Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften

Johanna Puhl (Universität zu Köln), Peter Andorfer (HAB Wolfenbüttel),
Mareike Höckendorff (Universität Hamburg), Stefan Schmunk (SUB Göttingen),
Juliane Stiller (MPIWG Berlin), Klaus Thoden (MPIWG Berlin)

DARIAH-DE Working Papers Nr. 11

Puhl, J.; Andorfer, P.; Höckendorff, M.; Schmunk, S.; Stiller, J.; Thoden, K.: Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2015
(DARIAH-DE working papers 11)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2015-4
URN: http://nbn-resolving.de/ urn:nbn:de:gbv:7-dariah-2015-4-4

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

---

Abstract: Das vorliegende Dokument beschreibt den aktuellen Diskussionsstand über ein Referenzmodell eines Forschungsdatenzyklus in den digitalen Geisteswissenschaften. Nach einer Bedarfserläuterung werden Begriffe, Funktionen und Abläufe eines solchen Datenzyklus näher analysiert und definiert. Das Dokument schließt mit einem Referenzmodell sowie einem Ausblick auf weiteren Evaluierungsbedarf und daraus resultierende Pläne in dem Forschungsprojekt DARIAH-DE.

Keywords: Forschungsdaten, Forschungsdatenmanagement, Digitale Geisteswissenschaften, Langzeitarchivierung, Forschungsdatenzyklus, Metadaten, Infrastruktur, Forschungsdatensammlungen, Referenzmodell

Research Data, Research Data Management, Digital Humanities, Longterm Preservation, Research Data LifeCycle, Metadata, Infrastructure, Research Data Collections, Reference Model

Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften

Johanna Puhl (Universität zu Köln), Peter Andorfer (HAB Wolfenbüttel), Mareike Höckendorff (Universität Hamburg), Stefan Schmunk (SUB Göttingen), Juliane Stiller (MPIWG Berlin), Klaus Thoden (MPIWG Berlin)

Johanna Puhl, Peter Andorfer, Mareike Höckendorff, Stefan Schmunk, Juliane Stiller, Klaus Thoden: „Diskussion und Definition eines Research Data LifeCycle für die digitalen Geisteswissenschaften". *DARIAH-DE Working Papers* Nr. 11.
Göttingen: DARIAH-DE, 2015. URN: **urn:nbn:de:gbv:7-dariah-2015-4-4.**

Die *DARIAH-DE Working Papers* werden von Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch herausgegeben.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Abstract

Das vorliegende Dokument beschreibt den aktuellen Diskussionsstand über ein Referenzmodell eines Forschungsdatenzyklus in den digitalen Geisteswissenschaften. Nach einer Bedarfserläuterung werden Begriffe, Funktionen und Abläufe eines solchen Datenzyklus näher analysiert und definiert. Das Dokument schließt mit einem Referenzmodell sowie einem Ausblick auf weiteren Evaluierungsbedarf und daraus resultierende Pläne in dem Forschungsprojekt DARIAH-DE.

Inhaltsverzeichnis

1. Einleitung **4**
2. Definitionsbedarf **6**
3. Ziele **8**
4. Begriffe **9**
4.1 Forschungsdaten und -sammlungen 9
4.1.1 Zur Unterscheidung von Primärdaten und Sekundärdaten 9
4.1.2 Der Unterschied zwischen Daten und Objekten 10
4.1.3 Der Unterschied zwischen Metadaten und Daten 11
4.1.4 Definition Forschungsdaten 12
4.2 Dateiformate und Objektklassen 15
4.3 Annotationen 18
4.4 Methoden & Tools 21
4.4.1 Methoden 21
4.4.2 Tools 22
5 Infrastrukturelle Funktionen **25**
5.1 Normalisierung 25
5.2 Vergabe persistenter Identifier 26
5.3 Metadaten-Anreicherung 27
5.4 Kuration und Speicherung (LZA) 29
5.4.1 Exkurs: das OAIS Modell 30
5.4.2 Notwendige Operationen und Empfehlungen für Langzeitarchivierung 32
5.5 Publikation 35
5.6 Peer-Review 36
5.7 Lizensierung 37
5.8 Rollen- und Rechtemanagement 38
6. Entwurf für ein Referenzmodell **41**
6.1 Ein generischer Workflow 42
6.2 Aktivitäten in einem Basis-Forschungsdatenzyklus 44
6.3 Datenmodell 45
7. Fazit **47**
8. Quellenverzeichnis **48**

1. Einleitung

Innerhalb des BMBF geförderten Infrastrukturprojekts DARIAH-DE wurde im Verlaufe der Arbeit eine grundsätzliche theoretische Unterfütterung und auch stärkere Formalisierung von geisteswissenschaftlicher Forschungsarbeit zur weiteren Ausarbeitung der von DARIAH-DE entwickelten Infrastruktur äußerst wichtig. Daher ist in der zweiten Förderphase die Arbeitsgruppe „Research Data Lifecycle" (RDLC) mit der Aufgabe betraut, ein Referenzmodell für einen digitalen Forschungsdatenzyklus in den Geisteswissenschaften zu erstellen. Um zu einer möglichst generischen Lösung zu gelangen, setzt sich die AG aus Mitarbeitern aller Projektbereiche von DARIAH-DE zusammen.

Zu Beginn des Entwicklungsprozesses wurden die jeweiligen Auffassungen eines Forschungsdatenzyklus aus den Geisteswissenschaften gegenseitig vorgestellt. Nach ausführlicher Diskussion von Konzepten und Begriffen, wie „geisteswissenschaftliche Forschungsdaten", Primär- und Sekundärdaten sowie Funktionen und Beziehungen im Datenmanagement, weist das nachfolgende Dokument nun folgende Struktur auf:

- Kapitel 2 diskutiert die Problemstellung eines Forschungsdatenzyklus. Dabei wird insbesondere die Notwendigkeit festgestellt, die in DARIAH zu entwickelnde Infrastruktur auf einem solchen theoretischen Modell aufzubauen, damit dezidiert die Geisteswissenschaften unterstützt werden können.
- Kapitel 3 beschreibt die Ziele des hier anvisierten Referenzmodells für einen Forschungsdatenzyklus der Geisteswissenschaften.
- Kapitel 4 und Kapitel 5 behandeln alle relevanten Begriffe und Funktionen, die im Kontext von geisteswissenschaftlichen Forschungsdaten und der Arbeit mit diesen notwendig sind.
- Kapitel 6 beschreibt einen Entwurf für ein Referenzmodell des Forschungsdatenzyklus für geisteswissenschaftliche Daten, in welchem die vorher diskutierten Funktionsweisen abgebildet werden sollen.
- Kapitel 7 schließt die Thematik ab und verweist auf eine Reihe konkreter Anforderungen an Forschungsdaten und auf Empfehlungen für interessierte Forschende, die Ihre Daten für den dauerhaften Zugang und Nachnutzung in einer Infrastruktur ablegen möchten.

Im vorliegenden Papier wird Geisteswissenschaften als Sammelbegriff und nach seiner institutionellen Bedeutung gebraucht. Dies rührt aus dem Fehlen einer fundierten Auseinandersetzung mit dem im DARIAH-DE Kontext häufig anzutreffenden Begriffspaar „Geistes- und Kulturwissenschaften". Eine eingehende Beschäftigung mit dem eigenen Gegenstand, seiner historischen Entwicklung, seinen Forschungsobjekten (Geist? Kultur?), seinen spezifischen Methoden (Hermeneutik? Dialektik? Analyse?), seinen Akteuren und vor allem auch seine Beziehung zum „Digitalen" wäre zwar notwendig und an der Zeit, kann aber hier nicht geleistet werden.[1]

Zu den Geisteswissenschaften zählen somit all jene Fächer, Disziplinen und Institute, die gemeinhin an den geisteswissenschaftlichen Fakultäten von Universitäten angesiedelt sind. Dabei fallen unter die hier verwendete Sammelbezeichnung Geisteswissenschaften auch die verschiedenen

[1] Für einen pointierten Einstieg zu diesem Thema siehe: Lauer 2013.

philologisch-kulturwissenschaftlichen, philosophischen, theologischen und kulturwissenschaftlichen Fakultäten.[2]

Da im vorliegenden Papier der Zyklus digitaler Forschungsdaten beschrieben werden soll, ist – sofern nicht ohnehin explizit angeführt – immer eine *digital arbeitende* Geisteswissenschaft gemeint.

[2] Für eine an den 2006 formulierten Empfehlungen des Deutschen Wissenschaftsrates orientierte Zusammenschau von insgesamt 96 geisteswissenschaftlicher Fächern siehe: Behrens et al. 2010, S. 182, und zu der erwähnten Definition des Wissenschaftsrates siehe: Wissenschaftsrat 2006, S. 17.

2. Definitionsbedarf

In den digitalen Geisteswissenschaften herrscht – genauso wie in allen anderen Disziplinen – eine beeindruckende Vielfalt von Definitionen für einen Forschungsdatenzyklus. Neben der in DARIAH-DE verwendeten Grafik der Data Documentation Alliance (DDI 2011), den *primitives* von John Unsworth (Unsworth 2000) sowie dem Zyklus von Boonstra et al. (Boonstra et al. 2004), existieren weitere mehr oder weniger praktisch umgesetzte Forschungsdatenzyklen und Methodensammlungen. Schon eine simple Google Bildersuche ergibt eine mannigfaltige Reihe von Schaubildern zu dem Thema:

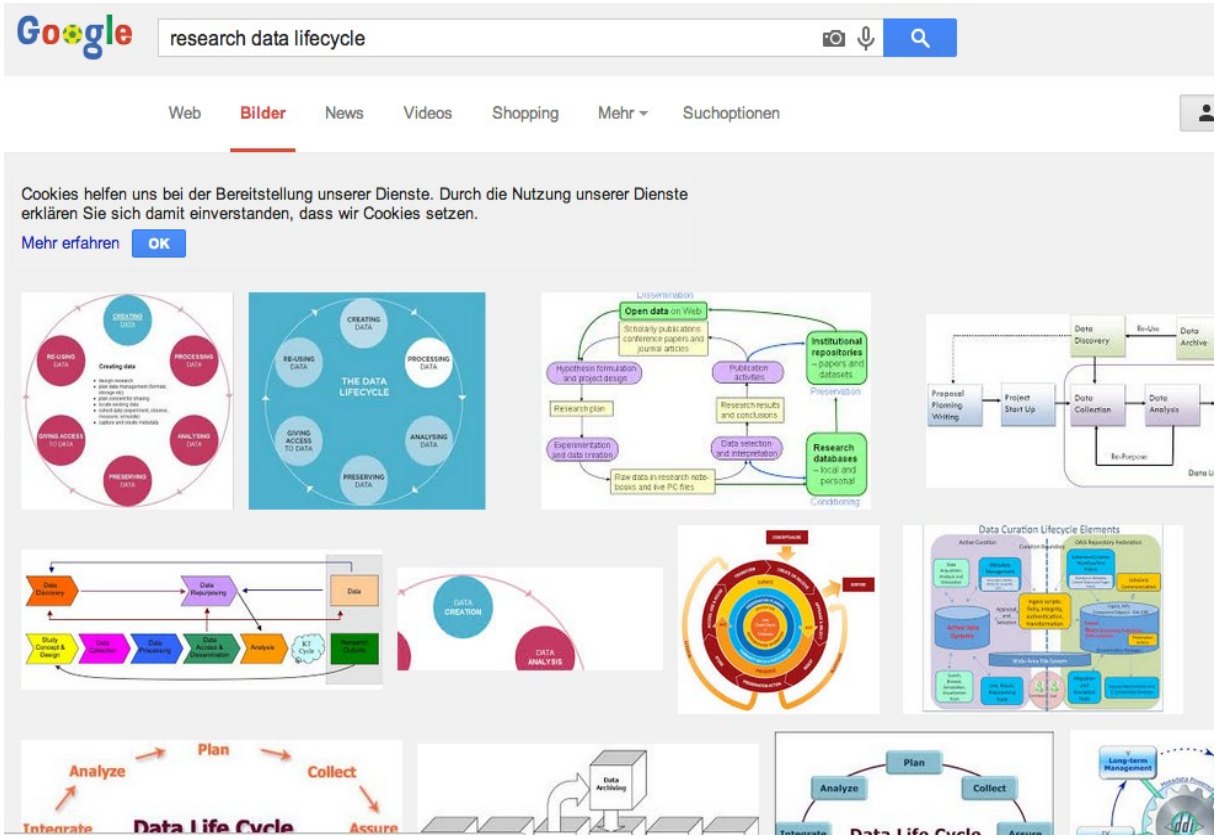

Abb 1: Eine Suche in Google nach "Research DataLifeCycle", Sommer 2014

Die in diesem Suchergebnis erscheinenden Bilder bestehen typischerweise aus einer Infografik, in der einzelne Tätigkeiten oder Aktivitäten eines Forschungsprozess in Form eines Zyklus oder gerichteten Prozess dargestellt werden. Das Kölner Datenzentrum für die Geisteswissenschaften, kurz DCH, zählt beispielsweise eine Reihe von Tätigkeiten auf, die ihren Fokus auf der Präsentations- und Nachnutzungsebene legen: *Problemstellung, Quellenauswahl, Erschließung, Auswertung, Ergebnisse, Aufnahme, Archivierung, Zugang, Präsentation, Nachnutzung*[3].

Das IANUS Projekt des Deutschen Archäologischen Instituts wiederum nennt die Tätigkeiten: *Datenerhebung, Datenspeicherung, Analyse, Auswahl, Langzeitarchivierung, Datenbereitstellung und Nachnutzung* als zentrale Bestandteile seines Forschungsdatenzyklus (IANUS 2012). Die beiden genannten Projekte befinden sich zur Zeit im Aufbau, d.h. in verschiedenen Stadien der

[3] Vgl. http://cceh.uni-koeln.de/projekte/dch/Forschungszyklus_dch_2013.png

Anforderungsanalyse und Implementation, wobei IANUS durch eine große Detailtiefe in der Bedarfserhebung und in den IT-Empfehlungen beeindruckt (IANUS 2014). Die genannten Ansätze zeigen eine starke Konzentration auf Datenerhebung, Analyse und Nachnutzung (IANUS) bzw. Zugang & Präsentation (DCH). Andere Projekte fokussieren sich fast ausschließlich auf den Aspekt der Kuration und Archivierung[4].

