Raman Ganguly

Workflow-Modell für das Datenmanagement

Handbuch Repositorienmanagement, Hg. v. Blumesberger et al., 2024, S. 103–120
https://doi.org/10.25364/97839033742327

Raman Ganguly, Universität Wien, Zentraler Informatikdienst, raman.ganguly@univie.ac.at |
ORCID iD: 0000-0002-9837-0047

Zusammenfassung

Das Workflow-Modell wurde an der Universität Wien entwickelt und dient der Unterstützung des Forschungsdatenmanagementsupports. Es soll darstellen, wie die Daten in eine zentrale Infrastruktur zur Aufbewahrung gelangen und wo welche Aufgaben und Verantwortungen der handelnden Personen liegen. Die Darstellung ist so generisch wie möglich, damit möglichst viele Anwendungsfälle abgedeckt werden können. Es soll keine konkreten Anforderungen darstellen, die umzusetzen wären, sondern wesentlichen Punkte, die das Datenmanagement an die Datenlieferant:innen stellt und umgekehrt. Die im Modell verwendeten vier Phasen können weiter differenziert werden. So kann dieses Modell auch als Basis für eine Workflowdarstellung im Bereich Open Educational Resources eingesetzt werden.

Schlagwörter: Forschungsdatenmanagement; Support; Datenaufbewahrung; Digitale Objekte; Langzeitarchivierung

Abstract

A Workflow Model for Data Management

The workflow model was developed at the University of Vienna with the purpose to support research data management. It is primarily used to show how the data are preserved in a central infrastructure and where which tasks and responsibilities lie. The representation is as generic as possible so that as many use cases as possible can be covered with it. It does not serve to implement a specific requirement, but is intended to represent the essential requirements of data management on the data suppliers and vice versa. The four phases used in the model can be further differentiated for specific use cases. This model has also served as the basis for a workflow representation in the area of open educational resources.

Keywords: Research data management; support; data preservation; digital objects; long-term preservation

1. Entstehung des Workflow-Modells

Seit dem Start des Repositoriums PHAIDRA an der Universität Wien 2008 können alle Mitarbeiter:innen und Student:innen dieses Service nutzen. Damals wurde angenommen, dass die meisten Personen Daten mittels Einzelupload über das Webinterface von PHAIDRA hochladen und diese Daten ebenfalls Objekt für Objekt mit den entsprechenden Metadaten beschreiben würden. Bei der Einführung des Services entstanden zwei Herausforderungen in der Kommunikation mit den Benutzer:innen: Erstens war vielen Benutzer:innen der Begriff Metadaten nicht geläufig, auch die Frage, wie die Daten am besten beschrieben werden können, war nicht gelöst. Zweitens war die Annahme falsch, dass das Webinterface für den Upload ausreichend sein würde. Schon bald zeigte sich, dass besonders an der Schnittstelle zwischen Datenproduzent:innen und Datenmanagement ein großer Aufwand an zusätzlichen Entwicklungen nötig ist.

Im weiteren Verlauf der Beratungen und Durchführung von Projekten stellte sich zusätzlich heraus, dass die Kommunikation schwierig ist. Das Team vom Datenmanagement kann sich teilweise nur bedingt den Prozess der Datenerstellung vorstellen bzw. weiß sehr wenig darüber, wie die Daten aussehen, die in das Repositorium gelangen sollen. Die Forscher:innen ihrerseits haben ein ähnliches Problem damit. Sie verstehen nur bedingt, was der Forschungsdatensupport von ihnen benötigt und wie die Daten an das Repositorium geliefert werden sollen, damit die Daten langfristig aufbewahrt werden können. Es war daher notwendig, eine anschauliche Methode für die Beratung zu entwickeln, die auf möglichst viele Bereiche anwendbar ist.

2. Anforderungen an das Workflow-Modell

Das Modell dient der Kommunikation mit den Personen, die keine bzw. nur wenig Vorerfahrung mit Datenmanagement haben, daher müssen in den Beratungen alle wichtigen Terminologien erklärt bzw. damit in Einklang gebracht werden. Weiters muss es, wie schon angesprochen, möglichst generisch sein, damit alle wissenschaftlichen Disziplinen und unterschiedlichen Fälle für die Übertragung der Daten in das Repositorium abgebildet werden können. Zusätzlich zur Einfachheit und Allgemeingültigkeit sollte das Modell auch alle entstehenden Aufwände transparent machen, damit diese klar sind und von beiden Seiten zeitlich und budgetär geplant werden können.

Zusammengefasst sind die Anforderungen an dieses Modell:

- Verständlichkeit
- generisch sein
- Transparenz
- Vollständigkeit

3. Entwicklung des Modells

Die erste Orientierung bietet das Open Archival Information System (OAIS)-Modell, das ein Referenzmodell für die Archivierung von Daten ist. Genau beschrieben ist es im sogenannten Magenta Book von Consultative Committee for Space Data Systems (CCSDS), welches ein Beratungskomitee für Weltraumdatensysteme ist.

