfügbarkeit analoger Forschungsdaten neben der „reinen“ Aufbewahrung der Daten mit geeigneter Technologie auch die Sicherstellung der langfristigen Auffindbarkeit und Interpretierbarkeit der Daten. Umso mehr trifft dies auf digitale Daten zu, die in unzähligen Formaten vorliegen können und abhängig von der Darstellung und Interpretierbarkeit durch sich verändernde Software sind. Digitale Daten müssen, um nachhaltig verfügbar gehalten zu werden, laufend überprüft und wenn notwendig konvertiert werden bzw. hinsichtlich gewählter Erhaltungsstrategie<sup>37</sup> behandelt werden. Die reine Datenspeicherung nützt wenig, wenn der originalgetreue Zugriff wegen veralteter Dateiformate oder nicht mehr verfügbarer Software nicht mehr möglich ist.

Die Herausforderung hierbei ist eine gute Dokumentation der Daten, die den Entstehungskontext, das wissenschaftliche Umfeld und die technischen Anforderungen ebenso berücksichtigt wie die Beschreibung der wissenschaftlichen Inhalte und Bedeutung der Daten.<sup>38</sup> Noch dazu sollte die Dokumentation in standardisierter und maschinenlesbarer Form erfolgen, um Interoperabilität der kommunizierenden Systeme zu gewährleisten. Die Strategien, Modelle und Systeme in Bereich der Langzeitarchivierung sollen im Folgenden beschrieben werden.

### 3.1 Erhaltungsstrategien in der Langzeitarchivierung

Je nachdem, welche Anforderungen an die langfristige Verfügbarkeit von Daten, deren Interpretierbarkeit und an die Möglichkeiten der Nachnutzung gestellt werden, gibt es verschiedene Strategien der Datenspeicherung. Die sichere Aufbewahrung der Daten als korrekte Abfolge von Nullen und Einsen (in Bits und Bytes) auf einem Speichermedium wird als „Bitstream Preservation“ bezeichnet.

---

<sup>37</sup> S. folgender Abschnitt 3.1.

<sup>38</sup> Weiterführende Informationen zur Dokumentation von Forschungsdaten finden sich im Beitrag von Dierkes, Kap. 4.1 in diesem Praxishandbuch.

Im Sinne der Langzeitverfügbarkeit digitaler Forschungsdaten muss jedoch zusätzlich zum physikalischen Erhalt der Daten auch die Darstellbarkeit und Interpretierbarkeit durch entsprechende Systeme gewährleistet werden. Ohne die Möglichkeit der Interpretation des vorhandenen Bitstreams können digitale Daten nicht dargestellt und damit nicht mehr genutzt werden. Der Erhalt der Lesbarkeit von Forschungsdaten kann dabei entweder durch die Migration der Ursprungsformate in aktuelle Formate erfolgen oder durch die Emulation oder Erhaltung der Ursprungs-umgebung der Datenentstehung.<sup>39</sup>

### **Bitstream Preservation**

Der rein physikalische Erhalt der Daten muss neben der sicheren Speicherung auch die Sicherstellung der Lesbarkeit vom physikalischen Datenträger beinhalten. Die Überprüfung der Aufbewahrung sollte also in zwei Richtungen erfolgen. Erstens muss garantiert werden, dass die auf dem Datenträger gespeicherten Daten auch nach längerer Zeit noch vorliegen und unverändert sind. Zweitens muss durch Erneuerung der Speichermedien bzw. -technologien garantiert werden, dass Daten mit Hilfe aktueller Hardware aufbewahrt und gelesen werden können.

Die Unversehrtheit der Daten kann durch die Nutzung von sog. „Fixity Checks“ in Form einer Checksummen-Bildung gewährleistet werden. Dabei wird aus den Daten mit Hilfe eines vorher definierten Algorithmus ein „Fingerabdruck“ generiert, welcher sich schon bei der kleinsten Änderung an den Daten ebenso verändert. Somit können mit Hilfe von Checksummen Änderungen an Dateien überprüft und die sichere Speicherung und Migration überwacht werden. Im Fall der redundanten Speicherung auf unterschiedlichen Speicherbereichen dient die Checksumme zusätzlich dazu, die Gleichheit der Daten zu gewährleisten, indem von Zeit zu Zeit Checksummen der Daten verglichen werden und bei Unregelmäßigkeiten eine gültige Kopie zum Einsatz kommt. Für einen validen Vergleich redundant gespeicherter Daten müssen dafür mindestens drei Kopien herangezogen werden. Erst damit wird es möglich, die korrekte(n) Datei(en) von der fehlerhaften zu unterscheiden (sofern nur bei einer Kopie Fehler aufgetreten sind).

Eine Erneuerung der Speichertechnologie kann entweder als Austausch vorhandener alter Hardware mit neuer Hardware des gleichen Typs geschehen (Refreshment) oder in der Nutzung neuer Hardwaretechnologien als Ersatz für alte nicht mehr gebräuchliche Hardware bestehen (Replication). In beiden Szenarien sollten Indikatoren, wie Fehlerraten beim Zugriff, durchschnittliche Zugriffshäufigkeit oder

---

<sup>39</sup> Vgl. auch im Folgenden, „Kapitel 8 – Digitale Erhaltungsstrategien.“ in Neuroth et al. 2016, Kap. 8.1–Kap. 8.33 (urn:nbn:de:0008-2010062472).

Alter der Hardware in Verbindung mit der Lebensdauerangabe des Herstellers als Entscheidungsgrundlage für die Migration berücksichtigt werden.

### **Formatmigration**

Die Formatmigration dient in der Regel dazu, Daten aus einem alten oder proprietären Format in ein aktuelles, standardisiertes Datenformat zu überführen. Der Fokus liegt hierbei auf dem Erhalt der Struktur und Informationen aus den alten Daten und nicht auf einer bitweisen Kopie der Daten. Das Ziel ist, die Darstellbarkeit in aktuellen und zukünftigen Systemen bzw. Anwendungen zu erhalten.

Grundlage einer möglichst verlustfreien Formatmigration ist eine Standardisierung der zu migrierenden Formate und die Kenntnis ihres Aufbaus. Darum sollte bei der langfristigen Aufbewahrung von Forschungsdaten auf die Verwendung offener, einfacher und standardisierter Formate geachtet werden. Je einfacher ein Datenformat gehalten ist, je höher ist die Wahrscheinlichkeit einer verlustarmen Formatmigration. Bei der Verwendung proprietärer Formate muss man sich darauf verlassen, dass eine Migration durch den jeweiligen Anbieter implementiert wird und diese dann auch genutzt werden kann.