Anhand der aufgeführten Beispiele wird bereits deutlich, dass die Betonung einzelner Funktionen bei jedem Research Data-Lifecycle stark variiert und auch entsprechend dem institutionellen und projektspezifischen Bedarf angepasst werden sollte.

Das in dem vorliegenden Dokument erarbeitete Referenzmodell ist der Versuch, gemeinsame Termini und Funktionen zu formulieren, welche möglichst präzise definiert und gleichzeitig möglichst generisch auf verschiedene (digitale) geisteswissenschaftliche Disziplinen übertragbar sein sollen. Um dies zu erreichen, wurden die in der AG vertretenen Mitglieder aufgefordert, ihr spezifisches Verständnis eines Forschungsdatenzyklus und seinen Bestandteilen oder Kernbegriffen möglichst genau zu formulieren. Aus den multiplen Formulierungen wurden dann Gemeinsamkeiten und Unterschiede herausgearbeitet, um zu einem interdisziplinär in den digitalen Geisteswissenschaften anwendbaren Modell zu gelangen. So konnten bereits in der internen Arbeit der AG Spezifika aus den Geschichts-, Literatur- und Informationswissenschaften berücksichtigt werden. Um das Spektrum noch zu erweitern, wurde ein erster Entwurf des Modells an weitere FachwissenschaftlerInnen innerhalb von DARIAH-DE zur kritischen Überprüfung versandt. Auf diese Weise wurde der Fokus um die Perspektive der Musikwissenschaften ergänzt.

Die folgenden Kapitel beleuchten detailliert die einzelne Aspekte, die hier berücksichtigt werden müssen und geben – soweit möglich – Umsetzungsempfehlungen.

[4] Vgl. http://4.bp.blogspot.com/-zTZlL8E-W9E/Un39VIp-9HI/AAAAAAAAAAA/VjFYaZ0QfD8/s1600/IU_ASIST_Panel.png

3. Ziele

Gemäß den Anforderungen an ein Referenzmodell zur Beschreibung von Software einerseits und ganz allgemein branchenspezifischen Prozessen andererseits, sollen in den folgenden Kapiteln **Begriffe, Funktionen und Abläufe** für einen **Forschungsdatenzyklus in den digitalen Geisteswissenschaften** beschrieben werden.

Das resultierende Referenzmodell dient sowohl als Empfehlung zur weiteren Implementation innerhalb von DARIAH-DE, als auch generell als Empfehlungsgrundlage für die Arbeit mit *Forschungsdaten* in anderen digital arbeitenden geisteswissenschaftlichen Vorhaben. Zum besseren Verständnis werden im ersten Teil die verwendeten **Konzepte** und **Begrifflichkeiten** und dann in den weiteren Kapiteln **Funktionen** und **Abläufe** definiert. Den oben diskutierten Anforderungen für Forschungsdatenzyklen in den Geisteswissenschaften entsprechend, soll folgende Definition gelten:

Ein **Referenzmodell** für Forschungsdatenzyklen in den Geisteswissenschaften soll ein integriertes System aus Forschungsdaten, deren Aufbereitung, sowie den zu ihrer Verarbeitung und Verwaltung benötigten Metadaten und Tools beschreiben. Dabei sollen in einem solchen System ein oder mehrere vollständige(r) Forschungsprozess(e) innerhalb der Geisteswissenschaften realisierbar sein. Ein solcher implementierter Forschungsdatenzyklus in einem System fächert sich in folgende Anforderungen auf:

- Forschungsdaten sollen in diesem System **verwaltet** werden
- Forschungsdaten sollen **weiter verarbeitet** werden können
- Forschungsergebnisse sollen **gespeichert** werden
- Forschungsergebnisse sollen **rekonstruierbar** sein
- Forschungsdaten sollen **wiederverwendbar** sein
- Forschungsdaten und (Zwischen-) Ergebnisse sollen **archiviert** werden
- und zur weiteren Benutzung **publiziert** werden

Ein solches System beinhaltet eine starke Komponente der Sicherung,Verwaltung und Publikation, sowie der Wiederauffindbarkeit und Rekonstruierbarkeit von Daten bzw. der daraus gewonnenen Ergebnisse. Sowohl die Verwaltung als auch die Rekonstruierbarkeit der Datenobjekte wiederum hängen stark von Art und Qualität der dahinter liegenden Metadaten sowie den sie generierenden automatischen Prozessen ab. D.h. sowohl **Datenformate**, als auch die auf ihnen arbeitenden **Tools** und alle verwendeten Metadatenstandards sollten einem hohen **Grad an Standardisierung und Normalisierung** unterliegen.

4. Begriffe

Die folgenden Kapitel diskutieren und definieren alle Konzepte und Begrifflichkeiten, die in einem Forschungsdatenzyklus der Digitalen Geisteswissenschaften eine Schlüsselrolle spielen. Insbesondere sind dies Konzepte, die direkt mit der Arbeitsweise in den digitalen Geisteswissenschaften zu tun haben und daher mit ihnen abzustimmen sind.

4.1 Forschungsdaten und -sammlungen

Forschungsdaten der digital arbeitenden und forschenden Geistes- und Kulturwissenschaften lassen sich eher in einem Annäherungsprozess umreißen als eindeutig definieren. Hier dienen insbesondere empirische Kriterien als Definitionskriterien für Forschungsdaten. Die folgenden Kapitel zählen einige wichtige Differenzierungen auf, die bei der Definition von und Arbeit mit Forschungsdaten eine Schlüsselrolle spielen.

4.1.1 Zur Unterscheidung von Primärdaten und Sekundärdaten

Der Unterschied von Primärdaten und Sekundärdaten ist im Falle von zyklischen, digitalen Forschungsprozessen (=des einen Primärdaten können des anderen Sekundärdaten sein) ein gradueller und kann nicht abschließend geklärt werden.

In der Theorie gelten folgende Arbeitshypothesen:

- Primärdaten sind Daten, die die Grundlage zur Beantwortung einer Forschungsfrage darstellen. Das Spektrum reicht von „rohen", also unbearbeiteten Daten bis zu aufbereiteten Daten.
- Primärdaten werden in der Regel im weiteren Forschungsprozess verarbeitet, d.h. zur Beantwortung dieser spezifischen Forschungsfrage angereichert.
- Sekundärdaten bezeichnen die Daten, die schon anderweitig als Quelldaten prozessiert worden sind und in erster Linie dem „Abgleich mit der Außenwelt" dienen.

Die Diskussion in den digitalen Geisteswissenschaften um Primär- und Sekundärdaten erweckt dabei den Eindruck, dass man versucht die Naturwissenschaften nachzubilden und deren scharfe Trennung zwischen beiden Datenarten zu übernehmen (vgl. Gradmann, Meister 2008). Aufgrund der anderen Arbeitsweise und vor allem aufgrund der viel heterogeneren Definition von Forschungsgegenständen, kann diese Trennung jedoch für die (digitalen) Geisteswissenschaften nicht vollzogen werden:

Je nach Forschungsfrage sind die „so genannten" Sekundärdaten des einen die Primärdaten des anderen. Es lässt sich weder eine scharfe Definition nach einer spezifischen Datenart noch nach Anwendungsfall oder Forschungsfrage aufstellen. Auch die Nestor Arbeitsgruppe zur digitalen Langzeitarchivierung zieht daher folgende Konsequenz:

> „Der begrifflichen Klarheit wegen sollte daher das Präfix „Primär-" nicht mehr verwendet werden und statt dessen nur noch von wissenschaftlichen Daten oder Forschungsdaten gesprochen werden." (Neuroth 2010, Kap.17:105)

Der Vollständigkeit und korrekten Zitierweise halber kann in den kommenden Kapiteln nicht vollständig auf den Begriff „Primärdaten" verzichtet werden, wo immer möglich, ist aber von „Forschungsdaten" als weit generischerem und zu weniger Missverständnissen führendem Oberbegriff die Rede. Auf den Begriff „Sekundärdaten" kann gänzlich verzichtet werden.

4.1.2 Der Unterschied zwischen Daten und Objekten

Grundsätzlich wird in der Literatur zu Forschungsdaten und Repositorien für diese bzw. für die digitale Langzeitarchivierung, der Begriff des digitalen *Objekts* verwendet. Dabei ist zu bemerken, dass dieser abstrakte Objektbegriff häufig an einem **Mangel an konkreten Definitionen** krankt. Beispielsweise definiert das OAIS Modell ein digitales Objekt als „An object composed of a set of bit sequences" (Nach Neuroth 2010, Kap 9.1, S.3). Die Deutsche Digitale Bibliothek (DDB) definiert ein digitales Objekt als ein Objekt,

> „... das in binärer Kodierung auf einem Datenträger vorliegt [...]. Ein digitales Objekt kann entweder ein Digitales Primärobjekt oder ein Digitalisat sein. In der DDB werden Digitale Objekte durch Medientypen klassifiziert." (DDB 2012, S.1)

Das Kompetenznetzwerk zur digitalen Langzeitarchivierung in Deutschland, *Nestor*, unterscheidet selbst zwischen *physischen*, *logischen* und *konzeptuellen* Objekten (Neuroth 2010, Kap 9.1, S. 4). Diese beschreiben verschiedene Perspektiven auf den Objektbegriff:

Physische Objekte können demnach als Folgen unterschiedlicher *Ladung* oder *pits* (je nach Art des Speichermediums) wahrgenommen werden, *logische* Objekte werden als Dateien (mit einem bestimmten Formattyp) oder auch ausführbare Programme (Z.B. *.app oder *.exe i.d.R. für ein bestimmtes Betriebssystem etc) definiert. *Konzeptionelle* Objekte beschreiben hingegen eine am ehesten inhaltliche Sicht auf Objekte: Hier ist nicht der Dateizusammenhang ausschlaggebend, ebensowenig das Format oder der Speicherort eines Objekts, die konzeptionelle Perspektive definiert den inhaltlichen Sinnzusammenhang eines Objekts gemäß seinem ursprünglichen Zweck und Kontext.

Sowohl in der Langzeitarchivierung als auch in einem System für einen Forschungsdatenzyklus muss das Paradigma gelten, mindestens das *logische* Objekt, also eine Datei mit all ihren Spezifika, zu verarbeiten und zu archivieren. Allerdings muss ein System, um ein *logisches* Objekt verarbeiten zu können, eine sehr genaue „Kenntnis" der dahinter liegenden Dateilogik oder auch der dahinter liegenden *Konzepte* enthalten. Je exakter diese Kenntnis ist – zusammen mit der intendierten Verwendungsweise einer Datei – desto eher nähert sich dieses Konzept dem Paradigma des *konzeptionellen Objekts*.

Es muss also sehr genau spezifiziert werden,

- wie physische Objekte gespeichert und verwaltet werden, damit sie als solche verarbeitet werden
- welche Dateiformate mit welcher Software verarbeitet werden können oder sollten (KEIN System kann alle Dateiformate und Verarbeitungsmöglichkeiten unterstützen)
- welche Verknüpfungen zwischen Dateien möglich und erlaubt sind, damit aus diesen Dateien *konzeptionelle* Objekte entstehen. Es müssen also alle erwartbaren Beziehungen

und Beziehungstypen zwischen Einzeldateien beschrieben sein, damit das System diese adäquat handhaben kann.

Diese Liste lässt sich mit zunehmender Komplexität einer Infrastruktur weiter führen. Die drei aufgeführten Punkte gelten dabei als absolutes Mindestmaß.

4.1.3 Der Unterschied zwischen Metadaten und Daten

Auf den ersten Blick erscheint der Unterschied zwischen Dateien (Daten) und Metadaten recht eindeutig:

> „Metadata is often called data about data or information about information." (NISO 2004).[5]

Weniger eindeutig ist jedoch, wie die Verknüpfung von Daten und Metadaten konkret erfolgt:

> „Metadata can be embedded within the information resource (as is often the case with web resources) or it can be held separately in a database." (Haynes 2004, S.8)

Aber auch die in den Metadaten gespeicherten Informationen selbst können – je nach Bedürfnis und Kontext – unterschiedliche Interessen bedienen und Auskünfte über den Inhalt der Daten geben, technische und administrative Spezifikationen bereitstellen und etwa Angaben zur Größe, zum Dateiformat oder zu Lizenzen beinhalten. Beispiele aus der Wirklichkeit illustrieren die Komplexität von Metadaten: So sind etwa einerseits einige Metadatenformate speziell dazu erschaffen worden, Beziehungen zwischen mehreren Dateien und Objekten zu beschreiben (bspw. gemäß oben beschriebenem *logischen* Objektparadigma), wie PREMIS oder METS aus dem bibliothekarischen Bereich.

Daneben existieren Metadatenformate, die selbst wieder als Forschungsobjekte gelten können. Gerade im Bereich der Editionswissenschaften und dem hier häufig verwendeten Standard TEI (TEI 2013) ist eine solche Perspektive auf Daten keine ungewöhnliche, da hier die Differenzierung zwischen Daten und Metadaten eine feinere ist.