In diesem Magenta Book wird Folgendes definiert:

> An OAIS is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community. It meets a set of such responsibilities as defined in this document, and this allows an OAIS Archive to be distinguished from other uses of the term 'archive'. The term 'Open' in OAIS is used to imply that this Recommendation, as well as future related Recommendations and standards, are developed in open forums, and it does not imply that access to the Archive is unrestricted.[1]

Es stehen hier die Organisation und die Personen im Vordergrund, die für die Datenarchivierung verantwortlich sind. Das Modell beschränkt sich nicht nur auf offene Daten, sondern das Open im Namen bezieht sich auch auf die Implementierung des Modells und nicht auf die Daten selbst. Das Workflow-Modell selbst ist sehr komplex und eher an Expert:innen des Datenmanagements gerichtet. Es ist sehr generisch und zeichnet den Datenfluss gut nach.

Das Fundament für das Workflow-Modell basiert auf diesem OAIS-Modell, es muss nur noch vereinfacht und auf den Bedarf der Kommunikation beim Support umgelegt werden. Es ist ein Modell für die Implementierung von Archivsystemen in einer Organisation und die Basis einer ISO-Norm, daher auch komplex und an ein Fachpublikum gerichtet. Das Workflow-Modell hat eine geringere Granularität, also eine höhere, abstraktere Sichtweise auf das Datenmanagement. Dennoch werden wichtige Begriffe übernommen wie z. B. Ingest und Data Management, beide werden im Folgenden noch genau beschrieben. Im OAIS-Modell wird zwischen

1 CCSDS (2012), S. 1.

Submission Information Package (SIP), Archival Information Package (AIP) und Dissemination Information Package (DIP) unterschieden[2], die wichtig für die Implementierung des Archivsystems sind, aber nicht für den Datenfluss eines im Einsatz befindlichen Archivsystems.

Beim OAIS-Modell werden drei Phasen oder Organisationseinheiten definiert: Producer, Management und Consumer[3]. Beim Workflow-Modell, das mehr den Fluss der Daten und nicht die Organisation widerspiegeln soll, wurde zusätzlich der Ingest als eigene Phase eingeführt. Hier soll insbesondere auf den Aufwand, der beim Ingest entsteht, hingewiesen werden.

4. Das Workflow-Modell

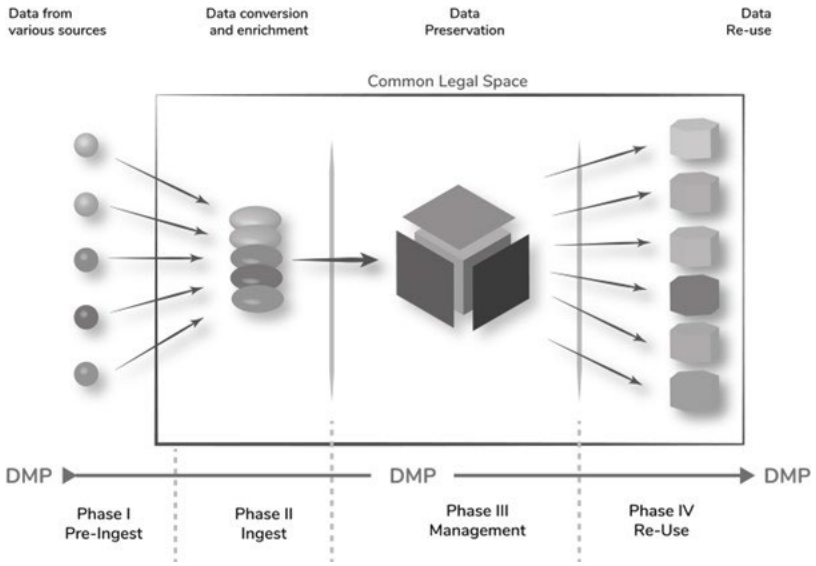

Abbildung 1: Workflow-Modell[4]

2 Ebd., S. 4-35.
3 Ebd. S. 2.
4 https://hdl.handle.net/11353/10.527220

Abbildung 1 zeigt das gesamte Workflow-Modell. Der Kern besteht aus vier Phasen: Pre-Ingest, Ingest, Management und Re-Use. Bei den Phasen gibt es unterschiedliche handelnde und zuständige Personen. Die Pre-Ingest-Phase wird von den Datenproduzent:innen durchgeführt und liegt auch in deren Verantwortlichkeit, die Management-Phase vom Datenmanagementteam. Die Phasen I (Ingest) und IV (Re-Use) sind jene Phasen, in denen die Daten dem System übergeben bzw. aus dem System geholt werden. Im Idealfall werden die abgerufenen Daten erneut verwendet und wieder in den Workflow eingespeist.

Sämtliche Prozessschritte umfassen das Datenmanagement. Hier muss zwischen dem Prozess und der Rolle des Datenmanagements unterschieden werden. Die Rolle des Datenmanagements wird üblicherweise von einer zentralen Einrichtung übernommen, die Daten langfristig zur Verfügung stellen kann. Der Prozess hingegen erstreckt sich über den gesamten Daten-Lifecycle, also von der Erstellung, über die Nachnutzung bis zum eventuellen Löschen von Daten. Viele Daten sollen jedoch für die Ewigkeit aufbewahrt werden, da entfällt natürlich die Löschung.

Im Workflow-Modell ist auch der Data-Management-Plan (DMP) eingezeichnet, der bereits vor der ersten Phase beginnt und nach der letzten Phase andauert. Der DMP umspannt den gesamten Zeitraum, vom Beginn, also der Planung der zu generierenden Daten, der Entstehung der Daten, bis zur Nachnutzung der Daten. Ein DMP wird aus Sicht eines Projekts erstellt und beschreibt auch, wie Daten über das Ende des Projekts aufbewahrt, geteilt und verwaltet werden müssen.