Mit Hilfe der Formatmigration bleiben Informationen relativ leicht durch aktuelle Systeme darstellbar. Jedoch birgt jede Migration die Gefahr von Informationsverlust in sich. Dies kann durch die Aufbewahrung der Originaldaten inklusive aller Migrationsschritte abgemildert werden, führt aber wiederum zu einem hohen Speicherplatzbedarf. Außerdem steht die Formatmigration nicht für alle Datenformate zur Verfügung.

### **Erhalt des Entstehungskontextes (Emulation und Computermuseum)**

Zum Informationserhalt aus Datenformaten, die nicht oder nur mit hohem Aufwand migriert werden können, gibt es die Möglichkeit den originalen Entstehungskontext zu erhalten (Hardware and Software Preservation) bzw. auf aktuellen Systemen künstlich wiederherzustellen (Emulation).

Die Erhaltung der originalen Hardware und Software als Erhaltungsstrategie ist keine adäquate Methode der Langzeitarchivierung und hat eher musealen Charakter. Sicherlich liegt in der Aufbewahrung der authentischen Umgebung ein wissenschafts- und technologiehistorischer Wert, jedoch hat diese Methode sowohl ein natürliches Ende aufgrund des physischen Zerfalls der Hardware als auch einen erheblichen Ressourcenfaktor (Platzbedarf und Kosten).

Eine vielversprechendere Strategie zur Vermeidung einer möglicherweise verlustbehafteten Formatmigration ist die Nachbildung der interpretierenden Software- bzw. Hardwareumgebung auf aktuellen Systemen. Diese sogenannte Emulation

kann dabei auf Anwendungsebene, auf der Ebene des Betriebssystems oder auf der Hardwareebene umgesetzt werden. Die Emulation auf Anwendungsebene sorgt dafür, dass die ursprünglichen Formate mit Hilfe der emulierten Software vollständig interpretierbar sind. Grundlage dafür ist allerdings die Kenntnis über Struktur und Konzepte des Originalformats bzw. der Originalsoftware. Diese Art der Emulation sollte nur in Ausnahmefällen für wichtige und vielfach genutzte Formate angewandt werden, da Anpassungen je Format und Zielumgebung notwendig sind. Die Emulation des originalen Betriebssystems oder der Hardware erhält die Möglichkeit, die Ursprungssoftware in dieser Umgebung weiterhin zu benutzen.

Die Erhaltung des Entstehungskontextes von Daten hat den Vorteil, dass sowohl die Information als auch die Struktur der Ursprungsdaten bestehen bleiben kann und eine Migration nicht notwendig ist. Auf der anderen Seite ist der Aufwand für eine Emulation sehr hoch und muss bei jedem Technologiewechsel erneut nachgezogen werden.

### 3.2 OAIS-Modell

In den letzten Jahren hat sich das „Open Archival Information System“ (OAIS) als Referenzmodell für die Langzeitarchivierung von Daten etabliert. Entstanden aus Standardisierungsaktivitäten zur Aufbewahrung von Daten aus Weltraummissionen<sup>40</sup> entwickelte sich OAIS zur Grundlage vieler Systeme und Workflows in der digitalen Langzeitarchivierung. OAIS beschreibt dabei sowohl einen Standard (ISO 14721) als auch ein Modell, welches das Zusammenwirken menschlicher und technischer Akteure innerhalb eines digitalen Langzeitarchivs als komplexes System beschreibt mit der Zielsetzung, digitale Inhalte dauerhaft aufzubewahren und definierten Nutzergruppen (Designated Communities) zur Verfügung zu stellen.<sup>41</sup>

Dabei ist das Modell weder auf bestimmte Formate, Objekttypen oder Systemarchitekturen festgelegt. Vielmehr ist OAIS offen und erweiterbar, um auf die Abläufe in Organisationen anpassbar zu sein. Es verfolgt damit einen ganzheitlichen Ansatz ohne Beschränkung auf die technische Sicht auf der einen oder auf die organisatorische Sicht auf der anderen Seite. Das OAIS-Modell betrachtet die Langzeitarchivierung als Zusammenspiel der in der Hauptsache digitalen Daten als Archivgut, die dem Archivierenden anvertraut werden und die dieser für definierte Nutzergruppen bzw. in definierten Nutzungsszenarien zur Verfügung stellt.

<sup>40</sup> Das 2003 als ISO 14721 verabschiedete OAIS-Referenzmodell wurde 2002 von der Data Archiving and Ingest Working Group des Consultative Committee for Space Data Systems (CCSDS) unter Federführung der NASA veröffentlicht. Weiterführende Informationen zur Entstehung des OAIS Modells finden sich in Brübach 2016, Kap. 4.3–Kap. 4.4, und Klump 2011, 118.

<sup>41</sup> The Reference Model for an Open Archival Information System (OAIS) (Volltext), s. <http://public.ccsds.org/publications/archive/650x0m2.pdf>. Deutsche Version: <http://d-nb.info/104761314X/34>.

Die Aufgaben Übernehmen, Bewerten, Erschließen, Bewahren und Bereitstellen des Archivguts aus dem klassischen Archivwesen werden auf die Anforderungen digitaler Daten und die Möglichkeiten digitaler Informationssysteme übertragen. Die Übertragung dieser Aufgaben auf ein digitales Archiv findet sich im *Funktionsmodell* des OAIS, welches aus sechs Aufgabenbereichen besteht, die den Ablauf der Langzeitarchivierung beschreiben (vgl. Abb. 1). Ergänzt wird das funktionale Modell um ein *Datenmodell*, welches anhand von Informationsobjekten das Archivgut selbst in drei Manifestationen beschreibt und Anforderungen an die Form und die Beschreibung dieser Informationsobjekte formuliert.

### Datenmodell

Ein Informationspaket wird als logischer Container betrachtet, der neben den Primärdaten (Content Information) selber zusätzliche, optionale Erhaltungsmetadaten (Preservation Description Information) enthalten kann. Weiterhin gehört zu einem Informationspaket die Verpackungsinformation, welche die Inhaltsinformation und die Paketbeschreibungsinformationen sowohl miteinander verbindet als auch voneinander abgrenzt und das Suchen nach der Inhaltsinformation ermöglicht.