So entstehen durch **TEI im (üblichen) Falle von Inline-Mark-Up** sowohl Daten (Inhaltsdaten = der ursprüngliche – häufig literarische – Text) als auch Metadaten (das Mark-Up der WissenschaftlerInnen) in der gleichen Datei. Selbst Office-Dateien bieten die Möglichkeit, in Ihnen sowohl administrative Metadaten abzuspeichern oder die dort hinterlegte Struktur von Kommentaren bis zur Überschriftenhierarchie als eine Art strukturelle Metadaten zu begreifen. Um die Verwirrung perfekt zu machen, wird die für Metadaten zumeist eingesetzte Auszeichnungssprache XML zudem gelegentlich in inhaltlichen Formatstandards, wie MPEG oder Open Office eingesetzt. Hier wird der eigentliche „Content" eines Dokuments, also der Binärstrom eines Bildes oder die ACSII-Zeichen eines Texts, welche später von einem Anwendungsprogramm dargestellt werden, in XML eingebettet und dann zusammen mit weiteren XML-Dokumenten in einem (z.T. komprimierten) Containerformat zusammengefasst. Auch dieses Beispiel zeigt die Schwierigkeiten, die bei dem Versuch, zwischen Daten und Metadaten klar zu trennen, entstehen können. Daher ist es sinnvoll eine klare Linie **vorzugeben**, wo die Grenze zwischen Dateien als Daten und Dateien als Metadaten verläuft:

Da Metadaten im Falle einer Forschungsinfrastruktur in erster Linie dafür eingesetzt werden, **administrativ** und **strukturell** digitale Objekte zu verwalten, fällt eine TEI-Datei oder eine MPEG-

[5] Für eine Übersicht verschiedener Definitionsentwürfe von „Metadaten" siehe: Miller 2011.

Datei hiermit eindeutig in das Gebiet der inhaltlichen Daten (Content) – ungeachtet der XML-Notation des Inhalts. Weiterhin sei an dieser Stelle angefügt, dass auch **bibliographische** Metadaten zum Teil als inhaltliche Metadaten bezeichnet werden, dann aber in das Feld der **deskriptiven** Metadaten gehören (ReMind 2005) und in diesem Fall auch für eine Forschungsinfrastruktur interessant sein können, da sie die Suchfunktionalität deutlich bereichern können. Eine dazugehörige administrative Metadatei würde in diesem Fall die Beziehung dieser inhaltlichen Daten beschreiben und ggf. nähere Informationen zu Inhalt und Struktur liefern, die ein System dann später auslesen kann.

Schwierigkeiten bereitet die Unterscheidung von Daten und Metadaten somit vor allem dann, wenn diese gemeinsam in einer Datei vorkommen, so dass ein System nicht maschinell erkennen kann, ob es sich hier um eine Inhaltsdatei (mit Informationen eines bestimmten Medientyps zur späteren Verarbeitung) oder um eine Datei mit Metadaten zur Verwaltung und Bereitstellung handelt[6], was gerade bei der Vergabe von Identifiern problematisch sein kann.

Es wird daher empfohlen, inhaltliche und alle anderen Metadaten nicht in der selben Datei vorzuhalten, sondern getrennt zu speichern. Eine Trennung zwischen Primärdatei (gescannter Text) und inhaltlichen Metadaten (Annotationen) in verschiedene Dateien wird ebenfalls empfohlen. Wenn letzteres nicht eingehalten werden kann, sollte die Datei, die sowohl digitalisierte Informationen als auch inhaltliche Anmerkungen dazu enthält, als Inhaltsdatei bereitgestellt werden.

4.1.4 Definition Forschungsdaten

Als Ausgangspunkt für eine Definition von digitalen geistes- und kulturwissenschaftlichen Forschungsdaten soll folgender Satz fungieren. Demnach wären Forschungsdaten:

> *Alle Daten die während der Forschung benötigt und erzeugt werden*

Präzisiert man diesen Satz nun um die Eigenschaften digital und geistes- und kulturwissenschaftlich, wären digitale, geistes- und kulturwissenschaftliche Forschungsdaten:

> *Alle **digitalen** Daten, die während **geistes- und kulturwissenschaftlicher Forschungen** benötigt und erzeugt werden.*

Möchte man den Zusammenhang zwischen Forschungsdaten und Forschungsfrage stärker betonen, so wären digitale, geistes- und kulturwissenschaftliche Forschungsdaten

> *Alle digitalen Daten, die **im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage** benötigt und erzeugt werden.*

Anstelle des wenig spezifischen „benötigt" empfiehlt es sich, hier über entsprechende Tätigkeitsbezeichnungen den Bezug zum Research Data Lifecycle zu stärken. Demnach wären digitale geistes- und kulturwissenschaftliche Forschungsdaten

> *Alle digitalen Daten, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage **gesammelt, erzeugt, beschrieben und/oder ausgewertet werden**.*[7]

---

[6] Zur Kritik speziell an TEI, siehe Schmidt 2012

[7] Bei Sahle, Kronenwett 2013, S. 79 werden „Primärdaten" beschrieben als: „Überreste und Artefakte der menschlichen Kultur [...] die entweder Gegenstand der geisteswissenschaftlichen Forschung sind oder

Formt man aus den bisherigen Definitionsfragmenten einen Satz, so ergibt sich:

> *Unter digitalen geistes- und kulturwissenschaftlichen Forschungs*daten *werden innerhalb von DARIAH-DE all jene Daten verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, beschrieben, ausgewertet und/oder erzeugt wurden.*

Wie zu sehen ist, tritt bei dieser Definition der zu definierende Begriff („Forschungsdaten") in leicht veränderter Form auch in der Definition („Daten") auf, womit es sich also um eine graduelle Zirkeldefinition handelt.

Sinnvoll dürfte hier eine Abgrenzung von „Daten" bzw. „Forschungsdaten" gegenüber „Quellen" bzw. „Materialien" und „Publikationen" sein. Eine funktionale Unterscheidung von Quellen/Materialien und Forschungsdaten macht im Kontext eines Resarch Data Lifecycles kaum Sinn, da Forschungsdaten nicht nur Ergebnis sondern auch Ausgangspunkt wissenschaftlicher Forschung sein sollen.

Von dieser Überlegung ausgehend wird deutlich, dass eine eindeutige Unterscheidung von Quelle/Material und Forschungsdatum nicht möglich ist, da Forschungsdaten immer Quellen/Material sein können. Mindestens genauso deutlich und unumstritten ist aber auch die Überlegung, dass es verschiedene Arten von Quellen und Forschungsmaterialien gibt, wobei sich eine Art im besten Fall durch eine Reihe artspezifischer Merkmale von einer anderen Art unterscheiden lässt.Lässt man diese Überlegungen nun in die bereits vorliegende Definition einfließen, so erhält man folgendes Ergebnis:

> *Unter digitalen geistes- und kulturwissenschaftlichen Forschungsdaten werden innerhalb von DARIAH-DE all jene* ***Quellen/und Materialien*** *verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und* ***in digitaler Form vorliegen****.*

Es wurde bereits eindrücklich darauf hingewiesen, dass Forschungsdaten nicht nur Quellen und Forschungsmaterial sind, sondern auch das Ergebnis von Forschungen wie beispielsweise Editionsprojekten darstellen.

Dieser wichtige Punkt könnte in der bisher entwickelten Definition stärker betont werden, indem nicht nur „Quellen/Materialien", sondern „Quellen/Materialien und Ergebnisse" genannt werden, auch wenn es sich bei diesen „Ergebnissen" strenggenommen auch wieder um Materialien handelt oder wenigstens handeln könnte. Eine vorläufige Definition von Forschungsdaten könnte in etwa wie folgt lauten:

> *Unter digitalen geistes- und kulturwissenschaftlichen Forschungs*daten *werden innerhalb von DARIAH-DE all jene* ***Quellen/Materialien und Ergebnisse*** *verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und in digitaler Form vorliegen.*

Überlegenswert wäre darüber hinaus die Frage, ob und wie der Aspekt der Datenspeicherung und -nutzung durch andere Forschende in einer Definition von Forschungsdaten thematisiert werden sollte:

Indizien zur Beantwortung von Forschungsfragen liefern können''.

> *Unter digitalen geistes- und kulturwissenschaftlichen Forschungs*daten *werden innerhalb von DARIAH-DE all jene Quellen/Materialien und Ergebnisse verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und in digitaler Form* ***zum Zwecke der Archivierung, Zitierbarkeit und zur weiteren Verarbeitung aufbewahrt*** *werden.*

Digital und / oder Maschinenlesbar

Die explizite Deklaration von Forschungsdaten als „Ergebnisse" umfasst auch die traditionellen Ergebnisse geistes- und kulturwissenschaftlichen Arbeitens, nämlich Veröffentlichungen wie Qualifizierungsarbeiten, Aufsätze und Monographien – sofern diese „digital" vorliegen. Die explizite Deklaration von Forschungsdaten als „Quellen/Materialien" umfasst außerdem prinzipiell alle digitalen Objekte, unabhängig von ihrem jeweiligen Dateiformat[8].

Der vorliegende Definitionsentwurf entspricht somit dem Wunsch nach einer möglichst breiten und umfassenden Definition von Forschungsdaten bzw. von digitalen geistes- und kulturwissenschaftlichen Forschungsdaten. Gleichzeitig können von diesem Definitionsentwurf allerdings kaum Richtlinien oder Vorgaben für die praktische Umsetzung eines Research Data Lifecycles unter Einbeziehung konkreter Storage-Infrastrukturen wie etwa dem DARIAH-DE-Repositoriums oder der Collection-Registry abgeleitet werden.

Dieses Defizit des vorliegenden Definitionsentwurfes könnte relativ einfach gelöst werden, indem man die Definition zumindest um den Begriff „maschinenlesbar" erweitert, wobei die konkrete Definition von „maschinenlesbar" schlussendlich von den technischen Spezifikationen der verwendeten Infrastruktur abhängen muss. Ein abschließende Definition von Forschungsdaten im Kontext des hier zu beschreibenden Research Data Lifecycles lautet also:

> *Unter digitalen geistes- und kulturwissenschaftlichen Forschungsdaten werden innerhalb von DARIAH-DE all jene Quellen/Materialien und Ergebnisse verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und/oder ausgewertet werden und in maschinenlesbarer Form zum Zwecke der Archivierung, Zitierbarkeit und zur weiteren Verarbeitung aufbewahrt werden können***.**

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen:

Technisch handelt es sich bei digitalen Forschungsdaten in den (digitalen) Geistes- und Kulturwissenschaften um Binärströme, wie in allen anderen (digitalen) Wissenschaften. Dabei ist gerade in einer Infrastruktur für die digitalen Geisteswissenschaften aufgrund der Vielfalt der Quellenarten und Objekttypen und aufgrund der Vielfalt der unterschiedlichen Fragestellungen auf ein konsistentes Datenmodell zu achten, damit keine Dateien und Verknüpfungen verloren gehen und auch der ursprüngliche Kontext einer Datensammlung erhalten bleibt.

[8] Eine Zusammenstellung verschiedener geisteswissenschaftlicher Quellentypen findet sich unter: https://dev2.dariah.eu/wiki/display/publicde/7.3+Zusammenstellung+geisteswissenschaftlicher+Quellentypen

4.2 Dateiformate und Objektklassen

Trotz der im vorigen Abschnitt erwähnten „Vielfalt der Quellenarten und Objekttypen" sowie der „Vielfalt der unterschiedlichen Fragestellungen", ist die Anzahl der in den Geisteswissenschaften mehrheitlich verwendeten Dateiformate (und Programme) überschaubar. Dies zeigen Untersuchungen die sowohl von DARIAH-DE selbst,[9] wie auch jenseits des DARIAH-Kontextes durchgeführt wurden.[10] Die konkreten Ergebnisse dieser Studien, welche an dieser Stelle aber nicht näher referiert werden sollen, bringt folgendes Zitat auf den Punkt: „What is a bit surprising perhaps is that still 6% of the respondents *do not use* office documents." (Kuipers, van der Hoeven 2009)

Eine nach weitgehend typischen geisteswissenschaftlichen Arbeitsschritten gegliederte Übersicht der beobachteten gängigsten Software und Dateiformate enthält nachfolgende Tabelle. Software und Dateiformate sind jeweils nach ihrer Popularität angeordnet.

Tabelle 1: Beobachtete Dateiformate in den klassischen Geisteswissenschaften

Arbeitsschritt	verwendete Software	verwendete Dateiformate
Bibliographieren	MS-Word, Zotero, Citavi, Endnote, MS-Excel, Libre/OpenOffice, Litlink	doc(x), odt, rtf, txt (weitere Outputformate von Zotero, Citavi etc.), xls(x)
Notieren/Exzerpieren	MS-Word, Citavi, OneNote, Libre/OpenOffice, Evernote	doc(x), odt, rtf, txt, bib, ris, xls(x)
Zählen, Ordnen, Analysieren	MS-Excel	xls(x), csv
Transkribieren	MS-Word, Citavi, Libre/OpenOffice	doc(x), odt, txt
Schreiben	MS-Word, Libre/OpenOffice, TeXnicCenter	doc(x), odt, tex
Präsentieren	MS-PowerPoint	ppt(x)
Ansehen, Speichern und Bearbeiten von Bildern	IrfanView, MS-Picture Manager, Photoshop, Picasa, gimp	jpeg, tiff, pdf, png, gif, svg
Aufnehmen, Abspielen und Bearbeiten von Tonaufnahmen	Audacity, Cool Edit Pro, Finale, Mediaplayer	wav, mp3, midi, flac, ogg

Dass die Digital Humanities dieses doch sehr enge Spektrum an Dateiformaten vergrößern, darf angesichts der fast schon programmatischen und das Fach definierenden systematischen und souveränen Verwendung des Computers erwartet werden. Anhand zweier prototypischer

[9] Vgl. dazu Stiller et al. 2015, sowie Andorfer 2015.

[10] Siehe ausführlicher dazu: Andorfer 2015.

Beispielprojekte soll die Fülle von verschiedenen eingesetzten Objekttypen aufgezeigt werden, die bei weitem noch nicht das ganze Spektrum der benutzten Objekttypen in den digitalen Geisteswissenschaften abdeckt.