Ein gemeinsamer rechtlicher Rahmen, der Common Legal Space der Daten, wird durch das Viereck gekennzeichnet, das sich vom Ingest bis zur Nachnutzung erstreckt. Beim Ingest sollen die Rechte so geklärt werden, dass die Daten ohne weiteres durch die anderen Phasen durchlaufen können. Besonders wichtig ist dies bei der Nachnutzung, denn es muss für die Nachnutzer:innen klar und eindeutig sein, in welcher Form sie die Daten nutzen können.

4.1. Pre-Ingest

In der Pre-Ingest-Phase werden die Daten erzeugt, daher sind in dieser Phase die handelnden und verantwortlichen Personen die Datenproduzent:innen. Meistens werden die Daten im Rahmen eines Forschungsprojekts erstellt, daher kann es sich auch um ein längeres Vorhaben handeln, wie zum Beispiel die Digitalisierung von Sammlungen. Der Fokus liegt auf der Erstellung von hochqualitativen Daten. Hier wird in der Regel noch keine Rücksicht auf eine spätere Datenaufbewahrung gelegt.

Die Aufbewahrung wird relevant, wenn bei der Erfassung gewisse Rahmenbedingungen erfüllt werden können, die einen späteren Ingest ermöglichen, aber keinen Einfluss auf die Qualität der Daten haben.

Einem DMP entsprechend ist es wie bereits oben vermerkt sinnvoll, bereits bei der Datenerfassung auf die rechtlichen Rahmenbedingungen für eine spätere Aufbewahrung zu achten. So können bei der Erhebung gleich die entsprechenden Einverständniserklärungen oder Rechte eingeholt werden, die bei einem späteren Zeitpunkt vermutlich mit einem erheblich höheren Aufwand erbracht werden müssten.

Auch kann beim Pre-Ingest auf die Dokumentation und Beschreibung (später als Metadaten verwendbar) sowie auf das Datenformat geachtet werden. Die Metadaten gleich bei der Entstehung zu erfassen, spart Aufwand und man sollte außerdem Formate verwenden, die für die langfristige Aufbewahrung geeignet sind, sofern dies möglich ist. All diese Maßnahmen können den Aufwand beim Ingest von Seiten der Datenproduzent:innen verringern. Der Ingest findet aus Sicht des Projekts meist gegen Ende statt, da meist erst dann die Daten zur Verfügung stehen. Gegen Ende des Projekts sind eingeplante Zeitpuffer jedoch meist schon aufgebraucht und manche Projektmitarbeiter:innen haben bereits Verpflichtungen in neuen Projekten. Dadurch werden die Ressourcen knapp.

4.2. Ingest

Beim Ingest werden die Daten von den Produzent:innen an das Datenmanagement übergeben. Die Daten werden dabei für die Aufbewahrung aufbereitet. Konkret bedeutet das, dass aus den Daten digitale Objekte werden. Nachdem Format und Rechte geprüft wurden, erfolgt eine Datenanreicherung.

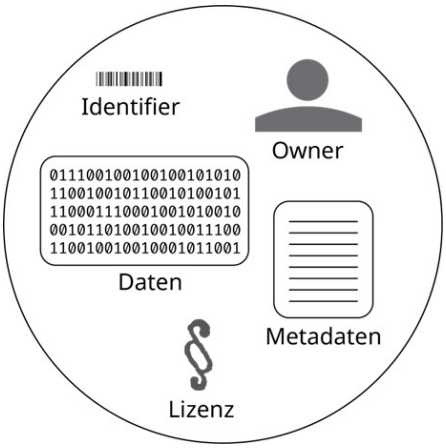

Abbildung 2: Digitales Objekt

Ein Digitales Objekt besteht aus den folgenden Teilen:

- den Daten selbst
- Persistenten Identifikatoren (PI)
- Metadaten, also der Beschreibung der Daten
- der Lizenz
- der/dem Rechteinhaber:in

Für jedes Objekt müssen möglichst früh die Urheberrechte geklärt und die Art der Nutzung klargestellt werden. Die eingeschränkteste Form für die Nachnutzung ist „alle Rechte vorbehalten", damit gelten die Regelungen des verankerten Urheberrechts. Von den meisten Forschungsförderern werden offene Lizenzen gefordert, sofern dies möglich ist. Hier werden in den meisten Fällen die Creative-Commons-Lizenzen vergeben. Die offenste Form, CC0, lässt Nachnutzer:innen alle Freiheiten, sogar ohne die Nennung des/der Urheber:in[5]. Nach dem österreichischen Urheberrecht ist diese Form der Nutzung nur bedingt möglich[6], somit ist CC BY die offenste Form. Hier gilt gleiches wie bei CC0, nur dass die Quelle der Daten, also die Namensnennung der Datenerfasser:innen gefordert ist.

5 Amini, S.; Blechl, G. et al. (2016)
6 Kucsko, G.; Zemann, A. (2017)

4.3. Management

Die Management-Phase ist die Kernphase der Aufgabe der langfristigen Aufbewahrung der Daten. Dabei werden die Daten über die geforderte Zeit in der geforderten Qualität aufrechterhalten und den Personen zur Verfügung gestellt, die das Recht haben, die Daten zu nutzen.