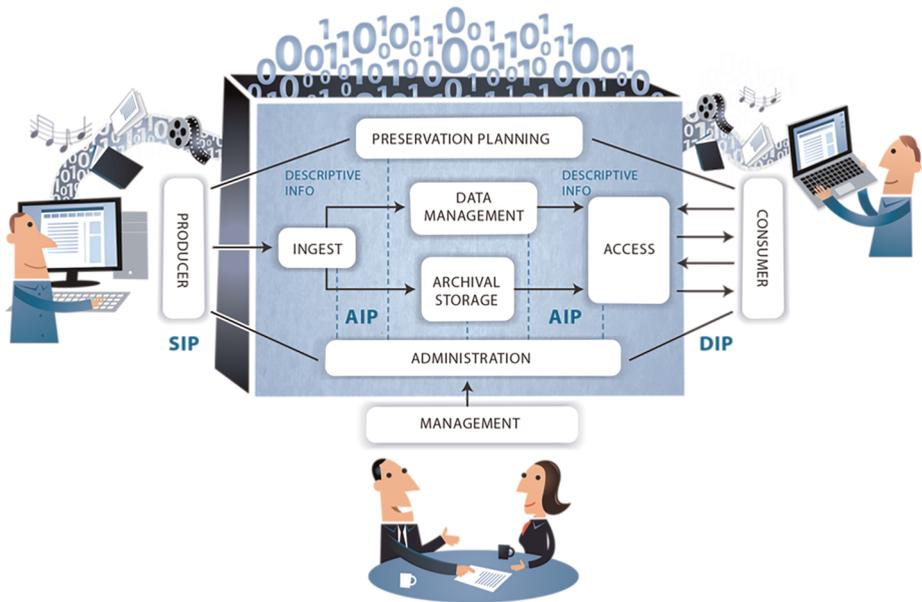
Das OAIS unterscheidet zwischen der Manifestation des entgegengenommenen Informationsobjektes als Submission Information Package (SIP), dem um archivische Metadaten ergänzten Objekt, dem sogenannten Archival Information Package (AIP) und den Repräsentationen dieser AIP für definierte Nutzungsszenarien, den sogenannten Dissemination Information Packages (DIP). Ein SIP wird durch den Produzenten zusammengestellt und zur Übernahme (Ingest) in das Archiv zur Verfügung gestellt. Aufgrund des Archivierungskonzepts des übernehmenden Systems wird daraus das AIP erstellt, wobei ein SIP sowohl 1:1 als AIP abgebildet werden, Teil eines größeren AIP (N:1) sein oder in mehrere AIP (1:N) aufgeteilt werden kann. Diese Entscheidung liegt in der Verantwortung des entgegennehmenden Archivs und in dessen Architektur begründet. Die Umformung des angebotenen SIP zu einem AIP kann bspw. die Umwandlung des gelieferten Datenformats in ein archivierungskonformes Format beinhalten. Da das OAIS Modell keine Aussagen zu Datenformaten trifft, kann es durchaus sinnvoll sein, vordefinierte Archivierungsformate im AIP zu nutzen. Bspw. könnte es die Vorgabe zur Archivierung von Texten als PDF/A geben, so dass während des Ingests alle Textformate in PDF/A konvertiert werden.<sup>42</sup>

Je nach Nutzungsinteresse kann beim Zugriff auf die Daten aus einem oder mehreren AIP ein DIP abgeleitet werden, welches das Archiv dann als Antwort auf

---

<sup>42</sup> Angaben zu geläufigen Archivierungsformaten finden sich im Abschnitt „Praxisbezug“ (s. Abschnitt 4: Datenkuration – Normalisierung/Standardisierung).

eine Anfrage an das OAIS dem Endnutzer zur Verfügung stellt. Der Anfragende erhält die Daten aus dem Archivsystem nicht, wie vormalig gespeichert, sondern als auf seine Bedürfnisse zugeschnittenes Informationspaket. Eine OAIS-konforme Umsetzung muss an dieser Stelle die Authentizität und Integrität der Informationen sicherstellen.



**Abb. 1:** Das Funktionsmodell nach OAIS, s. Abschnitt 3.2 OAIS-Modell (Urheber der Illustration ist digitalbevaring.dk, die Nutzung erfolgt unter CC BY 2.5 Denmark licence)

## Funktionsmodell

Die grundlegende Erhaltungsstrategie hinter dem OAIS-Modell ist die Formatmigration mit dem Ziel der Erhaltung der Information bei der langfristigen Aufbewahrung. Die Authentizität der Daten hat bei diesem Ansatz nachgeordnete Bedeutung. Insofern sind auch die einzelnen Module des Funktionsmodells Ausdruck dieser Erhaltungsstrategie.

Funktionsmodule im Bereich Archivgutverwaltung:

- Datenübernahme (Ingest)
- Datenaufbewahrung (Archival Storage)
- Szenario-basierter Datenzugriff (Dissemination/Access).

Das Funktionsmodul *Datenübernahme* umfasst alle Dienste und Funktionen die zur tatsächlichen Annahme und Verarbeitung eines Informationspakets durch das Archivsystem notwendig sind. Dabei wird das von einem Produzenten bereitgestellte SIP angenommen, geprüft und zu einem AIP weiterverarbeitet.

Das SIP sollte neben den Primärdaten zusätzlich Belege für die Authentizität und die Herkunft der Informationsobjekte liefern, die das Archiv als Teil des AIP dauerhaft übernimmt und erhält. Schließlich sollte die erfolgreiche Übergabe durch eine Bestätigung an den Produzenten abgeschlossen werden. Ebenfalls Teil der Übernahme ist die Qualitätssicherung der Daten. Es wird geprüft, ob der Transfer des SIP erfolgreich ohne Schreib- und Lesefehler erfolgt ist, dazu werden z. B. Checksummen und Systemprotokolle genutzt.

Zur Erzeugung eines AIP wird das SIP nach den Vorgaben des Archivs transformiert und um Metadaten angereichert. Dies kann bspw. eine Migration des Dateiformats, die Erzeugung zusätzlicher Repräsentationen und das Auslesen und Speichern zusätzlicher Metadaten (z. B. technische, administrative Metadaten) beinhalten. Ebenso kann die Struktur des Informationspakets bei der Übernahme verändert werden. Aus dem AIP werden die deskriptiven Metadaten übernommen und an die Datenverwaltung geliefert, um eine Recherche nach den archivierten Inhalten zu ermöglichen. Die Informationen zum Speicherort des AIP werden ebenfalls in die Datenverwaltung übernommen.