Im Falle eines Editionsprojekts von unsortierten Manuskripten eines frühneuzeitlichen Mathematikers ist die Grundlage eine Menge von digitalen Faksimiles und Transkriptionen in MS Word. Die Transkriptionen werden in XML-Dateien übertragen und gemäß einem XML-Schema ausgezeichnet. Nebenbei werden auch Biographien und Bibliographien (mit Referenzen zu Authority files im Internet) angelegt, zu denen aus den XML-Dokumenten verwiesen wird. Projektinterne Konventionen werden ebenfalls als Textdokumente abgelegt. Weitere Dokumentation über den Fortschritt der Arbeit erfolgt durch ein Versionskontrollsystem (Subversion).

Zusätzlich dazu werden in einer Diagramm-Software Diagramme gezeichnet, die thematisch und hierarchisch sortiert werden und zu der Online-Repräsentation der Faksimiles, zu Thesauri und anderen Online-Repositorien verlinkt werden. Da das Dateiformat der Diagramme wiederum eine Graphenstruktur (graphml) ist, können diese Dateien programmatisch zu einer RDF-Repräsentation weiterverarbeitet werden. Skripte, die entweder ein fester Bestandteil des Workflows sind oder für kleinere, einmalige Aufgaben benutzt werden, sind in XSL und Python geschrieben.

Das zweite Beispiel ist ein Projekt, das an der Modellierung wissenschaftlichen Arbeitens arbeitet. Ausgehend von Literaturrecherche (deren Ergebnisse in Programmen zur Literaturverwaltung festgehalten werden), Analyse bestehender Ontologien (die entweder in Form von Artikeln als PDF oder HTML oder direkt als Ontologien in RDF oder OWL vorliegen) und Interviews wird versucht, ein eigenes Modell zu entwickeln und dies auch in einer RDF-Repräsentation darzustellen. Weitere hier anfallende Objektklassen sind Audioaufnahmen von Interviews und die Transkription dieser. In einem zweiten Schritt wird eine maschinenlesbare Version dieses Modells mit webbasierten Annotationstools verbunden, so dass Webseiten mit dem Vokabular des Modells ausgezeichnet werden können, und beispielsweise Beziehungen zwischen Texten oder der Verlauf einer wissenschaftlichen Argumentation expliziert werden können.

Tabellarisch zusammengefasst ergibt sich demnach der folgende Überblick, in den die beobachteten Dateiformate aus den klassischen Geisteswissenschaften (Tabelle 1) zu integrieren sind:

Tabelle 2: Beobachtete Dateiformate in den Digital Humanities

Arbeitsschritt	verwendete Software	verwendete Dateiformate
Skripten/Programmieren	z.B. Notepad++, Eclipse, Oxygen, Emacs, vim	pl, py, java, xsl
Visualisieren	z.B. Gephi, GeoBrowser, yEd, Internetbrowser	gexf, kml, geojson, graphml, html, css
Quantitative Datenerfassung und Analyse	z.B.:Datenbanken (MySQL, SQLite, MongoDb, MS-Access), R, Tabellenkalkulation	sqlite, r, xlsx, ods
Arbeit mit 3D-Objekten	z.B. 3DHop, AutoCAD, Blender	ply, nxs, dxf, blend
Erstellen von Ontologien, Thesauri	z.B. Protegé, Tematres	owl, rdf
Datenmodelle	z.B. Schemadefinitionen für XML, Schemaeditoren wie Roma, Formatvorlagen	rng, rnc, xsd, dtd, dotx, ott
Annotieren	z.B. Bookmarklet	json

Es zeigt sich auch, dass in den Digital Humanities neben dieser breiten Palette von Dateiformaten verstärkt fachspezifische Metadatenstandards zum Einsatz kommen, um Interoperabilität der Daten zu gewährleisten. Viele Dateiformate sehen vor, dass die zugehörigen Metadaten direkt in der Datei gemeinsam mit dem eigentlichen Inhalt gespeichert werden. Eine andere Möglichkeit ist es, derartige Daten extern zu verwalten, beispielsweise in Datenbanken oder in sogenannten Sidecar-Dateien, wie beispielsweise XMP-Dateien, die die Metadaten zu einer zugehörigen Bilddatei vorhalten. Tabelle 3 zeigt beobachtete Metadatenstandards.

Tabelle 3: Beobachtete Metadatenstandards

Metadatenstandard	verwendet für/bei/in
TEI	Text (digitale Editionen, Korpuserstellung, …)
MEI	Noten (digitale Edition musikalischer Schriften, Partituren, …)
EXIF, XMP	Bilder

METS/MODS	DFG-Viewer, Bibliotheken
EAD	Archivwesen

Zuletzt sei vorsorglich darauf hingewiesen, dass bei der Archivierung von publizierten Daten bedacht werden muss, dass es sich unter Umständen um komplexe Systeme mit vielen Abhängigkeiten untereinander handeln kann. In solchen Fällen sind Lösungen wie virtuelle Maschinen und Webseitenarchive in Erwägung zu ziehen. Außerdem sollten Algorithmen oder softwarebasierte Methoden, mit denen Forschungsdaten manipuliert oder erzeugt wurden auch gespeichert werden.

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen: Die oben genannten Objektklassen können in verschiedenen Dateiformaten vorliegen. Dabei sind aus Gründen der besseren Langzeitarchivierbarkeit gut dokumentierte, nicht (oder nur verlustfrei) komprimierte Dateiformate in weit verbreiteten Standards[11] zu bevorzugen. Im Falle der Wahl zwischen proprietären und nicht-proprietären Dateiformaten (Hier sind häufig komplexere Inhalte wie audiovisuelle oder Datenbank- / 3D Formate betroffen) sollte die Entscheidung zugunsten nicht-proprietärer Standards (Lormant et al. 2005) getroffen werden.

Die weit verbreitete Verwendung von MS-Office Software respektive ihrer weniger restriktiven Äquivalente Libre- oder OpenOffice unter den eher traditionell arbeitenden GeistswissenschaftlerInnen erweist sich hier als vorteilhaft, da diese Programme neben ihren Standardformaten wie z.B. docx auch offene Formate wie etwa ODT generieren können. Selbiges trifft auch auf die beiden am häufigsten anzutreffenen Literaturverwaltungsprogramme Citavi und Zotero zu, welche ebenfalls eine Reihe empfehlenswerter Dateiformate exportieren können. Ebenfalls keine Probleme bereitet auch das Speichern von Bildern in unterschiedlichen Dateiformaten. Sofern also die Möglichkeit besteht, das Dateiformat wählen zu können, dann sollten die von DARIAH-DE empfohlenen Dateiformate[12] verwendet werden.

4.3 Annotationen

Annotationen sind Daten, die im Forschungsprozess generiert oder während dessen mit dem **Forschungsgegenstand verknüpft** werden, dabei können sie technisch die Form von Metadaten annehmen, dies ist aber keinesfalls zwingend notwendig oder immer möglich. Sie können die Form von Klassifizierungen, Anmerkungen, Kommentaren (u.ä.) haben.

> „Annotation is a practice with a rich and varied history in the humanities, one intertwined with, and therefore as complex as, the history of reading itself. Practiced extensively by artists and thinkers [...] annotation was also a formal pedagogy in the universities of Renaissance France and England. Indeed, the impulse to make marks on a page is so widespread and deeply rooted that it seems to reflect not a cultural formation, as is writing itself, but rather a natural one: the kinesthetic dimension of learning. In this respect, annotation is intuitive – a practice that readers would follow even without

[11] Eine vollständige Liste von Kriterien findet sich auf: IANUS 2014.

[12] Vgl. https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=38080370

> prior historical direction, and therefore wanders across the timeline of the humanities with irregular bursts and continuities". (Paradis et al. 2013)

Eine andere Definition von Annotationen stellt Folgendes fest:

> „Annotations are used in scholarly communication to organize existing knowledge and to facilitate the creation and sharing of new insights. They can become so important as to have scholarly value in their own right, and hence their transition from paper to digital scholarship has been, and remains today, crucially important". (Sanderson, Van de Sompel 2010)

In digitalen Forschungsumgebungen wurde das Annotieren zunehmend im Kontext von anderen Forschungspraktiken wie dem gemeinsamen Zugriff auf Inhalte, **kollaborativem Arbeiten oder der semantischen Arbeit mit Quellen** eingesetzt. In einigen Disziplinen werden Annotationen hier und heute mit Hilfe von kontrollierten Vokabularen erstellt und dann auch zunehmend maschinell interpretiert und verglichen[13].

Exemplarisch ist hier das CIDOC-CRM zu nennen, welches schon in den 90er Jahren entworfen wurde, um eine komplexe regelbasierte Wissensbasis zur semantischen Annotation und Dokumentation musealer aber auch archivarischer und bibliothekarischer Gegenstände zu erhalten. Durch die kontinuierliche Pflege und Weiterentwicklung des CIDOC-CRM gilt dies heute als eine der komplexeren und wirklich angewandten maschinenlesbaren Ontologien in den Geisteswissenschaften (Lampe et al. 2010), welche aber ursprünglich auf Papier entwickelt worden ist und bis heute zur Beschreibung von Objekten des kulturellen Erbes weit verbreitet eingesetzt wird.[14]

Annotationen brauchen nicht als Metadaten kenntlich gemacht sein, um für WissenschaftlerInnen von Relevanz und Nutzen zu sein. Gerade analoge „Annotationen" in Form von handschriftlichen Anmerkungen am Seitenrand historischer Quellen, können vermutlich nur mit hohem handwerklichem und intellektuellem Aufwand in technische Metadatenstandards auf XML-Basis überführt werden, so dass auf das historische Original verzichtet werden könnte. Daher ist es naheliegend, hier im Zweifel auf eine klare Trennung zwischen Forschungsdaten und Annotationen zu verzichten und beide dem Bereich der Forschungsdaten zuzuordnen. Daneben bestehen aber auch allgemeine Bestrebungen, digitale Annotationen in Form von Metadaten direkt eingebettet in die zu annotierenden Forschungsdaten (also Bild- & Textdaten) zu integrieren (metadatadeluxe 2015).

Wo genau Annotationen eingebettet werden und wie mit Ihnen umgegangen wird, hängt stark von den jeweiligen Disziplinen ab. Grundsätzlich lässt sich aber zur besseren technischen Handhabung in einem Repository feststellen, dass eine Trennung von Annotationen und anderen (technischen, administrativen, strukturellen) Metadaten mit Sicherheit zu befürworten ist, damit hier nicht unterschiedliche Konzepte vermischt werden und unnötige Schwierigkeiten bei der automatisierten Extraktion von technischen oder administrativen Metadaten auftreten.

---

[13] Die Vorgehensweise ist stark disziplinabhängig. Es kann aber angenommen werden, dass die Literaturwissenschaft kontrollierte Vokabulare eher nicht zum Annotieren verwendet, während dies in den Bibliotheks- und Museumswissenschaften zu einem großen Teil geschieht.

[14] Die aktuelle Version der CIDOC CRM Ontologie unter http://www.cidoc-crm.org/rdfs/cidoc_crm_v5.0.1.rdfs

Es lassen sich außerdem ganz grundsätzliche Ansätze beobachten, die (syntaktische) Definition einer disziplinübergreifenden interoperablen Form von Annotationen zum Ziel haben.

Ein solcher Ansatz ist das ***Open Annotation Data Model*** (W3C 2013a): Hier wird explizit ein Datenmodell beschrieben, welches versucht, einen generischen Standard für Notation und Austausch von Annotationen im WWW zu spezifizieren. Open Annotation inkludiert dabei sowohl DublinCore als auch die Linked Open Data Initiative FriendOfAFriend zur Notation von Personen, SKOS zur Spezifikation von Thesauri, RDF(S) zur Spezifikation von Ontologien und Ontologie-ähnlichen Wissensbasen, sowie die PROV Ontology (PROV-O) zur Modellierung von Herkunftsbezeichnungen und weitere Namespaces.

OpenAnnotation nennt explizit Annotationen im Netz, z.B. in sozialen Netzwerken, wie Youtube, Flickr etc., als Ausgangspunkt zur Entwicklung eines Austauschformats. Ziel ist ein erweiterbarer Standard zu Austauschbarkeit von Annotationen zwischen (web-) Plattformen. Dabei werden Annotationen verstanden als:

> „An annotation is considered to be a set of connected resources, typically including a body and target, and conveys that the body is related to the target" (Ebd.)

Der „body", also Körper, ist dann der Gegenstand, der über das „target", das Ziel, etwas aussagt, im einfachsten Fall kann dies ein Identifier sein, der eine Annotation mit einem Gegenstand verbindet. Insofern handelt es sich bei dem Datenmodell um ein Triple, welches dementsprechend auch in RDF gespeichert wird[15].

Neben dem sehr generischen Open Annotation Standard, der vor allem die Beziehung zwischen body und target beschreibt, sind viele weitere Dateiformate für Annotationen gebräuchlich. Um einige verbreitete Standards zu nennen, können Annotationen außerdem in folgenden Formaten notiert sein:

- JSON (annotator.js; via REST)
- XML Metadata Interchange (XMI)
- Weblicht TCF
- Webanno TSV
- Binary (jew. Annot.-Programm-spezifische Bearbeitungsform)

Da es sich hierbei jedoch nicht um Dateiformate handelt, die ausschließlich fachlichen Annotationen vorbehalten sind, sondern auch alle möglichen Arten von Informationen beinhalten können, kann hier ein „System" nicht automatisiert Annotationen als solche erkennen, sondern benötigt menschliche Hilfe, bspw. in Form eines Abfragesystems bei der Einlieferung von Forschungsdaten.

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen: Da es sich bei Annotationen um ein vielschichtiges Konzept handelt, bei dem sowohl auf technischer als auch auf fachwissenschaftlicher Seite noch viele Fragen unbeantwortet sind, können an dieser Stelle nur wenige grundlegende Empfehlungen ausgesprochen werden.

[15] Ein Beispiel für ein: OA Annotation findet sich auf: http://www.w3.org/community/openannotation/wiki/TextCommentOnWebPage

- Es wird eine Trennung von Annotationen (/Annotationsdaten) und vor allem technischen und administrativen Metadaten in unterschiedliche Dateien empfohlen.
- Nach Möglichkeit sollte der OpenAnnotation Standard zumindest als zusätzliches Austauschformat und zur späteren Verbreitung Verwendung finden.
- Wünschenswert ist außerdem für die pragmatische Verwendung in einer technischen Infrastruktur, dass im Ingest zumindest angegeben wird, ob und in welchen Dateien und Formaten sich Annotationen in einer Forschungsdatensammlung oder den Daten zu einem Forschungsvorhaben befinden.