Mit dieser Phase ist der langfristige Betrieb einer Infrastruktur verknüpft, mit der die Daten aufbewahrt werden können. In den meisten Fällen handelt es sich um Repositorien, in denen Daten vorgehalten werden. Diese Repositorien müssen für die Datenintegrität sorgen und eine Landingpage zur Verfügung stellen, über die die Metadaten permanent auffindbar sind. Eine Landingpage ist eine Webseite, die mit dem Persistenten Identifier verknüpft ist und somit die Persistenz des Ortes im Netz garantiert. Es ist die Seite, auf die verlinkt wird und die den Zugang zu den Daten ermöglicht. Diese Seite soll auch für Personen erreichbar sein, die keinen direkten Zugang zu den Daten haben dürfen, damit ist eine Zitierbarkeit in Publikationen gewährleistet, auch wenn es sich um geschlossene Daten handelt. Dies ist ein wichtiges Kriterium zur Erfüllung der FAIR-Data-Prinzipien.[7]

Die Infrastruktur geht weit über ein Repositorium hinaus, da es einerseits Daten geben kann, die nicht in Form von Dateien gespeichert sind, und andererseits auch die Langzeitarchivierung betrachtet werden muss. Mehr dazu im Kapitel 5.

4.4. Re-Use

In der Re-Use-Phase findet die zweite Übergabe statt. Hier werden die Daten an die Personen übergeben, die diese nachnutzen wollen. Es ist jene Phase, die die Notwendigkeit für das Datenmanagement begründet. Ohne potentiellen Re-Use ist der gesamte Aufwand für das Datenmanagement nutzlos, weil sonst nur Daten gesammelt werden. Oft sind die möglichen Re-Use-Szenarien am Ende eines Projekts nicht vollumfänglich erkennbar, dennoch sollte man bei der Auswahl der Daten für die Aufbewahrung an die Nachnutzung denken. Natürlich gibt es offensichtliche und/oder vorgeschriebene Szenarien, wie etwa die Nachvollziehbarkeit und/oder Reproduzierbarkeit der Ergebnisse.

7 Die FAIR-Data-Prinzipien sind ein Akronym für Findable, Accessible, Interoperable, Reuseable und bilden das Grundprinzip für das Datenmanagement bei der European Open Science Cloud (EOSC).

Im Lauf der Zeit können die Daten auch in einem ganz anderen Kontext wiederverwendet werden. So werden z. B. heute Jahrhunderte alte Logbücher aus der Schifffahrt für die Klimaforschung verwendet[8]. Niemand konnte damals den Wert der Daten für die heutige Forschung erkennen.

Werden Daten an Nachnutzer:innen übergeben, so werden Kopien der Daten weitergegeben. Das eigentliche Digitale Objekt verbleibt im Repositorium. Dies ist von der Art und Weise zwar völlig logisch, da im digitalen Raum immer Kopien weitergegeben werden, es entstehen daraus aber verschiedene Probleme: Da es sich um eine Kopie handelt, kann nicht mehr nachvollzogen werden, was mit dieser Kopie alles passiert, z. B. die Verbreitung der Kopie auf anderen Kanälen. Hier ist die Lizenz besonders wichtig, da diese bestimmt, was erlaubt ist und was nicht. Aus Sicht des Repositorien-Managements muss darauf vertraut werden, dass sich die Nachnutzer:innen entsprechend der Lizenz rechtskonform verhalten. Es liegt nicht in der Verantwortung des Managements, die Einhaltung zu kontrollieren.

Da es sich um eine Kopie der Daten handelt, kann nicht verhindert werden, dass die Daten außerhalb des Repositoriums verbreitet werden, falls dies die Lizenz zulässt. Daher ist eine nachträgliche Änderung der Lizenz nicht zulässig, da nicht nachvollziehbar ist, wann die Lizenzen geändert wurden und unter welcher Lizenz dann die jeweilige Kopie steht.

Ein weiteres Thema ist die Authentizität der Daten. Wird die Kopie der Daten, die außerhalb des Repositoriums liegt, geändert, so kann dies nicht nachvollzogen werden. Nur durch eine korrekte Zitierung der Daten, die wieder auf das Original im Repositorium verweist, kann die Korrektheit überprüft werden.

5. Die Phasen aus Sicht der Daten

Nun betrachten wir dieses Modell aus Sicht der Daten und was mit den Daten in den jeweiligen Phasen geschieht. Zunächst wird geklärt, welche Daten in das Datenmanagement überführt werden können und wie die Daten im Ingest zu digitalen Objekten werden. Anschließend wird beschrieben, welche Maßnahmen getroffen werden müssen, damit die Qualität der Daten erhalten bleibt. Der Re-Use wird in dieser Darstellung nur am Rande betrachtet, da es sich aus Sicht des Modells um eine reine Kopie handelt. Dennoch wird hier unabhängig vom Modell noch dargestellt, wie Infrastrukturen des Datenmanagements den Re-Use erleichtern können.

8 Becker, R. (2019)

5.1. Pre-Ingest: Art von Daten, die entstehen

In einem Forschungsprozess kann jede erdenkliche Art von Daten entstehen bzw. verwendet werden. Erst der Kontext der Forschung definiert das Forschungsdatum. Besonders gut kann man das im Datenmanagement von institutionellen Infrastrukturen an Universitäten beobachten. Mit der großen Vielzahl an unterschiedlichen Disziplinen kommt es auch zu einer hohen Variation an unterschiedlichen Arten von Daten.