Für den Datenzugriff auf archivierte Inhalte richtet die oder der Endnutzende (Consumer) Anfragen an das OAIS und erhält Antworten in der Form eines oder mehrerer DIP. Zur Recherche nach relevanten Informationen und Generierung von einmaligen oder regelmäßigen Anfragen stellen OAIS Systeme geeignete Tools zur Verfügung, bspw. einen Suchindex. Die Funktionseinheit Zugriff (Access) stellt die beschriebene Funktionalität zur Verfügung. Dabei werden eventuell bestehende Einschränkungen des Zugriffs berücksichtigt (z. B. Filterung personenbezogener Daten). Aus den gewünschten AIP werden DIP erzeugt und an die oder den Endnutzenden online oder offline ausgeliefert. Hierbei können die DIP Repräsentationen der Informationsobjekte enthalten, die durch Transformation (Konvertierung in definierte Ausgabeformate) oder Bearbeitung (Ausschnitte, Bildbearbeitung) entstanden sind.

Administrative Funktionsmodule

- Datenmanagement
- Systemverwaltung
- Preservation Planning

Die administrativen Module, insbesondere die Systemverwaltung, beschreiben im Wesentlichen Aufgaben, die der Betreiber eines OAIS-Systems zu verantworten hat. Zu den Aufgaben im Datenmanagement gehört die Verwaltung und Aufbereitung der deskriptiven und archivarischen Metadaten. Services des Datenmanagements sind z. B. das Ausführen von Suchabfragen und die Ausgabe von Ergebnismengen, das Ausführen von ereignisbasierten, regelmäßigen Datenabfragen oder das Aus-

führen von verarbeitenden Algorithmen, die über die abgerufenen Daten laufen. Weiterhin können auf Grundlage der archivarischen Metadaten, Endnutzer-Zugriffs-Statistiken, Endnutzerabrechnungen, Sicherheitskontrollen, Ablaufpläne sowie Reports zum Monitoring erstellt werden.

Das Preservation Planning umfasst sowohl die Beobachtung des technologischen Fortschritts, als auch die Entwicklung und Umsetzung der Erhaltungsmethoden. Veraltete Datenformate müssen in aktuelle Formate konvertiert werden. Dabei werden durchgeführte Erhaltungsmaßnahmen dokumentiert, es wird auf die Erhaltung der Integrität geachtet und Rechtsverbindlichkeiten werden berücksichtigt.

## 4 Praxistransfer – Hilfestellungen für die Praxis

Im Abschnitt Praxistransfer möchten wir ein paar Arbeitsmittel an die Hand geben, die im Alltag bei der Entscheidung über die Art und Dauer der Aufbewahrung von Forschungsdaten unterstützen sollen. Außerdem werden Systeme gelistet, die sich für die langfristige Speicherung von Forschungsdaten anbieten, inklusive Vor- und Nachteile.

### Datenspeicherung

Um den sicheren physischen Erhalt von Forschungsdaten zu gewährleisten, sollten folgende Empfehlungen eingehalten werden:

- Verwendung von mindestens drei redundanten Kopien der Daten und Generierung von Checksummen aus den Originaldateien.
- Speicherung der Daten auf heterogenen, aber standardisierten Speichermedien, die im besten Fall auch noch organisatorisch und räumlich verteilt sind.
- Regelmäßige Migration der Daten auf neue (aktuelle) Speichermedien mit integrierten Fixity Checks (Überprüfen der Checksummen) während des Umkopierens. Dabei kann die Migration sowohl auf Grundlage eines Datenträgeraus-tauschs (Refreshment) als auch als Technologiewechsel (Replication) stattfinden.

### Datenkuration – Selektion

Die Archivwürdigkeit von Forschungsdaten kann man mit Hilfe folgender Checkliste überprüfen.<sup>43</sup> Dabei erhebt die Liste keinen Anspruch auf Vollständigkeit. Sie kann

---

<sup>43</sup> Je mehr Fragen mit Ja beantwortet werden können, umso höher ist die Archivwürdigkeit der Forschungsdaten einzuschätzen. Eine Priorisierung der einzelnen Fragen ist durch die Reihenfolge

lediglich eine grobe Hilfestellung bei der Bewertung von Forschungsdaten sein. Die letztendliche Entscheidung über die Relevanz von Forschungsdaten obliegt der Wissenschaft respektive den Forschenden selber.

- Bestehen Vorgaben Dritter (Fördergeber, Datenpolicies, Richtlinien der Forschungseinrichtung), die es notwendig machen, die Daten langfristig aufzubewahren?
- Hat man die notwendigen Nutzungsrechte an den Daten? Unter welchen Bedingungen besitzt man die Daten?
- Sind die erhobenen Daten einmalig und nicht reproduzierbar oder sind die Kosten der Reproduktion höher als die Kosten der Langzeitaufbewahrung?
- Liefert die Datenerhebung durch den technologischen Fortschritt voraussichtlich keine besseren Ergebnisse?
- Gibt es ein hohes Nachnutzungsinteresse an den Forschungsdaten?
- Wurden die Daten noch nicht vollständig wissenschaftlich untersucht?
- Sind die Daten charakteristisch oder untypisch für ein Forschungsgebiet bzw. handelt es sich um einmalige Forschungsergebnisse?
- Haben die Daten möglicherweise eine allgemeine oder regionale historische Bedeutung?
- Ist die Datenqualität technisch und inhaltlich gut?
- Sind deskriptive Metadaten vollständig vorhanden oder können generiert werden?
- Können die notwendigen Erhaltungsmetadaten (Referenz-, Provenienz-, Kontext- und Persistenz-Informationen sowie Angaben zu Zugriffsrechten) geliefert werden?

### **Datenkuration – Normalisierung/Standardisierung**

Im Laufe der Zeit haben sich für verschiedene Arten von Dokumenten quasi Standards herausgebildet, die weit verbreitet sind und von vielen Systemen und Anwendungen unterstützt werden. Sollte die Wahl des Datenformats unabhängig von proprietären Formaten aus Messinstrumenten oder aus individueller (Hersteller-)Software sein, dann empfiehlt es sich die Forschungsdaten in folgende Formate zu transformieren.<sup>44</sup> Weiterhin sollten möglichst einfache Strukturen und Formate gewählt werden, die am besten durch Menschen lesbar und interpretierbar sind (z. B.

---

nicht impliziert, da diese im Zweifelsfall durch die Forschenden bzw. die Forschungscommunity selbst vorgenommen werden muss.