4.4 Methoden & Tools

Gemäß der Diskussion in der AG sollten Methoden und Tools in den digitalen Geisteswissenschaften sich gegenseitig reflektieren und Ihr Einsatz auf den unterschiedlichen Arten geisteswissenschaftlicher Forschungsdaten angemessen wieder gegeben werden. Im folgenden Kapitel werden beide Begriffe in diesem Kontext definiert und in Beziehung zueinander gesetzt.

4.4.1 Methoden

Eine Methode ist ein standardisiertes und von der wissenschaftlichen Community anerkanntes Vorgehen, von dem pragmatische Verfahren abgeleitet werden können, welche wiederum auf Forschungsdaten (= dem Forschungsgegenstand) angewendet zur Beantwortung von wissenschaftlichen Fragestellungen verwendet werden (Reiche et al. 2014). Ein Verfahren ist demnach:

> „ein planvoller, systematischer und praktischer Umgang mit Forschungsgegenständen, dem die jeweilige Forschungsmethode stets übergeordnet bleibt. Ein wissenschaftliches Verfahren, wie z.B. der Vergleich mehrerer Textfassungen oder die Auswertung von Metadaten, kommt somit immer erst vor dem Hintergrund des sinnstiftenden Horizonts der jeweils gewählten Methode zum Einsatz. Zugleich erlauben solche Methoden der Forschung die Anwendung mehrerer Verfahren – durchaus auch eine Abfolge bestimmter Verfahren – um ein sorgfältig umgrenztes Erkenntnisinteresse und damit mindestens ein verbundenes Forschungsziel zu erreichen." (ebd. S. 6)

Es können folgende Feststellungen getroffen werden:

- **Ein Tool wird zur konkreten Umsetzung eines *Verfahrens*** verwendet, damit ein oder mehrere Methoden umgesetzt werden können.
- **Verfahren und Tools stehen in einer N:M Beziehung**: Ein Verfahren kann meist durch mehrere Tool umgesetzt werden. Andererseits kann auch ein Tool zur Ausführung mehrerer Verfahren verwendet werden.
- Auch **Tools und Methoden stehen in einer N:M Beziehung**: Eine Methode kann sich in eine Kette von Einzelverfahren gliedern, für die jeweils EIN Tool verwendet wird. Andersherum kann EIN Tool in unterschiedlichen Methoden zum Einsatz kommen (Abbildung 2).
- Weiterhin werden mithilfe der Tools **Objekte oder auch Forschungsdaten in Hinblick auf ein bestimmtes Verfahren hin analysiert oder auch editiert**. So können weitere Forschungsdaten entstehen.

Die Diskussion um Methoden und Verfahren der digitalen Geisteswissenschaften werden von einer breiten Community weltweit getragen. Grundsätzlich verfolgen diese Diskussionen das Ziel, das Fachgebiet „Digitale Geisteswissenschaft" vor allem in Bezug zu relevanten anderen Disziplinen zu definieren und so ihr Profil zu schärfen. Eine Herangehensweise ist dabei die Sammlung von konkreten Verfahren und Methoden in den Geisteswissenschaften und die Definition ihrer Beziehung[16]

Wichtigster Aspekt für ein Referenzmodell ist an dieser Stelle, nicht die Diskussion über die Sinnhaftigkeit einzelner Methoden sondern v.a. eine möglichst allgemeingültige Empfehlung zur Auswahl von Methoden für ein System (in diesem Fall für einen Forschungsdatenzyklus der DH). Hier kann als Basis für Empfehlungen die Nutzungshäufigkeit und Verbreitung von Datentypen einerseits als auch die Auswertung der Häufigkeit verwendeter Verfahren in möglichst vielen Teildisziplinen dienen. Naheliegenderweise soll es sich dabei um ein Methodenset handeln, das auch durch die eben erwähnten Sammlungen von Methoden der Digital Humanities abgedeckt und mit entsprechend verbreiteten Tools umgesetzt wird.

Auch sollte darauf geachtet werden, dass die gewünschten Methoden und Tools sich entsprechend ergänzen und gemäß den in einem Repository vorliegenden Datentypen gewählt werden.

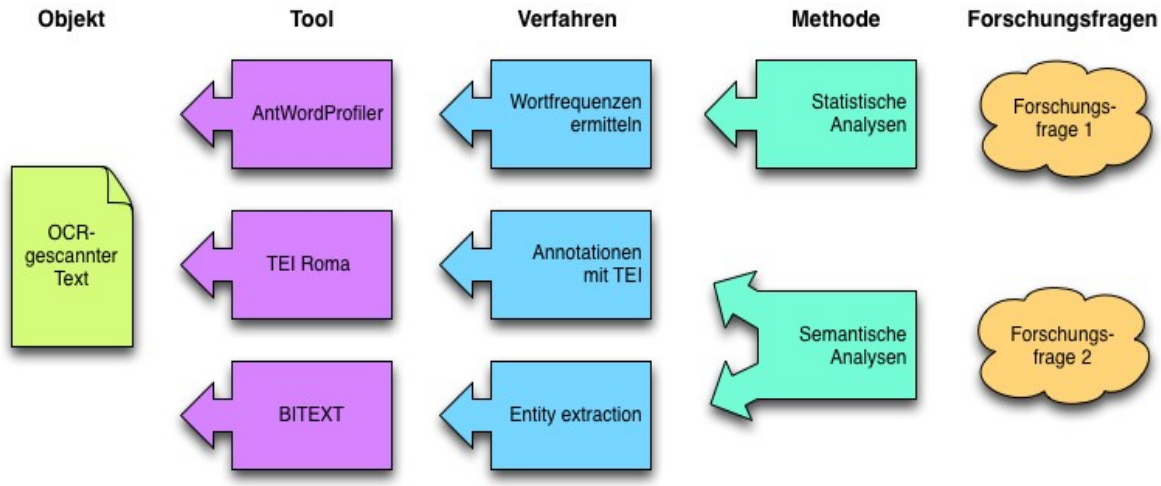

Abb. 2: Beziehung zwischen Objekten, Methoden und Tools

4.4.2 Tools

An dieser Stelle erscheint es sinnvoll, eine Auswahl von konkreten Tools zu treffen, welche in einer Mindestzahl von erwünschten Methoden vorkommt.[17] Denkbar ist die Auswahl gemäß den Disziplinen einer Zielgruppe, die eine Infrastruktur nutzen soll. Je vielfältiger diese sind, desto vielfältiger ist auch die Menge und Art der Tools, welche im Rahmen einer Infrastruktur unterstützt

[16] Vgl. dazu Reiche et al. 2014. Ein Verzeichnis von Verfahren und Methoden stellt beispielsweise TaDIRAH dar: https://github.com/dhtaxonomy/TaDiRAH. Ein Verzeichnis von Tools wird von DiRT angeboten: http://dirtdirectory.org/. Im Rahmen des NEDIMAH Projekts wurde schließlich versucht, mithilfe einer komplexen Ontologie die Beziehungen von Tools und Arbeitsweisen in den Digital Humanities abzubilden: http://www.nedimah.eu/workgroups/linked-data-and-ontological-methods

[17] Eine relativ breite aber nicht unbedingt spezifisch geisteswissenschaftliche Liste bietet die DiRT-Registry: http://dirtdirectory.org/.

werden müssen. Auch eine Sortierung gemäß vorliegender Medientypen erscheint möglich aber nicht zwingend sinnvoll.

Eine erste, explorative Sammlung von Software, die in den Geisteswissenschaften verwendet wird, sowie eine Befragung von Geisteswissenschaftlern hat gezeigt, dass grundsätzlich folgende – eher als generisch denn als genuin geisteswissenschaftlich zu bezeichnenden – Tools eingesetzt werden:

- Transkriptionstools
- XML Editoren zur Erzeugung inhaltlicher und struktureller Metadaten
- Annotationstools, welche konkret TEI oder Open Annotation als Austauschformat implementieren und validieren
- Annotationstools, welche speziell das Annotieren unterschiedlicher Medientypen erlauben
- Bildanzeige und -verarbeitungsmöglichkeiten
- Tools für die geospatiale Darstellung und Verarbeitung von Daten
- Unterstützung diverser statistischer Verfahren auf Text-, Bild- und Audiomaterial
- Software für linguistische Technologien (Spracherkennung, Frequenzanalysen, logikbasierte Analysen, linguistische Annotation…)
- Software zur Visualisierung solcher Verfahren und Daten
- Einbezug von externen Wissensbasen, also Ontologien und Taxonomien, die mit den entsprechenden Sprachen (RDF, RDF(S), OWL) in maschinenlesbarer Form vorliegen

Außerdem unterstützende Tools, die ganz allgemein in einer wissenschaftlichen Infrastruktur wünschenswert wären, wie

- Diverse Kollaborationsfunktionen (Wiki, Chat, gemeinsames Editieren in einem Texteditor…)
- Organisation von Bibliographien
- Dateisammlungen generieren und teilen
- Ggf. Ticketsysteme, Kalendersysteme oder andere Tools zur Organisationsunterstützung

Da sowohl die interne Sammlung der AG als auch die Umfrage nur begrenzt Einblick in die Fülle der tatsächlich verwendeten Tools geben kann, wird im folgenden Projektverlauf eine Analyse veröffentlichter Artikel hinsichtlich der zitierten Tools vorgenommen.

Umsetzungsempfehlung

Da die Beziehung von Tools und Methoden sehr komplex ist und im Feld der digitalen Geisteswissenschaften noch nicht hinreichend ausgearbeitet wurde, zumal nicht in konkreter Form, d.h. als Ontologie mit der genauen Definition von Beziehungen zwischen konkreten Softwareprodukten und Methoden, kann eine Infrastruktur vorerst nur eine Reihe unterstützender Tools, wie sie in der letzten Auflistung aufgezählt wurden, anbieten, um die kollaborative und begleitende Tätigkeiten bei der Erfüllung eines Forschungsprojekts zu erleichtern.

Hier ist durchaus zu diskutieren, was das dezidiert Geisteswissenschaftliche an einer solche Basisinfrastruktur ist, bzw. inwiefern eine solche Infrastruktur nicht WissenschaftlerInnen aller Disziplinen ohne Bezug zu den Digital Humanities zugute kommt. Allerdings eröffnete sich dabei das bereits zu Beginn dieses Dokumentes angesprochene Problemfeld der fehlenden allgemeinen Reflektion der Begriffe Geistes- und Kulturwissenschaften, digitale Geisteswissenschaften bzw.

Digital Humanities, von der Frage nach etwaigen spezifischen Methoden einmal ganz zu schweigen.

Perspektivisch sind Entwicklungen in Bezug auf konkrete Ontologien zu berücksichtigen und auch entsprechend maschinenverwertbar in eine Infrastruktur zu integrieren. Daneben soll an dieser Stelle auf entsprechende redaktionell betreute Empfehlungen im Web verwiesen sein.[18]

[18] Liste von Empfehlungen im DARIAH Wiki: https://dev2.dariah.eu/wiki/pages/editpage.action?pageId=38080370

5 Infrastrukturelle Funktionen

Die Motivation eines Research Data Life Cycle ist im Prinzip die Formalisierung eines Zyklus von in sich geschlossenen inhaltlichen Tätigkeiten / Anwendungen, die von einer Infrastruktur unterstützt werden müssen, damit ein Forschungsunterfangen – von der Erstellung von Forschungsdaten bis hin zu Nachnutzung der Daten – möglichst reproduziert, wenigstens aber intellektuell nachvollzogen werden kann.

Dieses Kapitel beschreibt **Funktionen**, die für einen Forschungsdatenzyklus in einer Infrastruktur abgedeckt werden sollen. Stichworte sind hier: Ingest, Verwaltung, die Vergabe von persistenten Identifiern und Metadaten, Dissemination, Publikation, Kuration.

Außerdem ist eine gewünschte Kernfunktion des RDLC der **modulare Aufbau** , d.h. die Erweiterbarkeit um bzw. Austauschbarkeit von Bestandteilen, also Methoden, Formaten, Tools, Metadaten. Eine Kerneigenschaft all dieser Funktionen ist, dass sie normalisierend wirken sollen. Was genau unter Normalisierung verstanden wird, beschreibt das folgende Kapitel:

5.1 Normalisierung

Normalisierung wird allgemein verstanden als der Prozess der Vereinheitlichung zur Steigerung praktischer Lösungsansätze. Normalisierung wird hier als erweiterter Begriff aus dem Feld der Langzeitarchivierung verwendet und bezeichnet dort Prozesse der Vereinheitlichung von Dateiformaten, aber auch von unterschiedlichen Metadatenstandards zur besseren Verwaltung und Auffindbarkeit[19]. Insofern sind Normalisierungsprozesse von ganz grundsätzlichem Interesse für den erfolgreichen Aufbau einer Infrastruktur für Forschungsdaten. Insbesondere um Abläufe möglichst automatisiert abbilden zu können, ist diese unumgänglich:

Es ist unabdingbar, Standards und Formate für alle Forschungsdaten, die im System verarbeitet werden können, festzulegen oder zumindest zu spezifizieren, für welche Formate und Standards welche Verarbeitungsschritte erfolgen. So kann es zum Beispiel durch Normalisierung möglich sein, anhand der verwendeten wissenschaftlichen Methode im System automatische Rückschlüsse auf den verwendeten Medientyp und das verwendete Metadatenformat zu erlauben oder anders herum anhand des Medientyps oder Metadatenformats Schlussfolgerungen auf eingesetzte Verfahren zuzulassen. Beide Arten von Rückschlüssen könnten WissenschaftlerInnen bei der Recherche nach Forschungsdaten und Forschungsprojekten zu einem bestimmten Feld immens unterstützen.