Daten können in Form von Textdokumenten, Bildern, Audio/Video-Daten, Messreihen, 3D-Objekten, Software, Datenbanken usw. vorliegen. Alle diese Daten sind unterschiedlich komplex und müssen teilweise in verschiedenen Infrastrukturen aufbewahrt werden. Dabei können erfahrungsgemäß diese drei Typen unterschieden werden: Dateien, Software und Datenbanken. Bei der folgenden Betrachtung handelt es sich um eine modellhafte Vereinfachung, bei der Big Data zunächst einmal ausgeblendet wird. Beim Thema Big Data entstehen neue Anforderungen an die Infrastruktur, die den Rahmen hier sprengen würden.

5.1.1. Dateien

Bei Daten kann es sich um Texte, Bilder, Audio oder Video-Daten handeln. Worum es sich genau handelt, bestimmt das Format und kann meist über die Dateiendung erkannt werden. So sind Textdateien, die mit Microsoft Word geschrieben wurden, mit der Endung .docx gekennzeichnet. Auch .pdf ist ein sehr übliches und bekanntes Format, das für das Datenmanagement wichtig ist.

Wesentlich ist die Abgeschlossenheit der Datei. Sie kann mit einer entsprechenden Software geöffnet und verarbeitet werden. Hier gilt es zu unterscheiden, ob Daten in einem offenen oder geschlossenen Format gespeichert werden. Bei einem offenen Format ist allgemein einsehbar, wie die Spezifikation, also die Struktur der Datei, ist. Unterschiedliche Hersteller:innen von Software können und dürfen entsprechende Programme schreiben, die eine Datei in diesem Format verarbeiten kann. Bei geschlossenen Formaten handelt es sich meist um Formate von Firmen, die oft nur mit der Software des Herstellers verarbeitet werden dürfen. Die Wahl des Formats hat eine direkte Auswirkung auf die Interoperabilität und wie einfach oder schwierig es ist, die Daten nachzunutzen.

5.1.2. Software

Bei der Software stellt sich die Frage, ob der Software-Code archiviert werden soll, oder ob die Software selbst in Betrieb gehalten werden soll. Aus der Praxis gesprochen, ist diese Frage früh zu klären, da unter dem abstrakten Begriff Archivierung

oder sogar langfristige Zurverfügungstellung sehr unterschiedliche Dinge verstanden werden.

Soll nur der Software-Code aufbewahrt werden, kann die Software wie Dateien behandelt werden. Im Prinzip handelt es sich bei Software-Code nur um Text, der in einem offenen Dateiformat gespeichert ist. Wichtig ist, dass neben dem Code auch noch eine umfangreiche Dokumentation abgelegt wird. Diese geht in den meisten Fällen über die klassischen Metadaten hinaus, da erklärt werden muss, welche Schritte und welche Infrastrukturen notwendig sind, um die Software in Betrieb zu nehmen. Dafür haben sich in der Software-Entwicklung gängige Praktiken etabliert (z. B. Versionierung), die vom Datenmanagement übernommen werden sollten. Es ist auch ratsam, ein Tool für die Versionskontrolle zur Verfügung zu stellen. Nach derzeitigem Stand der Technik ist Git[9] am sinnvollsten einzusetzen. Auf dem Versionskontrolle-Tool Git aufbauend gibt es Tools mit Webinterfaces, die sich gut für das Datenmanagement eignen. GitHub[10], GitLab[11] und Bitbucket[12] sind die drei bekanntesten Tools, wobei GitLab auch unter einer Open-Source-Lizenz zur Verfügung steht.

Von klassischen Repositorien aus kann man eine Verbindung mit einem Git-Repositorium herstellen, in dem auch die Dokumentation zur Software vorhanden sein kann. Der Begriff Repositorium wird bei der Versionskontrolle anders verwendet als im Datenmanagement. Repositorium wird hier als Einheit verstanden, die eine Software vorhält, und nicht als übergeordnetes System, in dem die Objekte liegen. Im Repositorium kann es dann zu mehreren Versionen einer Software kommen und jede Version hat einen eindeutigen Identifikator, der wiederum die Basis für ein digitales Objekt in einem klassischen Datenrepositorium sein kann. Somit können die Vorteile aus beiden Bereichen miteinander verknüpft werden. Es gibt eine Landingpage mit Metadaten, DOI (falls gewünscht), Owner usw. und man kann den Source-Code so nutzen, wie es in der Software-Entwicklung üblich ist. Man findet im Internet viele Beispiele zur Funktionsweise von Git. Einen guten Überblick über den Aufbau und die Möglichkeiten von Git findet sich auf Developer-Info-Seiten von IBM[13].