**44** Aktuelle Listen mit Empfehlungen für Standardformate zur Archivierung von Daten finden sich online bspw. im Katalog archivischer Dateiformate (KAD) unter: [https://kost-ceco.ch/cms/kad\\_main\\_de.html](https://kost-ceco.ch/cms/kad_main_de.html), auf der Webseite der ETH Zürich unter: <https://documentation.library.ethz.ch/display/RC/Archivtaugliche+Dateiformate>, und auf der Webseite der Library of Congress unter: <https://www.loc.gov/preservation/resources/rfs/TOC.html>.

Präferenz von Textdokumenten gegenüber der binären Darstellung). Die Tabelle enthält eine Auswahl gängiger Dokumenttypen und eine Empfehlung für ein stabiles und archivierungsfähiges Format.<sup>45</sup>

**Tab. 1:** Gängige Dokumenttypen samt Formatbezeichnung und Kürzel.

Dokumenttyp	Formatbezeichnung	Dateinamen-erweiterung
3D-Anwendung	COLLADA Digital Asset Exchange	*.dae
	Wavefront OBJ	*.obj
	Polygon File Format	*.ply
	Extensible 3D	*.x3d
Audio	Waveform Audio	*.wav
Bild/Rastergrafik	Windows Bitmap	*.bmp
	JPEG 2000 part 1	*.jpg
	Open Microscopy Environment – Tagged Image File Format	*.ome.tiff
	Portable Network Graphics	*.png
Tagged Image File Format	*.tif	
GIS (Geoinformationssystem)	Geography Markup Language	*.gml
PDF (Portable Document Format)	Acrobat PDF/A – Portable Document Format 1a	*.pdf (PDF/A-1a)
	Acrobat PDF/A – Portable Document Format 1b	*.pdf (PDF/A-1b)
	Acrobat PDF/A – Portable Document Format 2a	*.pdf (PDF/A-2a)
	Acrobat PDF/A – Portable Document Format 2b	*.pdf (PDF/A-2b)
	Acrobat PDF/A – Portable Document Format 2u	*.pdf (PDF/A-2u)
Unabhängiges textbasiertes Format	Character-Separated Values	*.csv
	Hypertext Markup Language	*.html
	Markdown	*.md
	Standard Generalized Markup Language	*.sgml
	Text file	*.txt

<sup>45</sup> Angelehnt an die LZV-Dateiformatliste des Hochschulbibliothekszenrum des Landes Nordrhein-Westfalen (hbz), s. <https://www.hbz-nrw.de/produkte/langzeitverfuegbarkeit/langzeitverfuegbarkeit-fuer-hochschulen/lzv-dateiformatliste>.

Dokumenttyp	Formatbezeichnung	Dateinamen-erweiterung
	Extensible Hypertext Markup Language	*.xhtml
	Extensible Markup Language	*.xml
Vektorgrafik	Scalable Vector Graphics	*.svg
Video	Motion JPEG 2000	*.mj2, *.mjp2
	Matroska Multimedia Container (FF video codec 1)	*.mkv (FFV1)
Webarchivierung	Web ARChive	*.warc

### Langfristige Verfügbarkeit

Je nach Anwendungsszenario, Ansprüchen an die langfristige Verfügbarkeit von Forschungsdaten und technischer Infrastruktur gibt es verschiedene Systeme, Forschungsdaten langfristig aufzubewahren. Aufgrund der Vielzahl der Alternativen zur langfristigen Speicherung von Forschungsdaten und der sehr unterschiedlichen Anforderungen und Kosten ist es notwendig, sich schon frühzeitig im Datenlebenszyklus Gedanken darüber zu machen, welche Zielgruppen und Nachnutzungsszenarien mit der Aufbewahrung der Daten erreicht werden sollen.

### OAIS-konforme Systeme

OAIS-konforme Systeme zeichnen sich dadurch aus, dass sie Workflows zur Archivierung vieler Formattypen zur Verfügung stellen und die Funktionsmodule des OAIS-Modells, die dauerhafte Archivierung von digitalen Informationsobjekten und die Erhaltung dauerhaften Zugangs, implementieren. Die Validierung und Charakterisierung der Formattypen erfolgt in der Regel durch Einbindung externer Tools wie z. B. DROID<sup>46</sup> und JHOVE<sup>47</sup> oder die Nutzung von Formatdatenbanken wie z. B. PRONOM.<sup>48</sup> Genauso können im Allgemeinen neben den standardmäßig implementierten Auslieferungsformen auch Viewer für die verschiedenen Objekttypen als Plugins angebunden werden. Grundsätzlich können OAIS-konforme Systeme als „light archive“ oder als „dark archive“ betrieben werden. Ein „light archive“ bedeutet, dass der Zugang zu den Materialien im Archivsystem für die Nutzer über ein Discovery System, wie z. B. einen Katalog/OPAC erfolgen. Dieser Index kann sowohl im Sys-

<sup>46</sup> Digital Record Object Identification (DROID) ist ein Open Source Tool zur automatischen Formaterkennung von Dateien. S. <http://digital-preservation.github.io/droid/>.

<sup>47</sup> S. <https://jhove.openpreservation.org/>.

<sup>48</sup> S. <https://www.nationalarchives.gov.uk/PRONOM/Default.aspx>.

tem selbst betrieben oder auch extern angebunden werden. Beim „dark archive“ gibt es keine öffentliche Bereitstellung. Anwendungsfälle für das „dark archive“ sind z. B. Materialien, die zunächst nicht für eine Veröffentlichung vorgesehen sind, oder für welche bereits eigene Präsentationslösungen etabliert sind.

Rosetta<sup>49</sup> ist ein kommerzielles Produkt der Firma Ex Libris, das als OAIS-konformes Langzeitarchiv in Zusammenarbeit mit der Nationalbibliothek von Neuseeland entwickelt wurde und seit 2009 auf dem Markt verfügbar ist. Grundsätzlich ist Rosetta als System für Digital Preservation und Digital Asset Management mit individuell konfigurierbaren Workflows für alle Arten von Daten- und Dateiformaten nutzbar. Es kann sowohl als zentrale Installation mit mehreren Mandanten (Institutionen) als auch dezentral als lokale Installation betrieben werden. Rosetta setzt auf ein flexibles Datenmodell, um alle Objekttypen abbilden zu können. Dieses Datenmodell orientiert sich dabei am Objektmodell des PREMIS-Standards und umfasst vier Level: Intellektuelle Entität, Repräsentation, File und Bitstream. Für alle vier Ebenen werden langzeitarchivierungsrelevante Metadaten geschrieben. Diese Metadaten folgen dabei konzeptuell den von PREMIS vorgegebenen Entitäten: Objects, Events, Agents, Rights. Als Format für die Abbildung der Metadaten auf allen Ebenen wird der Metadata Encoding & Transmission Standard (METS)<sup>50</sup> eingesetzt, sodass für jedes AIP eine Metadatendatei existiert. Im Bereich der deskriptiven Metadaten unterstützt Rosetta primär Dublin Core.<sup>51</sup> Es ist aber möglich, Metadaten in anderen Standardformaten oder eigenen Originalformaten als Source-Metadaten nach Rosetta zu übernehmen. Langzeitarchivierung als Dienstleistung bieten im Bereich der Bibliotheken in Deutschland SLUB Dresden,<sup>52</sup> TIB Hannover,<sup>53</sup> der Bayerische Bibliotheksverbund<sup>54</sup> und das hbz in Köln<sup>55</sup> an.