Ganz grundsätzlich ist hier festzustellen, wie hoch der Bedarf nach Normalisierung ist und in welchem Ausmaß er befriedigt werden kann. Dies gilt implizit für jede der folgenden genannten Funktionen und wird jeweils am Schluss der Kapitel in einer „Umsetzungsempfehlung" diskutiert.

[19] Vgl. zur Normalisierung von Metadaten: DP4Lib 2014 bzw. speziell zur Formatnormalisierung: Archivematica 2014.

5.2 Vergabe persistenter Identifier

Der Nutzen von eindeutigen und persistenten Identifiern in Repositories und Nachweissystemen sollte mittlerweile als nachgewiesen gelten (Tonkin 2008). So können Links und Verweise in einem System langfristig überdauern und behalten auch nach – beispielsweise – Technologiebrüchen, Serverumzügen oder Firmenübernahmen weiterhin ihre Gültigkeit. Auf diese Art ist eine dauerhafte Verfügbarkeit und ein dauerhafter Zugriff auf Daten in einer Infrastruktur sicher gestellt

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen:

Das oben diskutierte Objekt- und Forschungsdatenparadigma und die Differenzierung zwischen Daten und Metadaten nach ihrer inhaltlichen Relevanz in getrennten Dateien, führen zu einer relativ klaren Empfehlung zur Vergabe von Identifiern. Demnach sollten Identifier unbedingt **auf Dateiebene** vergeben werden, um langfristig die Nachweisbarkeit und den Zugriff durch Software oder Webtools auf diese zu gewährleisten. Ergänzend ist eine Identifier-Vergabe auf Objektebene erstrebenswert. Je nach vorgesehener Verschachtelung und Detailtiefe von Objekten bis zur Dateiebene kann hier eine große Zahl an Identifiern vorgesehen werden.

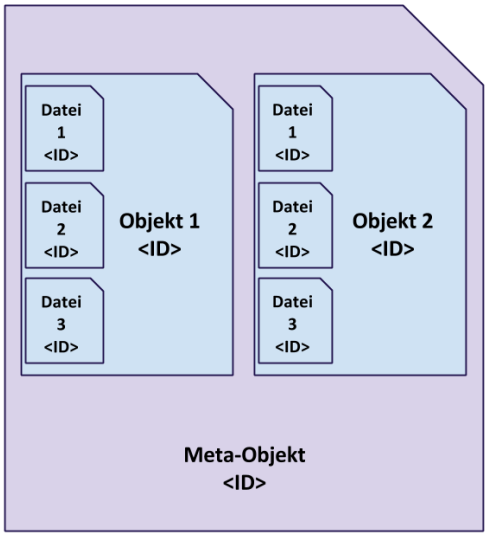

Abb. 3: Identifiervergabe bei komplexen Objekten in den Metadaten

Alle vergebenen Identifier können sodann im Forschungsdatensystem zu Nachweis und Weiterverarbeitung einzelner Objekte verwendet werden.

Die in DARIAH-DE verwendete PID-Technologie[20] unterstützen den Einsatz von Identifieren schon in frühen Phasen eines Forschungsprozesses, wenn noch nicht notwendigerweise fest steht, ob die referenzierten Daten alle später erhalten bleiben sollen. Dies erleichtert den in DH-Projekten erstrebten frühzeitigen Austausch über Objekte und die Kollaboration mit diesen Forschungsdaten als eine informelle Art des Peer-Reviews.

Auch die Bündelung von einzelnen Dateien zu komplexen Objekten via PIDs in den Metadaten ist gegeben, so dass Dateien in unterschiedlichen Forschungskontexten neu zusammen gestellt und analysiert werden können (Aurast 2014).

[20] Vgl. https://de.dariah.eu/pid-service

5.3 Metadaten-Anreicherung

Grundsätzlich wird zwischen deskriptiven, administrativen, technischen und strukturellen Metadaten[21] unterschieden. Häufig ist gerade der Bereich der deskriptiven Metadaten noch einmal unterteilt in bibliografische (Titel, Autor, Inhaltsangabe) und andere inhaltliche Anmerkungen (gerade im Fall von Annotationen und TEI-Notationen). Während aber bibliografische Angaben der besseren Auffindbarkeit dienen, dienen Annotationen und inhaltliche Metadaten eher der intellektuellen Beschäftigung mit einem Text. Während also inhaltliche Metadaten das Ergebnis eines intellektuellen Prozesses darstellen und häufig sehr fachspezifisch strukturiert sind, können administrative, technische und strukturelle Metadaten häufig maschinell erhoben und gespeichert werden. Bibliografische Angaben, also klassischer Weise deskriptive Metadaten liegen dazwischen: Sie müssen einmal manuell eingefügt werden, dienen aber dennoch eher der Auffindbarkeit und Verwaltung in einem System und nicht so sehr der inhaltlichen Auseinandersetzung.

Für die Abbildung von **Workflows** in den digitalen Geisteswissenschaften ist daher ein generischer Metadatenstandard wünschenswert, der vor allem administrative, technische und strukturelle Informationen enthält und in den ggf. andere Metadaten überführt werden können[22].

Angestrebt wird daher auch eine möglichst **automatisierte Vergabe von Metadaten** sowie eine Validitätsprüfung von bereits vorhandenen administrativen, technischen und strukturellen Metadaten im Ingest von Forschungsdaten[23]. Inhaltliche Metadaten können weder automatisch überprüft werden noch gibt es für sie einen einheitlichen **disziplinübergreifenden Metadatenstandard**, in den diese überführt werden können.

An Metadaten sind Informationen relevant, die:

- den Kontext eines Forschungsprojekts,
- beteiligte Institutionen,
- Personen,
- Datumsangaben,
- zusammengehörige Dateien,
- Informationen zu Dateiformaten
- Informationen zu erstellender Software
- Referenzen zu Versionen einzelner Dateien
- deren Identifier
- rechtliche Informationen

etc. beinhalten.

Technische, administrative oder strukturelle Metadaten sollen in einer Infrastruktur folgenden Zwecken dienen:

- der Abbildung eines Research Data LifeCycle,
- der Wiederauffindbarkeit von Forschungsdaten (über PIDs und Referenzen)
- der inhaltlichen Zuordnung der Daten zu einzelnen Projekten

[21] Vgl. http://www.digitalisierung.ethz.ch/metadaten.html.

[22] Für den Versuch des Mappings von Metadatenfeldern aus einem Schema in ein anderes vgl: http://dev3.dariah.eu/schereg/.

[23] Im Bereich der administrativen und ggf. strukturellen Metadaten, siehe Kapitel Kuration.

- ggf. der Unterscheidung ihrer Art in Inhalts- und Metadaten
- der Unterscheidung weiterer Datentypen, z.B. des Projektantrags und angegliederter Online-Ressourcen, wie Wikis.

Es bieten sich also Metadatenstandards an, deren Zweck insbesondere die Abbildung von Workflows im Forschungsumfeld ist. Leider sieht die Situation hier für den interdisziplinären Bereich der digitalen Geisteswissenschaften eher mager aus: Mehrere wissenschaftliche Organisationen haben bereits versucht, eine Übersicht für Metadatenstandards aus unterschiedlichen Disziplinen zu sammeln und zu referenzieren. (DCC 2015 oder UMN 2015 oder auch Riley 2010) Auch in DARIAH-DE hat es bereits einen solchen Versuch gegeben, der allerdings in erster Linie Vorschläge zur Erweiterung von (dem eigentlich inhaltlich verwendeten) TEI Standard enthielt (Aurast 2014). Letztlich besteht aber die Anforderung darin, einen Metadatenstandard zu finden, der unabhängig von inhaltlich verwendeten Metadaten für Forschungsdaten die oben aufgezählten Informationen speichert und strukturiert.

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen:

Grundsätzlich wird für den Bereich der digitalen Geisteswissenschaften der Ansatz verfolgt, dass zur inhaltlichen Beschreibung der Daten jegliche Schemata möglich und verwendbar sein sollen. Es wird daher auf Einschränkungen bei der Annahme von Dateien in unterschiedlichen Metadatenformaten verzichtet. Jedoch sollten Empfehlungen zur Verwendung bestimmter Metadatenstandards ausgesprochen werden und aus diesen administrative, technische und strukturelle und auch deskriptive Informationen extrahiert werden können.

Weiterhin sollte gelten:

- Als Minimalstandard für deskriptive Metadaten in der Infrastruktur sollte DublinCore gelten.
- Da der Informationsgehalt von DublinCore allerdings minimal ist, sollte nach Möglichkeit ein weiterer Standard verwendet werden, der möglichst DublinCore Elemente aufnehmen kann, aber weit darüber hinausgeht und in Metadatenform nicht nur deskriptive Informationen sondern für jedes „Forschungsdatum“ eine vollständige Übersicht aller damit verbundenen Informationen enthält.
- Hier sind sowohl W3C Prov (W3C 2013b) als auch PREMIS (PREMIS 2012) oder der Metadatenstandard der Data Documentation Inititative (Hoyle 2013) vorstellbar.
- Die genaue Verwendung des jeweiligen Metadatenstandards in einer Infrastruktur, also welche Felder mit welchen Informationen gefüllt werden, ist eng an das Arbeitsfeld der Kuration gekoppelt und kann an dieser Stelle nicht entschieden werden.

Wie auch im Fazit dieses Dokuments diskutiert, ist die Entscheidung für oder gegen einen Metadatenstandard, der Provenienz oder auch Workflows modellieren kann, in DARIAH-DE noch nicht gefallen und erfordert die tiefere Evaluation bestehender Ansätze, welche im kommenden Jahr vorgenommen werden soll.

5.4 Kuration und Speicherung (LZA)

Die deutsche Forschungsgesellschaft empfiehlt in Ihren „Empfehlungen der Deutschen Forschungsgemeinschaft zur Sicherung guter wissenschaftlicher Praxis" eine Aufbewahrungsfrist von Primärdaten (also Daten, die die Grundlage zur Beantwortung einer Forschungsfrage bilden) von 10 Jahren (DFG 2013). Auch in DARIAH-DE sollten **10 Jahre als absolute Untergrenze** für eine Langzeitarchivierung von Forschungsdaten verstanden werden. Grundsätzlich wird in einer solchen Komponente – wie auch in der Grafik des Projektantrages von DARIAH – zwischen der Kuration und der langfristigen Aufbewahrung von Daten differenziert (Lazorchak 2011).

Ein erste grobe Definition des Kurationsbegriffs bezeichnet folgende Tätigkeiten als Kuration von Objekten, so wie es Gedächnisinstitutionen für das ihnen überlassene Erbe praktizieren. Zur Kuration zählen folgende Operationen auf Objekten für die Archivierung und Bereitstellung.:

- die thematische Zuordnung,
- die Gliederung,
- die (deskriptive) Metadatenanreicherung
- die Auswahl

Die Speicherung selbst ist als eher technische Prozedur zu betrachten und soll zu allererst die Erhaltung des Bitstreams, die Gewährleistung der Wiederauffindbarkeit und Speicherung aller relevanten Referenzen garantieren. Neben der Bitstreampreservation sind weitere Funktionen zur (technischen) Erhaltung des Zugriffs auf Dateien denkbar, wie das extensive Speichern technischer Metadaten, die Anwendung von Formatmigration oder sogar – bei Bedarf – die Emulation veralteter Software auf entsprechenden Servern. Solche Operationen werden in Kapitel 5.4.2 diskutiert.

Im Allgemeinen wird die so eben vollzogene begriffliche **Trennung zwischen Kuration (Pflege) und Speicherung (Erhaltung, Aufbewahrung)** nicht immer so eindeutig gesehen. Im Gegenteil: Mal bezeichnet gerade im anglo-amerikanischen Raum der Begriff „Curation" eher das Forschungsdatenmanagement insgesamt sowie alle diesem Begriff zugeordneten konzeptionellen Tätigkeiten, mal wird „Curation" synonym mit „Preservation" verwendet, was eher die technischen Probleme des Zugangs zu digitalem Erbe fokussiert (Lazorchak 2011). Wenn man nicht alle Tätigkeiten und Funktionen eines Research Data LifeCycles mit einem Begriff abdecken möchte, so ist im Rahmen der DARIAH-DE Implementation eines Forschungsdatenzyklus der Begriff der Kuration wohl am ehesten mit folgenden Tätigkeiten beschrieben:

> „curation" [...] concentrates on underpinning activities of building and managing collections of digital assets and so does not fully describe a more broad approach to digital materials management." (Ebd.)

Zur pragmatischen Implementation eines Forschungsdatenzyklus sind viele der genannten „kurativen" Tätigkeiten schon zu Beginn eines „Zyklus" – bei der Einlieferung der Daten in ein System – von den Forschenden selbst zu bewältigen oder zumindest vorzubereiten. Hierbei handelt es sich um Tätigkeiten, wie das Eintragen deskriptiver Metadaten, die richtige Referenzierung zu anderen Objekten und Projekten, sowie ggf. die Anreicherung mit weiteren Informationen, bspw. das Anbringen von zusätzliche Verweisen, rechtlichen Nutzungsbedingungen, Projektkontext, etc. Dabei empfiehlt sich der folgende Ansatz:

Die – eher inhaltliche, manuelle – Aufgabe der Kuration kann aufgrund des damit verbundenen Arbeitsaufwandes schnell eine abschreckende Wirkung entfalten. Daher ist bei einem automatisierten Ingest von Daten in eine Infrastruktur, die *manuelle Erweiterung* von Metadaten und ggf. die *manuelle Verknüpfung* von Daten zu komplexen Datenobjekten als möglichst gering zu veranschlagender Zusatzaufwand zu kennzeichnen.