9 git – fast-version-control; Webseite des Projekts: https://git-scm.com/
10 https://github.com/
11 https://about.gitlab.com/
12 https://bitbucket.org/
13 https://developer.ibm.com/tutorials/d-learn-workings-git/

Soll die Software in Betrieb gehalten werden, wird es kompliziert. Eine Software ist nie abgeschlossen und fehlerfrei, und hängt sie auch von anderen Software-Elementen ab. Sie benötigt ein Ökosystem, in dem sie laufen kann. Das Ökosystem besteht aus Software und Hardware, die sich wiederum im Laufe der Zeit ändert. Erfahrungsgemäß sind die Zyklen der Änderungen von Hard- und Software, die die Infrastruktur des Ökosystems bilden, teilweise sehr kurz (ein bis zwei Jahre). Daher entsteht ein hoher Aufwand, die Software in Betrieb zu halten, da Änderungen im Ökosystem sich auf die Software auswirken können. Dieser Aufwand kann nicht mehr vom Projekt getragen werden, da es weit über das Projektende hinausreicht. Für eine zentrale Einheit des Datenmanagements, die die Software noch nie zuvor gesehen hat, ist es ein sehr hoher Aufwand, den Betrieb zu übernehmen. Es kann auch nur schwer die Qualitätssicherung übernommen werden, wenn das Domänenwissen der Software, also das Wissen, wofür die Software eingesetzt wurde, nicht mehr vorhanden ist.

5.1.3. Datenbanken

Datenbanken sind noch komplexer als Software. Auch hier stellt sich vorrangig die Frage, ob die Datenbank in Betrieb gehalten werden soll oder nicht. Zusätzlich muss geklärt werden, was genau unter dem Begriff Datenbank verstanden wird und welches Datenbankkonzept verwendet wird.

Grundsätzlich sind Datenbanken Systeme, in denen Daten nach einer bestimmten Struktur, dem Datenmodell, abgelegt werden. Oft sind die Datenbanken mit einer Software verknüpft, die für die Eingabe und Darstellung der Daten verantwortlich ist. Nicht nur die Datenbank als solche, sondern auch die dazugehörige Software für die Ein- und Ausgabe von Daten muss in Betrieb gehalten werden. Auch wenn es nur die Datenbank betrifft, muss hierfür die Betreuung des Betriebs mit eingerechnet werden.

5.2. Ingest

Beim Ingest wird die Art der Daten geprüft und, wie diese aufbewahrt werden können. Hierbei ist die Expertise des Datenmanagements gefragt. Beim Datenmanagement kommt es auf eine Balance zwischen technischen und nicht-technischen Aufgaben an. Nur wenn beide Bereiche gemeinsam betrachtet werden, können Lösungen für Infrastrukturen gefunden werden. Der Ingest sorgt nun dafür, dass Daten in die Infrastrukturen für die Aufbewahrung überführt werden. Das Knowhow des nicht-technischen Bereichs kommt aus dem Bibliotheks- und Archivwesen, das

eine lange Tradition hat, Wissen zu bewahren. Dieses muss nun mit den neuen digitalen Techniken kombiniert werden, um digitale Daten aufzubewahren. Erst die Kombination schafft die Möglichkeit, der Flüchtigkeit von digitalen Daten entgegenzuwirken, damit diese auch nach mehreren Generationen nachnutzbar sind.

Daten haben ihren Ursprung in den jeweiligen wissenschaftlichen Domänen. Nur mit ihrer Hilfe können aus den Daten digitale Objekte erstellt werden, da das Datenmanagement selbst keine Beschreibung und notwendigen Informationen für die Metadaten hat bzw. erstellen kann. Wenn verstanden wird, wie die Daten der einzelnen Domänen entstehen und wie deren Prozesse ausgestaltet sind, können die Infrastrukturen, in denen die Daten entstehen, mit den Infrastrukturen zusammengeführt werden, in denen die Daten aufbewahrt werden. Auch können so automatisiert Metadaten erstellt werden. Ziel ist es, beim Ingest einen möglichst geringen Aufwand zu haben und bereits bestehende Metadaten bei der Erzeugung des digitalen Objektes zu nutzen.

Die große Herausforderung hierbei ist, dass sowohl Bibliotheken und IT-Services als auch die Fachdomain ihre ganz eigenen Traditionen und Arbeitsweisen haben. Nur durch eine intensive Zusammenarbeit kann Verständnis für die jeweils andere Tradition entwickelt und eine gemeinsame Sprache gefunden werden, die das Fundament für die Entwicklung von kombinierten Infrastrukturen für die Entstehung, Analyse und Aufbewahrung von Forschungsdaten ist.

5.3. Datenmanagement

Was die Daten anbelangt, geht es beim Datenmanagement um die Erhaltung der Qualität über eine definierte Zeit. Der Zeitraum kann kurz, aber auch sehr lang sein. Dabei stellt sich die Frage, welche Qualität über einen gewissen Zeitraum aufrechterhalten werden soll. Je länger dieser Zeitraum ist, umso komplexer ist die Aufgabe. Das liegt daran, dass das digitale Zeitalter noch sehr jung ist und wir noch nicht erahnen können, wie die Daten in hundert oder mehr Jahren gespeichert und gelesen werden, und wir auf keine Erfahrungen aus einer längeren Vergangenheit zurückgreifen können. Von einer Stabilität wie bei der geschriebenen Informationsweitergabe auf Papier, Tontafeln oder sonstigen Trägermaterialien können wir im digitalen Raum nicht ausgehen. Es kommt auch zu permanenten Veränderungen bei der Software und bei der Hardware, mit denen digitale Daten genutzt und abgerufen werden.