Archivematica<sup>56</sup> ist ein weiteres OAIS-konformes, universell einsetzbares Langzeitarchivierungssystem, welches die gesamte Breite der Langzeitarchivierungsprozesse abdeckt. Das System wird in Großbritannien als zentrale Lösung zur langfristigen Aufbewahrung von Forschungsdaten eingesetzt. Es verfügt über keine eigene Endnutzeroberfläche, sondern nur über eine Verwaltungsoberfläche. Daher ist es vorwiegend als „dark archive“ zu verstehen, welches die Auslieferung von DIP bspw. an Repositorien ermöglicht. Das System beherrscht die gängigen Metadaten-

49 S. <https://www.exlibrisgroup.com/de/produkte/Rosetta>.

50 S. <http://www.loc.gov/standards/mets/>.

51 S. <https://www.dublincore.org/>.

52 S. <https://slubarchiv.slub-dresden.de/>.

53 S. <https://wiki.tib.eu/confluence/display/lza/Digitale+Langzeitarchivierung+an+der+TIB>.

54 S. <https://www.bib-bvb.de/web/digitales-langzeitarchivierungssystem/home>.

55 S. <https://www.hbz-nrw.de/produkte/langzeitverfuegbarkeit/langzeitverfuegbarkeit-fuer-hochschulen>.

56 S. <https://www.archivematica.org/en/>; Dokumentation: <https://www.archivematica.org/en/docs/archivematica-1.6/contents/>.

formate (z. B. PREMIS, METS, Dublin Core). Die Open-Source-Software ist nicht mandantenfähig, weswegen der Betrieb einer zentralen Instanz für einen Verbund nicht out-of-the-box unterstützt wird. Archivemata ermöglicht sowohl Hosting- als auch On-Premise-Lösungen. Hosting ist vor allem über die Services Arkivum<sup>57</sup> und ArchivesDirect<sup>58</sup> verfügbar. Im nationalen Kontext wird das System in größerem Maßstab durch den Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV)<sup>59</sup> als Angebot für die digitale Langzeitarchivierung genutzt.<sup>60</sup> Zu den internationalen Anwendern von Archivemata gehören zum Beispiel die University of British Columbia<sup>61</sup> oder das Museum of Modern Art (MoMA)<sup>62</sup> in New York.<sup>63</sup>

Ebenso wie Rosetta bietet auch Preservica<sup>64</sup> ein kommerzielles Langzeitarchivierungssystem nach dem OAIS-Modell. Zentrale ebenso wie dezentrale Lösungen sind mit Preservica möglich. Das System ist mandantenfähig, d. h. man kann eine Anwendung mit verschiedenen Partitionen und Rechten fahren. Hierfür bietet der Hersteller drei Lösungen an, Software as a Service (SaaS), Hosting und On-Premise (Cloud Edition und Enterprise Edition). Es handelt sich prinzipiell um eine kommerzielle Lösung; diese ist aber erweiterbar durch Open Source Tools. Zudem verfügt sie über einen eigenen Software Development Kit (SDK) und offene Programmierschnittstellen (APIs). Allerdings stehen sowohl die Dokumentation als auch die Nutzerhandbücher des Systems und die Foren zum Austausch der Anwender nur für Kunden zur Verfügung. Wie andere OAIS-Systeme auch, setzt Preservica Standard-File Format Registries und Migrationstools wie DROID, PRONOM und Linked Data Registries ein, um Erhaltungsmaßnahmen und Migrationspfade für mehr als 1200 Dateiformate zu automatisieren. Deskriptive Metadaten können aus mehreren Standard-Schemata ausgewählt werden (Encoded Archival Descriptor – EAD<sup>65</sup> – 2002, Metadata Object Description Schema – MODS<sup>66</sup> – 3.4, Dublin Core 1.1). Alternativ können auch nutzerdefinierte deskriptive Metadatenschemata verwendet werden, z. B. XML Schema Definitions. Auf allen Ebenen werden sowohl deskriptive als auch langzeitarchivierungsrelevante Metadaten im XIP-Format verwaltet. Diese können bei der Bildung von SIP-Packages in METS und PREMIS-Metadaten konvertiert wer-

---

57 S. <http://arkivum.com/>.

58 S. <http://www.archivesdirect.org/>.

59 S. <https://www.kobv.de/>.

60 S. <https://www.kobv.de/services/archivierung/lza/>.

61 S. <https://www.ubc.ca/>.

62 S. <https://www.moma.org/>.

63 S. <https://www.artefactual.com/clients/>.

64 S. <http://preservica.com/>.

65 S. <https://www.loc.gov/ead/>.

66 S. <http://www.loc.gov/standards/mods/>.

den. Preservica wird z. B. an der Wellcome Library,<sup>67</sup> an der Yale University<sup>68</sup> und der Danish National Library<sup>69</sup> zur Archivierung diverser Objekttypen eingesetzt.

### (Fachspezifische) Forschungsdatenrepositorien

Neben OAIS-konformen Systemen zur Langzeitarchivierung, die für die langfristige Aufbewahrung unterschiedlicher Objekttypen (z. B. Dokumente, AV-Medien, Retrodigitalisate, Forschungsdaten, Verwaltungsdaten) konzipiert sind, gibt es speziell auf Forschungsdaten ausgerichtete Infrastrukturen und Repositorien. Insbesondere sind hier die fachspezifischen Datenrepositorien zu nennen, die zur Publikation von Forschungsdaten entstanden sind.<sup>70</sup> Die Publikation der Forschungsdaten auf einem solchen Fachrepositorium beinhaltet auch die langfristige Speicherung der Daten. Jedoch gilt es an dieser Stelle zu beachten, dass es sich mehrheitlich um reine Bitstream Preservation handelt.