Sollte dieser Arbeitsaufwand in einem System obligatorisch werden oder zu späteren Zeitpunkten erneut vorgenommen werden müssen, sollte vorab alles unternommen werden, damit diese Arbeit den NutzerInnen erleichtert werden kann, so dass der Abschreckungseffekt ausbleibt. Hier sind unbedingt entsprechende Empfehlungen aus der internationalen Community, also Vereinigungen wie *DPC* (Digital Preservation Coalition), *DCC* (Digital Curation Centre), *LOC* (Library of Congress) etc. zu berücksichtigen.

Im Unterschied dazu sollten *automatische* Funktionen zur Sicherung der langfristigen Verfügbarkeit aller digitalen Forschungsdaten und – soweit möglich – automatische Kurationsstrategien hingegen so umfangreich und autonom wie möglich implementiert werden.

5.4.1 Exkurs: das OAIS Modell

Grundsätzlich soll das Open Archival Information System (kurz: OAIS) Modell (OAIS 2012) als Grundlage zur langfristigen Speicherung und Verfügbarkeit der Forschungsdaten dienen. Das OAIS Modell wird gemeinhin als Referenzstandard für die Langzeitarchivierung digitalen Erbes herangezogen und beschreibt den vollständigen Archivierungs- und Bereitstellungspozess von der Aufnahme von Dateien in ein Archivsystem (**Ingest**), ihrer Aufbereitung und Anreicherung (**Preservation Planning**) bis hin zu Ihrer langfristigen Speicherung (Storage) einerseits bzw. ihrer Publikation / Verbreitung (**Access**) andererseits.

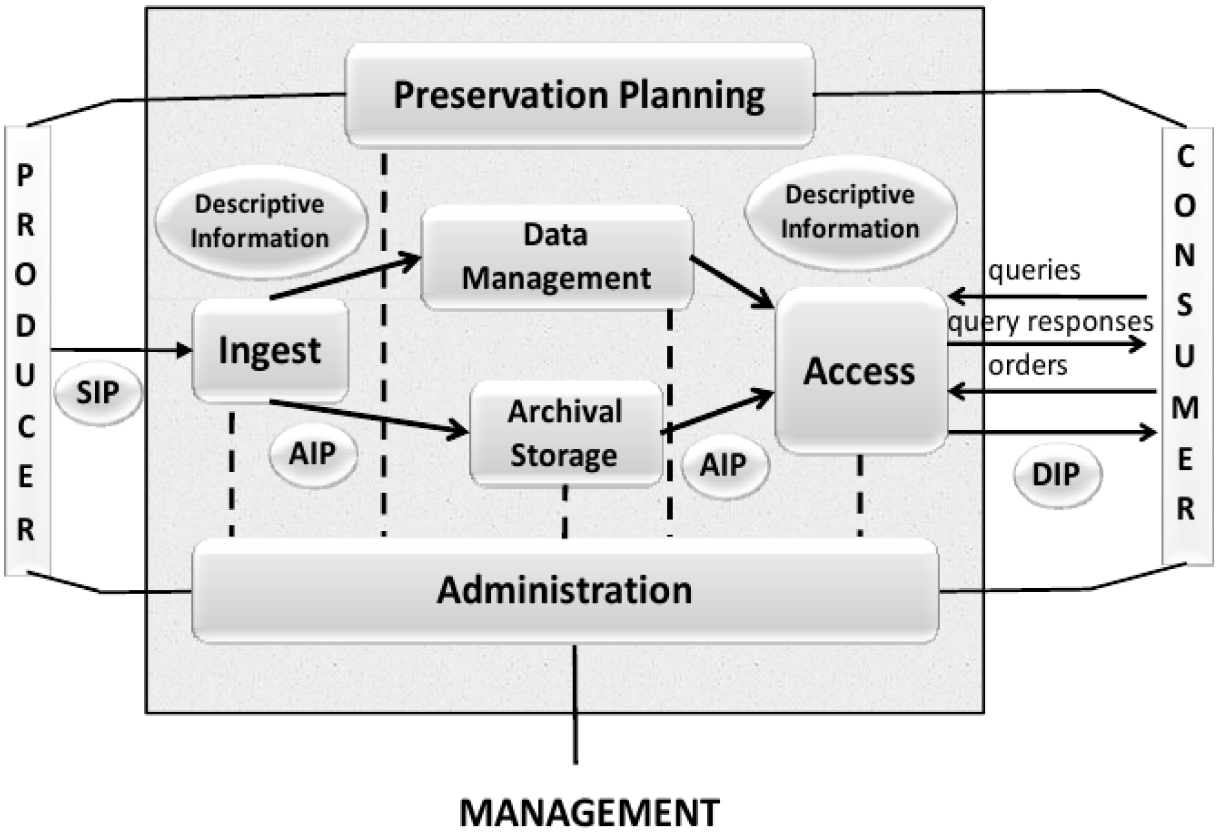

Abb. 4: Das OAIS Modell, die bekannteste Grafik (The OAIS functional model, Ebd. S. 4-1)

Dabei ist durchaus die Tätigkeit, die im OAIS Modell als Data Management und Preservation Planning beschrieben wird, nicht nur für Aspekte der Langzeitarchivierung interessant sondern muss auch im Kontext eines Systems zur Verarbeitung und Bereitstellung von Forschungsdaten beachtet werden.

Die Arten von Informationsobjekten, die OAIS beschreibt, sind

- das Submission Information Package (**SIP**), als Paket zusammengehöriger Daten im Ingest
- das Archival Information Package (**AIP**) als das Informationsobjekt, welches im Langzeitspeicher abgelegt wird, und das
- Dissemination Information Package (**DIP**), welches Objekte beschreibt, die zur Veröffentlichung und Verbreitung erstellt wurden

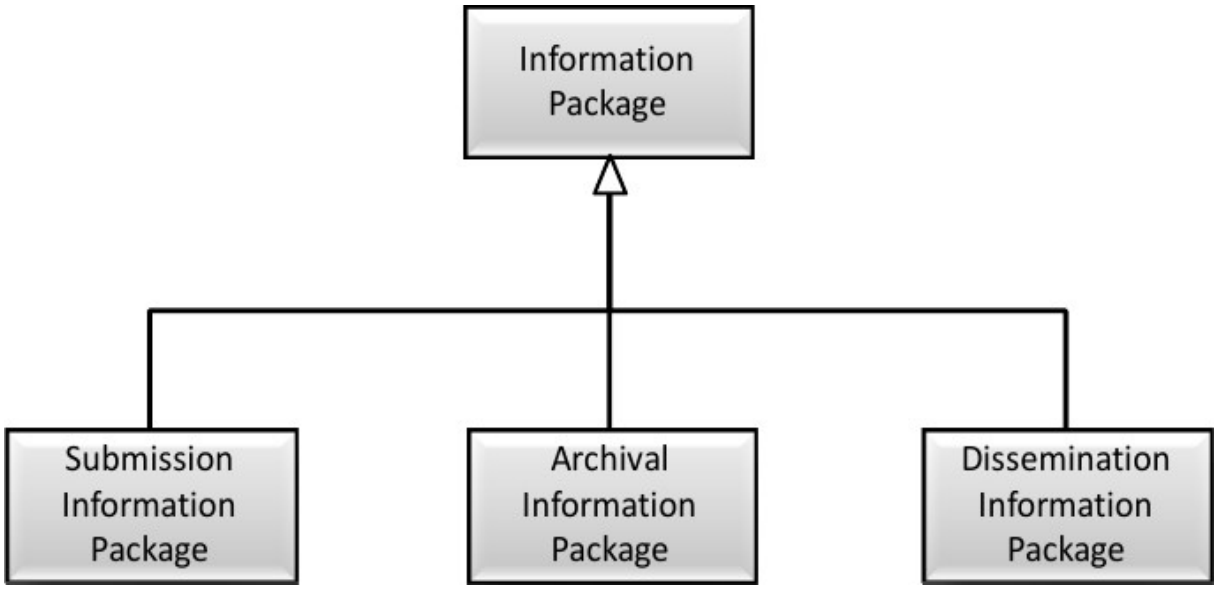

Abb. 5: Pakettypen im OAIS Modell. (Ebd. Figure 4-14: Information Package Taxonomy. S. 4-35)

Dabei besteht ein Informationsobjekt im OAIS Modell grundsätzlich aus „Content Information" und „Preservation Description Information", beide zusammen werden durch „Descriptive Information" als ein Objekt referenziert (Ebd.). „Content Information" bezeichnet in erster Linie den Begriff für das inhaltliche Datenobjekt selbst in Addition mit „Representation Information", was die darin enthaltenen Metadaten zur richtigen Interpretation im Dateisystem meint (bspw. Angabe des MIME-Types im Header). „Preservation Description Information" enthält hingegen alle notwendigen Informationen, welche zur langfristigen Erhaltung des Objekts notwendig sind. Hier werden Informationen über Herkunft, Dateiformat, eine Prüfsumme und auch ein Identifier empfohlen.

5.4.2 Notwendige Operationen und Empfehlungen für Langzeitarchivierung

Gemeinhin werden drei oder mehr aufeinander aufbauende Strategien zur Langzeitarchivierung digitaler Daten beschrieben, welche hier nur ganz kurz skizziert werden sollen[24]:

- **Bitstreampreservation**:
  Als Minimallösung gilt es, den Binärstrom einer Datei als solches zu sichern und diese Sicherung überprüfbar zu gestalten. D.h. Bestandteile der Bitstreampreservation sind die Sicherung gegen **Dateibeschädigung** – ob durch Kopierprozesse oder Hardwareüberalterung. Dabei werden folgende Maßnahmen ergriffen:
  - **redundante** Datenhaltung der gleichen Daten an mehreren Speicherorten,
  - regelmäßiges Umkopieren auf aktuelle Datenträger (also **Hardware-Migration** der Daten von veralteten Datenträgern auf aktuelle Alternativen),
  - und die Überprüfung der korrekten Datenhaltung durch das Vergleichen von mitgespeicherten Fingerprints oder **Prüfsummen**.
- **Metadatenextraktion**:
  Da nach einer gewissen Zeitspanne nicht mehr sichergestellt werden kann, ob aktuelle Betriebssysteme die gespeicherten Dateiformate noch erkennen und auch weil ein

[24] Vgl. dazu beispielsweise Ausführungen zu „Preservation Service Levels" in DPC 2008, S. 64, oder Klimpel, Keiper 2013, S. 35 ff.

gewisser Informationsverlust bei evtl. späterer Dateiformatmigration unvermeidbar ist, ist eine möglichst umfassende Speicherung **technischer Metadaten** sinnvoll. Diese sollten gleich zu Beginn (also beim Ingest der Daten in ein System) mithilfe technischer Prozesse extrahiert und mit den Daten mitgespeichert werden.

- **(Format-) Migration:**
  Da der Fall absehbar ist, dass Daten älterer und ggf. proprietärer oder anderweitig problematischer Dateiformate in Zukunft nicht mehr von Anwendersoftware geöffnet werden können, ist langfristig eine Migration der betroffenen Dateien in aktuelle, idealerweise weit verbreitete, nicht proprietäre Formatstandards sinnvoll[25]. Hier handelt es sich um ein langfristiges (auch personelles und finanzielles) Engagement, da Migrationsentscheidungen auch im laufenden Archivbetrieb von geschultem Personal getroffen und umgesetzt werden müssen.
- **Emulation**:
  Die Strategie der Emulation geht über die reine Formatmigration noch einen Schritt hinaus, und die Forschung hierzu befindet sich eher in einem **experimentellen** Stadium. Dabei wird angenommen, das Dateiformate zu spezifische Eigenschaften haben, so dass der Informationsverlust oder auch der Verlust von **Editierbarkeit** durch Anwendungsprogramme durch Formatmigration zu große Ausmaße annimmt. Daher wird empfohlen, ganze Softwarepakete zusammen mit den zu archivierenden Dateien zu speichern und diese auf Rechnerarchitekturen, ggf. auf einer virtuellen Maschine, die die Rechnerarchitektur der **Originalumgebung der Software** nachbildet, auch für die Zukunft lauffähig zu halten. Es ist leicht zu erfassen, dass diese Strategie die Aufwändigste der genannten Strategien ist, um einzelne Dateien zugreifbar zu halten, weswegen es durchaus umstritten ist, für den Erfolg von Emulationsstrategien Vorhersagen zu treffen.

Neben der adäquaten Empfehlungen für Langzeitarchivierungsstrategien müssen außerdem Empfehlungen für die Eignung von in den Geisteswissenschaften weit verbreiteten Dateiformaten zur Langzeitarchivierung ausgesprochen werden. Denn neben der nicht immer umgesetzten Strategie „Migration", bei der dann eine Infrastruktur selbständig Dateien in langfristig empfohlene Dateiformate migrieren kann, können auch die einliefernden WissenschaftlerInnen selbst schon Einiges zur langfristigen Nutzbarkeit ihrer Daten beitragen, in dem sie ihre Dateien in den empfohlenen Dateiformaten abliefern.

Dabei gibt es durchaus konfligierende und teilweise auch sehr **differenzierte Empfehlungen** für die Langzeitarchivierungstauglichkeit von Dateiformaten: Verschiedene Organisationen empfehlen für bestimmte Medientypen (Bild-, Text-, Multimediadateien) jeweils unterschiedliche Dateiformate zu Langzeitarchivierung. So empfiehlt das Florida Digital Archive in seinen „FDA File Preservation Strategies by Format" (FLVC 2012a) sowohl JPEG2000 part1 als auch **TIFF (Baseline**) als geeignete Dateiformate zur Langzeitarchivierung von Bildern. Tiff (compressed) aber auch **JPEG 2000** part 2 bzw. lossy compressed wird hingegen dezidiert eine niedriges Vertrauenslevel ausgesprochen. Da sowohl JPEG2000 als auch Tiff mit und ohne Kompression angeboten werden, aber nur JPEG 2000 auch verlustfreie Kompression anbietet und somit u.U. auch als Ansichtsformat in webbasierten

[25] Eine Liste von Kriterien für geeignete Dateiformate für die digitale Langzeitarchivierung findet sich hier: NDIIPP 2013.