Auch bisher mussten wir die Träger der Information erhalten, damit die Qualität der Daten stabil bleibt. Übertragen auf die digitale Welt ist die Hard- und Software das Medium, das es zu bewahren gilt. Nur sind die Daten nun unabhängig von deren

Medium und müssen daher getrennt betrachtet werden. Anders als im analogen Raum können sich die Daten verlustfrei von Medium zu Medium im virtuellen Raum bewegen. Es kommt immer darauf an, ob der Zielort die Daten auch verarbeiten und deren Inhalt preisgeben kann. Die Auflösung von Raum bedeutet aber nicht eine völlige Lösung von physikalischen Rahmenbedingen. Auf der physikalischen Ebene sind die 0 und 1, in denen Daten gespeichert werden und auf der Hardware vorliegen, die Basis der Daten. Die Formate definieren logische Sinneinheiten, die der Serie von 0 und 1 eine Struktur verleihen. Erst durch das Format können wir wissen, ob es sich bei der 0 und 1 um ein Dokument, ein Bild, ein Video usw. handelt. Mit Hilfe von Software werden dann die Inhalte dargestellt. Nur Software, die das jeweilige Format auch kennt, kann dies leisten. Dazu braucht es die entsprechenden Programme.

5.3.1. Bitstream

Die Reihe von 0 und 1, die in einer Hardware gespeichert ist und die Basis der Sinneinheiten bildet, wird Bitstream genannt. Für eine Aufbewahrung der Daten muss sichergestellt werden, dass dieser Bitstream nicht verändert wird. Die Library of Congress spricht sogar davon, dass die Erhaltung des Bitstreams ein Eckpfeiler der digitalen Aufbewahrung[14] ist.

Veränderungen können durch Prozesse passieren, etwa durch das Kopieren oder die Übertragung von Daten. Auch durch Fehler in Speichermedien kann es zum sogenannten „Bit-Rot“ kommen. Dies sind Fehler im Bitstream, die sich bei der Speicherung auf dem physikalischen Medium im Lauf der Zeit einschleichen können[15].

Bei der Bitstream Preservation geht es darum, den Bitstream regelmäßig daraufhin zu prüfen, ob es zu Veränderungen gekommen ist. Dies kann mittels eines Hashwerts[16] geprüft werden. Ein Hashwert kann als Prüfsumme verwendet werden, da er eindeutig ist. Verändert sich der Bitstream, so verändert sich auch der Hashwert und Fehler können erkannt werden. Die Integrität muss anschließend wieder hergestellt werden. Daher ist es notwendig, bei der Bitstream Preservation Kopien der Daten zu haben, die in unterschiedlichen Datenpools gehalten werden. Alle Datenpools müssen dahingehend regelmäßig geprüft werden, ob die Integrität der Daten noch vorhanden ist.

14 Library of Congress (n.d.)
15 https://en.wikipedia.org/wiki/Data_degradation
16 https://de.wikipedia.org/wiki/Hashfunktion

5.3.2. Formate

Genauso wie die Software verändern sich auch die Formate. Es werden nicht nur neue Formate entwickelt, sondern auch alte Formate werden obsolet, oder es kommt zu neuen Versionen bestehender Formate. Z. B. gibt es die Dateiendung .doc, die für das Dokumentenformat von Microsoft Word steht, schon sehr lange. Das Format selbst hat sich im Laufe der Zeit stark verändert, da neue Funktionen in das Programm Word aufgenommen wurden, die auch im Format abgebildet werden mussten. Ein Beispiel wäre die Library of Congress, welche die Änderungen dokumentiert, da sie es für ihr eigenes Datenmanagement benötigt[17].

Die alten Formate können von neuerer Software oft nicht mehr gelesen werden. Damit man auf die Inhalte der Daten zugreifen kann, muss man entweder die alte Software erhalten oder das Format auf die neue Version migrieren. Beides ist mit einem erheblichen Aufwand verbunden.

Bei der Erhaltung der Software muss berücksichtigt werden, dass alte Software auch ein altes Betriebssystem benötigt. Es muss das gesamte alte Ökosystem erhalten werden. Mit neuen Technologien funktioniert es schon recht gut, dies zu erhalten. Z. B. helfen sogenannte Virtuelle Maschinen dabei, ein altes Softwaresystem zu betreiben. Eine Virtuelle Maschine simuliert eine Hardware[18] und man kann auf einem Computer mehrere Betriebssysteme gleichzeitig laufen lassen. Auch kann man diese Technik dazu nutzen, um ein Betriebssystem zum Laufen zu bringen, das schon älter ist. Nur geht dies nicht uneingeschränkt: Die alten Apple-Computer beispielsweise hatten einen anderen Prozessor und in solchen Fällen wird es schon sehr schwierig, das Betriebssystem in einer virtuellen Maschine zum Laufen zu bringen.

Formate in eine aktuellere Version zu migrieren, ist eine weitere Möglichkeit, die Lesbarkeit der Datei aufrecht zu erhalten. Damit das möglich ist, muss das Format offen sein, so dass das Ausgangsformat auch vollständig in seiner Struktur bekannt ist. Außerdem muss bei der langfristigen Aufbewahrung darauf geachtet werden, dass möglichst nur offene Formate verwendet werden. Damit kann man die Abhängigkeit von Format und Softwarehersteller trennen. Ansonsten kann es passieren, dass es keine Software mehr gibt, die das Format lesen kann, falls der Hersteller die Software aufgibt oder in Konkurs geht.