Zur Archivierung und Publikation sogenannter Long-Tail-Forschungsdaten oder Daten, zu denen keine fachspezifischen Angebote vorhanden sind, stehen fachübergreifende Repositorien zur Verfügung. Genannt werden sollen an dieser Stelle Figshare,<sup>71</sup> Zenodo,<sup>72</sup> Dryad<sup>73</sup> und RADAR<sup>74</sup> als aktuell prominente übergreifende Aufbewahrungssysteme. Figshare und Zenodo sind klassische Publikationssysteme. Hier ist die Archivierung der Daten keine eigenständige Funktionalität, sondern geschieht aufgrund der Tatsache, dass publizierte Daten verfügbar gehalten werden müssen.

Das Research Data Repositorium (RADAR) ist das Ergebnis des gleichnamigen DFG-geförderten Projekts eines disziplinübergreifenden Projektteams (FIZ Karlsruhe,<sup>75</sup> Karlsruher Institut für Technologie – KIT<sup>76</sup>, Fakultät für Chemie und Pharmazie

67 S. <http://blog.wellcomelibrary.org/2011/07/preserving-our-digital-assets-1-sdb4/>.

68 S. <https://yaledailynews.com/blog/2015/12/10/libraries-utilize-preservica/>.

69 S. <https://preservica.com/resources/press-releases/state-and-university-library-of-denmark-collaborates-with-preservica-to-safeguard-history-of-danish-cultural-heritage>.

70 Einen Überblick und Unterstützung bei der Suche nach einem geeigneten Repositorium bietet re3data unter: <http://www.re3data.org/>, und RIsources der DFG unter: <http://risources.dfg.de/>. Eine ausführliche Liste der gängigen Repositorien, gegliedert nach Fachbereichen, wird vom Open Access Directory (OAD) oder dem Directory of Open Access Repositories (OpenDOAR) bereitgestellt. OAD: [http://oad.simmons.edu/oadwiki/Data\\_repositories](http://oad.simmons.edu/oadwiki/Data_repositories), OpenDOAR: <http://v2.sherpa.ac.uk/opendoar/>.

71 S. <https://figshare.com/>.

72 S. <https://zenodo.org/>.

73 S. <https://datadryad.org/stash>.

74 S. <https://www.radar-projekt.org/display/RD/Home>.

75 S. <https://www.fiz-karlsruhe.de/>.

76 S. <https://www.kit.edu/>.

der LMU,<sup>77</sup> Leibniz-Institut für Pflanzenbiochemie – IPB<sup>78</sup> und TIB Hannover<sup>79</sup>). Ziel des Projekts war der Aufbau einer Infrastruktur für die Datenarchivierung und -publikation in der öffentlichen (dauerhaften) Domäne. RADAR ist disziplinübergreifend konzipiert und bietet eine zentrale Anlaufstelle zur Archivierung und Publikation vielfältiger Daten und Dateiformate. Je nach gewähltem Service Level (Archivierung mit und ohne Publikation der Daten) kann das System auf unterschiedliche Art genutzt werden. Es gibt ein zweistufiges Geschäftsmodell mit unterschiedlichem zeitlichem Horizont: Archivierung der Daten mit und ohne Datenpublikation. Das reine Archivierungsangebot umfasst dabei die langfristige Speicherung von Datenpaketen für eine vom Kunden festgelegte Haltefrist (5–15 Jahre). Nach Ablauf der Haltefrist kann diese verlängert oder die Daten gelöscht werden. Werden Daten publiziert, gilt eine Haltefrist von mindestens 25 Jahren, wobei eine unbegrenzte Aufbewahrung angestrebt wird. Auf jeden Fall werden die Daten nicht gelöscht. Die Langzeitverfügbarkeit der Daten wird über eine reine Bitstream Preservation sichergestellt. Diese ist abgesichert über MD5-Checksums, die beim Ingest, allen Kopiervorgängen und beim Ausliefern überprüft werden. Während der Aufbewahrungsfrist verändert RADAR die gespeicherten Datenpakete nicht mehr, sondern sichert ausschließlich deren physischen Erhalt.

## Fazit

Die langfristige Speicherung von Daten im Forschungsumfeld ist abhängig vom speziellen Projekt und den gestellten Zielsetzungen. Da das Angebot an Speichermedien auch in Zukunft sehr breit, allerdings auch unterschiedlich kostenintensiv sein wird, wird es bei der physikalischen Speicherung von Daten immer eine Abwägung zwischen den Faktoren Speicherkosten, Speicherkapazität und benötigten Zugriffsgeschwindigkeiten geben. Weiterhin können rechtliche Randbedingungen Auswirkungen auf die für ein Projekt verwendbaren Speicher Einfluss haben, was z. B. bei der Verarbeitung von personenbezogenen Daten eine Rolle spielt. Die Datensicherheit bedingt weiterhin Überlegungen zur Erkennung korruptierter Daten und deren Wiederherstellung durch geeignete Back-up-Strategien.

Diese Aspekte sollten schon vor dem Beginn einer Forschung ausreichend geklärt werden. Genauso wie auch eine möglich Nachnutzung erhobener Daten schon von Anfang an mitgedacht werden sollte.

Um dies in die aktuellen Entwicklungen im Forschungsdatenmanagement zu integrieren, ist die Kooperation verschiedener zentraler Einrichtungen mit den For-

---

<sup>77</sup> S. <https://www.cup.uni-muenchen.de/>.

<sup>78</sup> S. <https://www.ipb-halle.de/>.

<sup>79</sup> S. <https://www.tib.eu/de/>.

schenden wichtig. Beim Aufbau einer Infrastruktur zur Beratung zum Thema Forschungsdatenmanagement sollte daran gedacht werden, dass Fachleute im Bereich Speichersysteme und Datenkuration in die Planung einbezogen werden, ebenso wie Fachleute im Bereich von Metadaten und Standards in der Langzeitarchivierung. Serviceeinrichtungen im Bereich Rechnerinfrastruktur sollten eine Liste der Services inklusive der Kosten veröffentlichen, so dass sich die Forschenden bereits im Vorfeld der Beantragung von Fördermitteln dazu ein realistisches Bild machen können.

## Literatur

Letztes Abrufdatum der Internet-Dokumente ist der 15.11.2020.

- Apel, Jochen, Fabian Gebhart, Leonhard Maylein und Martin Wlotzka. 2018. „Offene Forschungsdaten an der Universität Heidelberg: Von generischen institutionellen Repositorien zu fach- und projektspezifischen Diensten.“ *O-Bib. Das Offene Bibliotheksjournal* 5 (2): 61–71. doi:10.5282/o-bib/2018H2S61-71.
- BMI. 1987. „Grundsätze zur Durchführung der Sicherheitsverfilmung von Archivalien: Bek. d. BMI v. 13.05.1987 – ZV 1 M 325 100–213.“ *Gemeinsames Ministerialblatt*, hg. v. Bundesministerium des Innern, – 38 (16): 284–292.
- Borges, Georg und Jan Geert Meents, Hg. 2016. *Cloud Computing: Rechtshandbuch*. München: C. H. Beck.
- Clelland, C. T. et al. 1999. „Hiding messages in DNA microdots.“ *Nature* 399: 533–534.
- Corti, Louise, Veerle Van den Eynden, Libby Bishop und Matthew Woollard. 2014. *Managing and Sharing Research Data. A Guide to Good Practice*. London: SAGE.
- Cabrera Valdes, Victoria und James L. Bischoff. 1989. „Accelerator 14C dates for early upper paleolithic (basal Aurignacian) at El Castillo Cave (Spain)“ *Journal of Archaeological Science* 16 (6): 577–584. doi:10.1016/0305-4403(89)90023-X.
- Depuydt, Leo, 1999. „Rosetta Stone“. In *Encyclopedia of the Archaeology of Ancient Egypt*, hg. v. Kathryn A. Bard, 686–687. London: Routledge.
- Deutsche Forschungsgemeinschaft (DFG). 2013. *Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission ‚Selbstkontrolle in der Wissenschaft‘*. Ergänzte Auflage. Weinheim: Wiley VHC. doi:10.1002/9783527679188.oth1.
- Deutsche Forschungsgemeinschaft (DFG). 2019. „Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex.“ [https://www.dfg.de/download/pdf/foerderung/rechtliche\\_rahmenbedingungen/gute\\_wissenschaftliche\\_praxis/kodex\\_gwp.pdf](https://www.dfg.de/download/pdf/foerderung/rechtliche_rahmenbedingungen/gute_wissenschaftliche_praxis/kodex_gwp.pdf).
- Factor, Michael, Kalman Meth, Dalit Naor, Ohad Rodeh und Julian Satran. 2005. *Object storage: The future building block for storage systems*. In *2005 IEEE International Symposium on Mass Storage Systems and Technology*, Sardinia, Italy, 2005: 119–123. doi:10.1109/LGDI.2005.1612479.
- Hennessy, John L., David A. Patterson und Andrea Arpaci-Dusseau. 2007. *Computer Architecture: A Quantitative Approach*. 4. Auflage. Amsterdam, Heidelberg [u. a.]: Elsevier.
- Klump, Jens. 2011. „Langzeiterhaltung digitaler Forschungsdaten.“ In *Handbuch Forschungsdatenmanagement*, hg. v. Stephan Büttner, Hans-Christoph Hobohm, Lars Müller. Bad Honnef: Bock u. Herchen: 115–119
- Leggett, Elizabeth R. 2014. *Digitization and digital archiving: a practical guide for librarians*. Lanham, Plymouth: Rowman & Littlefield.

- Majonica, Rudolf. 2007. *Das Geheimnis der Hieroglyphen: die abenteuerliche Entschlüsselung der ägyptischen Schrift durch Jean François Champollion; mit zahlreichen dokumentarischen Abbildungen und zeitgenössischen Grafiken*, Überarb. und erw. Neuausg. München: Dt. Taschenbuch-Verlag.
- Mohamed, H. S. Al Risi, T. L. Jin, J. Kosel, S. N. Piramanayagam und R. Sbiaa. 2020. „Controlled spin-torque driven domain wall motion using staggered magnetic wires.“ *Appl. Phys. Lett.* 116: 032402. doi:10.1063/1.5135613.
- nestor – Kompetenznetzwerk digitale Langzeitarchivierung. *Publikationen* [https://www.langzeitarchivierung.de/Webs/nestor/DE/Publikationen/publikationen\\_node.html](https://www.langzeitarchivierung.de/Webs/nestor/DE/Publikationen/publikationen_node.html).
- Neuroth, Heike, Achim Oßwald, Regine Scheffel, Stefan Strathmann und Mathias Jehn, Hg. 2016. *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung, Version 2.3. 2.*, aktualisierte Druckauflage. Glückstadt, Göttingen: Werner Hülsbusch Fachverlag für Medientechnik und -wirtschaft.
- Neuroth, Heike, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump und Jens Ludwig, Hg. 2012. *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*, Boizenburg, Göttingen: Werner Hülsbusch.
- Oßwald, Achim, Regine Scheffel und Heike Neuroth. 2012. „Langzeitarchivierung von Forschungsdaten. Einführende Überlegungen.“ In *Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme*, hg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump und Jens Ludwig, 13–21. Boizenburg, Göttingen: Werner Hülsbusch.
- Ovshinsky, Stanford R. 1968. „Reversible Electrical Switching Phenomena in Disordered Structures“ *Phys. Rev. Lett.* 21 (20): 1450–1453.
- Parkin, Stuart S. P., Masamitsu Hayashi und Luc Thomas. 2008. „Magnetic Domain-Wall Racetrack Memory“ *Science* 320 (5873): 190–194. doi:10.1126/science.1145799.
- Rapp, Franziska, Stefan Kombrink, Volodymyr Kushnarenko, Matthias Fratz und Daniel Scharon. 2018. „SARA-Dienst: Software langfristig verfügbar machen.“ In *O-Bib. Das Offene Bibliotheksjournal* 5 (2): 92–105. doi:10.5282/o-bib/2018H2S92-105.
- Rfll. 2016. *Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland*. Göttingen. <http://www.rfii.de/?wpdmdl=1998>.
- Rfll. 2017. *Entwicklung von Forschungsdateninfrastrukturen im internationalen Vergleich. Bericht und Anregungen*. Göttingen. <http://www.rfii.de/?wpdmdl=234>.
- Tanenbaum, Andrew S. und James Goodman. 2001. *Computerarchitektur: Struktur, Konzepte, Grundlagen*. 2. Auflage. München: Prentice Hall, Pearson Studium.
- Verheul, Ingeborg und IFLA. 2006. *Networking for Digital Preservation: Current Practice in 15 National Libraries*. (IFLA Publications; 119) München: K. G. Saur.