Neben einem offenen Format ist auch wichtig zu wissen, ob das Format die Daten komprimiert hat. Falls die Daten komprimiert sind, ist es notwendig zu wissen, ob

17 Vgl. https://www.loc.gov/preservation/digital/formats/fdd/fdd000509.shtml
18 https://de.wikipedia.org/wiki/Virtuelle_Maschine

diese verlustfrei oder verlustbehaftet komprimiert sind. Für die Aufbewahrung wird empfohlen, wenn möglich, ein offenes und unkomprimiertes (bzw. verlustfrei komprimiertes) Format zu wählen. Da dies nicht immer möglich ist, muss das Datenmanagement Kompromisse in diesem Bereich eingehen, vor allem bei Videodaten.

5.4. Re-Use

Beim Re-Use wird dem/der Nachnutzer:in eine Kopie der Daten zur Verfügung gestellt. Bei Dateien in einer Größe, die leicht über das Internet verbreitet werden kann, ist dies ein einfaches Verfahren. Es wird meistens über Download-Links geregelt. Bei Software und Datenbanken ist das schon etwas schwieriger.

Wenn die Daten in einer Software integriert sind, muss entweder die Software so zur Verfügung gestellt werden, dass sie den/die Nachnutzer:in auf einer eigenen Infrastruktur in Betrieb nehmen kann, oder die Software ist in einer lauffähigen Version vorhanden, die auch die Nachnutzer:innen verwenden können.

Bei Datenbanken ist es ähnlich, auch hier sind die Daten nicht direkt über eine Datei erreichbar. Es werden das Datenbanksystem benötigt, die Struktur, in der die Daten abgelegt sind, und natürlich die Daten selbst. Auch hier gibt es die Möglichkeit, die Struktur und Daten zur Verfügung zu stellen. Dann müssen die Nachnutzer:innen das Datenbanksystem selbst betreiben und die Struktur sowie die Daten in das Datenbanksystem importieren. Natürlich kann auch das Datenmanagement die Datenbank im Betrieb halten und Nachnutzer:innen direkt auf die Daten zugreifen lassen.

Das Team des Datenmanagements muss sich Wege überlegen, wie die Daten von den Archivsystemen zu den Computersystemen übertragen werden können. Hier bedarf es einer Integration zu den Forschungsinfrastrukturen. Im Idealfall hat man für den Ingest der Daten bereits die Schnittstellen aufgebaut.

6. Conclusio

Das Workflow-Modell ermöglicht die Kommunikation mit den Forscher:innen und ist ein Werkzeug für den Support und die Beratung. Forscher:innen sind mit den neuen Anforderungen der Forschungsförderer zum Teil überfordert und sie wissen nicht, wie Datenmanagement funktioniert und was im Bereich Datenmanagement von ihnen erwartet wird. Das Modell erklärt sehr gut, wo und wann die Aufwände für die Aufbewahrung von Daten entstehen und wer dafür verantwortlich ist. Dabei ist es wesentlich, die Anforderungen von ihnen auf dieses Modell zu übertragen. Es

kommt immer darauf an, wie die Daten entstehen und wohin sie fließen. Das Modell nimmt, wie bereits beschrieben, die Perspektive der Daten ein und dokumentiert ihren Fluss von der Entstehung bis zur Nachnutzung. Das Modell kann auch als Grundlage zur Vermittlung der Anforderungen des Datenmanagements herangezogen werden.

Bibliografie

Amini, Seyavash; Blechl, Guido; Hamdi, Djawaneh et al. (2015): Cluster E: FAQs zu Creative-Commons-Lizenzen unter besonderer Berücksichtigung der Wissenschaft. https://phaidra.univie.ac.at/o:459183

Becker, Rachel (2019): Why Century-Old Ship Logs Are Key to Today's Climate Research. The Verge, 03.05.2019. https://www.theverge.com/2019/5/3/18528638/southern-weather-discovery-ship-logs-climate-change (abgerufen am 08.03.2023)

CCSDS. The Consultative Committee for Space Data Systems (ed.) (2012): Recommendation for Space Data System Practices. Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2. Magenta Book. https://public.ccsds.org/pubs/650x0m2.pdf (abgerufen am 08.03.2023)

Kucsko, Guido; Zemann, Adolf (2017): CC0 1.0 Universal – Beurteilung der Verzichtserklärung und der Lizenzerteilung im Rahmen der Fallback-Klausel nach österreichischem Recht. https://phaidra.univie.ac.at/o:528411

Library of Congress (n.d.): Bit Level Preservation and Long Term Usability. In: Digital Collections Management Compendium. https://www.loc.gov/programs/digital-collections-management/digital-formats/bit-level-preservation-and-long-term-usability/ (abgerufen am 10.02.2023)

Raman Ganguly hat seinen fachlichen Hintergrund in der Softwareentwicklung und Medientechnik. Er leitet die Abteilung IT Support für Research am Zentralen Informatikdienst der Universität Wien und ist für die Entwicklung und den Betrieb von Datenmanagement-Infrastruktur verantwortlich. Seit 2011 beschäftigt er sich mit der Archivierung von digitalen Daten aus der Forschung und Lehre mit dem Schwerpunkt der langfristigen Verfügbarhaltung. Er ist der technische Leiter des an der Universität Wien entwickelten Open-Source-Archivierungssystems PHAIDRA und der technischen Koordination für den internationalen Verbund von PHAIDRA bestehend aus 21 Institutionen. Raman Ganguly berät wissenschaftliche Bibliotheken bei technischen Fragen zum Datenmanagement und ist Vortragender bei den Universitätslehrgängen Data Librarian und Data Steward.