Anwendungen verwendet werden kann, galt JPEG 2000 lange Zeit als das Dateiformat der Wahl für Bilder. In jüngster Zeit werden allerdings vermehrt Stimmen laut, die die überkomplexe Struktur und nicht sehr weite Verbreitung des JPEG 2000 Standards als Kritikpunkte werten und somit eher zur Nutzung von (Baseline) Tiff raten (Succeed 2014) (LeFurgy 2013).

Ein anderer Aspekt der Problematik betrifft das Feld der **Editier- und Nachnutzbarkeit**: Unter der Prämisse, dass die abgelegten Daten nicht nur zur Ansicht sondern tatsächlich zur Weiterverwendung gespeichert werden, muss der Aspekt der Editierbarkeit ein viel größere Rolle spielen: Allgemeinhin wird z.B. für die Langzeitarchivierung von Textdokumenten **PDF/A** als Dateiformat empfohlen, meist in dem vollen Bewusstsein, dass PDF ein Dateiformat zur (plattformunabhängigen) Darstellung, nicht aber zur weiteren Editierbarkeit eines Dokuments darstellt.

Gerade die in den Digital Humanities nach wie vor weit verbreitete Nutzung von (häufig unterschiedlichen) **Office-**Produkten und den daraus resultierenden Dateiformaten macht es hier schwierig, speziell die Nachnutzbarkeit von Office-Formaten zu ignorieren. Im Gegensatz zu Office-Dokumenten speichert das vielfach empfohlene PDF-Format keinerlei Strukturinformationen, wie Fußnoten, Kapitelüberschriften, Kapitelhierarchien etc. Eine Konversion eines Office-Dokuments in PDF bedeutet also grundsätzlich immer einen substantiellen Verlust hinsichtlich der Editierbarkeit, da sämtliche Formen intertpretativer Angaben zu graphischen Anweisungen umgesetzt werden und zwar so, dass dieser Prozess nicht direkt umkehrbar ist. Exemplarisch wird bei diesem Problemfeld eine **zweigleisige Vorgehensweise** empfohlen: Zur Bewahrung der optischen Erscheinung eines Dokuments kann man die Konvertierung nach PDF/A grundsätzlich empfehlen, zur Erhaltung der Editierbarkeit und der Dokumentstruktur ist daneben eine Erhaltung des Originaldokuments sowie ggf. eine Konvertierung in ein einheitliches (markup-basiertes) offenes Office-Format (bpsw. ODT[26] o.ä.) zu überlegen (Paradigm 2008).

Umsetzungsempfehlung

Es werden Empfehlungen auf zwei unterschiedlichen Ebenen ausgesprochen: Zum einen wird von DARIAH-DE eine nicht exklusive Liste von empfohlenen Dateiformaten vorgehalten, unabhängig davon, ob innerhalb einer Infrastruktur Tools zu ihrer Weiterverarbeitung angeboten werden können und sollen[27].

Zum anderen wird der Bedarf nach einem Datenmodell zur Dokumentation komplexer Forschungsdaten in einer Infrastruktur festgestellt und dessen Konzeption angekündigt. Hier soll definiert werden, was eigentlich „unterstützt" im Rahmen von Kuration und Speicherung bedeutet. Es zeigt sich zunehmend der Bedarf nach einem Datenmodell in einem adäquaten Beschreibungsstandard, in dem eine Infrastruktur Informationen über Objekte vorhält**.** Zum einen müssen für jedes Datenobjekt **Basisinformationen** nachgehalten werden (Erstellungsdatum, Ersteller, Dateiformat, Größe, Prüfsumme, bei mehreren Dateien: Pfad und Identifier für alle zusammengehörenden Dateien). Dies lässt sich auch über „einfache" Metadatenstandards, wie DublinCore realisieren. Daneben müssen sowohl **kurative Informationen** (Provenienzinformationen über verbundene Forschungsprojekte, Referenzen zu inhaltlich

[26] Über die Archivierbarkeit von Officeformaten: ADUK 2009.

[27] Vgl: https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=38080370

verbundenen Dateien) als auch Informationen zur Langzeitarchivierung vorgehalten werden. Bspw. können technisch relevante Informationen zu Dateien mithilfe von Tools, wie JHOVE[28] extrahiert und in Metadaten abgespeichert werden, so dass sie in einem Langzeitarchiv wertvolle Zusatzinformation zur Weiterverarbeitung der vorgefundenen Daten liefern können. Andere Informationen müssen unter Umständen **manuell erhoben** werden.

Die Wahl der sowohl kurativen als auch langzeitarchivierungsrelevanten Metadatenfelder kann sinnvollerweise durch einen schon bestehenden Standard abgenommen werden und nur noch an einigen Stellen ergänzt werden. Zur Debatte steht aktuell bspw der PREMIS Standard der Library of Congress (PREMIS 2012). Darüber hinaus ist zu überlegen, wie ein solcher Metadatenstandard mit allen weiteren Features eines Forschungsdatenzyklus interagiert und ggf. in einem komplexeren Datenmodell zur Abbildung des vollständigen Forschungszyklus aufgehen kann.

Die Arbeit der AG soll im zweiten Projektjahr ein robustes, verlässliches Datenmodell auf Basis eines solchen Standards hervorbringen, welches uneingeschränkt zur Dokumentation des hier beschriebenen Lebenszyklus verwendet werden kann.

5.5 Publikation

Grundsätzlich sollte angestrebt sein, die in einem solchen Forschungsdatensystem produzierten Ergebnisse entsprechend früh und nachhaltig zu publizieren.

Dabei muss die wissenschaftliche (textbasierte) Publikation nicht in dem gleichen System erfolgen in dem auch die Forschungsdaten selbst erstellt oder modifiziert wurden, es sollte aber darauf geachtet werden, dass bei Abschluss des Projekts eine **wechselseitige und persistente Referenzierung** erfolgt. Diese Referenzierung sollte technisch mithilfe von eindeutigen, persistenten Identifiern erfolgen, inhaltlich sollte auch ein natürlich-sprachlicher Hinweis bspw. auf den Titel und Autor einer Sammlung **wissenschaftlicher Daten** bzw. Titel und Autor einer **Publikation**, die mit diesen Daten verbunden ist, gegeben werden. Auch damit verbundene Projekttitel, Projektträger und Laufzeiten sind an dieser Stelle sinnvoll unterzubringen.

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen:

- Alle Forschungsprojekte, die in einer Infrastruktur ihre Forschungsdaten ablegen und dort referenzieren, sollten zum Ende ihrer Projektlaufzeit dazugehörige Publikationen nachweisen und entsprechende **URIs** in einer entsprechenden persistenten Infrastruktur hinterlegen.
- Die mit einem Forschungsprojekt verbundenen Daten sollten zumindest innerhalb der Forschungsinfrastruktur am Projektlaufzeitende **öffentlich verfügbar** sein, so dass ihre Nachnutzbarkeit gewährleistet ist
- Ebenso soll die Verwendung entsprechender standardisierter Metadatenfelder und **Suchfunktionen** in einer **Publikationsoberfläche** für entsprechende Auffindbarkeit und Sichtbarkeit sorgen.

[28] Vgl. http://jhove.sourceforge.net/

5.6 Peer-Review

Langfristig hat ein Forschungsergebnis nur dann wissenschaftlich Bestand, wenn ein Peer-Review Verfahren fester Bestandteil des Forschungsprozesses ist. Dabei ist zu berücksichtigen, dass der Forschungsprozess in den Digitalen Geisteswissenschaften grundsätzlich viele Ansatzpunkte für ein Community-Feedback bietet:

Während das klassische Peer-Review Verfahren vorsieht, dass ein Forschungsprojekt einerseits vor der Förderbewilligung oder aber andererseits ein Artikel mit Forschungsergebnissen vor der Veröffentlichung von mehreren Fachvertretern auf Qualität geprüft wird, werden Projekte der digitalen Geisteswissenschaften bereits in der Konzeptionierungsphase und von da an **immer wieder der Revision unterzogen**. Das liegt zum einen in der Natur des Forschungszweiges – die die Abstimmung zwischen Fachwissenschaft und Informatik notwendig macht – zum anderen in den oft interdisziplinär angelegten Projektstrukturen. Die Zusammenarbeit zwischen Informatik und Geisteswissenschaft kann dabei sehr unterschiedliche Formen annehmen, die vom eher praktisch angelegten technischen Support bis hin zur gemeinsamen Entwicklung neuer Tools und Methoden reichen kann.

Die **Kollaboration** mit anderen Fachdisziplinen wird hingegen immer zu einem gegenseitigen Austausch von Kooperationspartnern mit unterschiedlichen Perspektiven auf einen Datensatz führen, der den Fokus des Gesamtprojektes erweitert und den Rahmen des Projektes **ständig neu definiert**. In beiden Fällen handelt es sich bei der Abstimmung um eine Rückmeldung von non-Peers, denen somit innerhalb des Forschungsfeldes der DH eine wachsende Bedeutung beigemessen werden muss (Cavanagh 2012).

Abgesehen von den oben beschriebenen eher automatisch auftretenden Revisionen durch non-Peers, haben digitale GeisteswissenschaftlerInnen auch Möglichkeiten, ihren Forschungsprozess bewusst für andere fachlich interessierte Kollegen zu öffnen. Vor allem in digitalen Infrastrukturen wie dem TextGrid-Lab ist diese Möglichkeit technisch sehr einfach umzusetzen, indem weitere User für ein Projekt freigeschaltet werden können und nur eingeschränkte (Lese-, Kommentier-) Rechte erhalten. Auf diese Weise kann die nicht abgeschlossene Forschung bereits kommentiert und somit einer Revision unterzogen werden. Da immer der Administrator eines Projekts die Gastaccounts freischaltet, hat dieser auch die volle Kontrolle darüber, wem – Peer oder non-Peer – er Zugang zu den Forschungsdaten verschafft und wen er damit zu Kommentaren einlädt.

Eine weitere Möglichkeit zur frühen Öffnung eines Forschungsprojektes bietet das **wissenschaftliche Bloggen**. Ob auf einer eigenen Domain, einem Blogportal wie hypotheses.org oder einem Communityblog wie dem DHd-Blog, das Veröffentlichungsformat eignet sich insbesondere dafür, kurze Zwischenergebnisse zu präsentieren und Kommentare von Lesern zu erbitten. Auf diese Weise können von kurzen Extrakten bis hin zu Artikelserien mehr oder weniger umfangreiche Einblicke in die Forschung gegeben werden.

So kann Peer (oder non-Peer) Review bereits vor Abschluss des Projektes geleistet und in den Prozess einbezogen werden. Ein Mittelweg zwischen dem kollaborativen Review in der Forschungsumgebung und dem Crowd-Review im Blog kann in der Veröffentlichung von Artikeln in sozialen Netzwerken wie Academia.edu gesehen werden. Hier können Artikel gepostet werden, die während eines Forschungsprojektes entstehen. Auch hier gibt es die Möglichkeit, Kommentare

zu erbitten. Da ein solches Netzwerk in Interest Groups gegliedert ist, kann dies als Community-Review bezeichnet werden. Ob diese Community ausschließlich aus Peers oder auch aus non-Peers besteht, liegt in dem Falle daran, wie der /die jeweilige NutzerIn sich seine / ihre Netzwerkumgebung gestaltet.

Umsetzungsempfehlung

Es bleibt insgesamt festzuhalten, dass in einem RDLC der Digital Humanities weder ein fester Zeitpunkt noch ein einzelnes standardisiertes Verfahren für Reviews auszumachen ist. Es ist vielmehr zu **berücksichtigen**, dass ein Projekt in den digitalen Geisteswissenschaften – sei es auf systemimmanente oder vom Forscherenden forcierte Weise – **ständig in Revision** begriffen ist.

Aus diesem Grund ergeben sich eher unscharfe Wünsche für die technischen Umsetzung: So ist ein Verfahren, mit dem anderen WissenschaftlerInnen noch während Durchführung eines Forschungsprojekts Zugang zu den darin verwendeten Daten und darauf stattfindenden Analysen gewährt wird, sicher sehr wünschenswert. Ebenso erfordert ein solches Verfahren eine gewisse Detail- & Implementierungstiefe von Rollen- und Rechtemanagement, welches im Falle von DARIAH-DE durch die dortige Autorisierungs- und Authentifizierungs-Infrastruktur abgedeckt sein sollte.

5.7 Lizensierung

Rechtliche Fragestellungen, sowohl zu Bearbeitungs- und Verwertungsrechten des bereitgestellten Materials als auch zu verwendeten Softwareprodukten und Bibliotheken sollten in jedem Fall frühzeitig diskutiert und Beschlüsse getroffen werden.

Tendenziell ist hier ein mehrstufiges Verfahren sinnvoll und erstrebenswert: Beim Aufbau einer Infrastruktur für DH Forschungszyklen sollten eingangs nur Quellen als Forschungsdaten publiziert werden, deren Verwertungsrechte abgelaufen sind oder deren Rechtstatus anderweitig als **frei nutzbar** gesetzt ist (Bspw. die diversen CC-BY, BSD und DIPP Lizenzen) oder ohnehin als **gemeinfrei** gilt. Daneben lassen sich nicht frei-nutzbare Forschungsdaten in einem System mit entsprechenden Nutzungseinschränkungen[29] belegen, so dass sie nicht (oder nur eingeschränkt) publiziert werden und nur von einer bestimmten, vorher authentifizierten Nutzergruppe eingesehen werden können.

Nach einer gewissen Produktionsreife können weiterhin komplexere Lizensierungsfunktionen für einzelne Teile von Forschungsdatensammlungen angestrebt werden, wie sie z.B. in Großbritannien von JISC und JISC assoziierten Verbänden[30]. Auf diese Weise lassen sich schnell und unkompliziert relativ große Datenmengen, die sich für die geisteswissenschaftliche Analyse eignen, in einem System verwalten, ohne dass von Beginn an alle administrativen Funktionen des Systems implementiert sein müssen. Diese können in einem mehrstufigen Prozess nach und nach eingefügt werden.

[29] Vgl. Kapitel 5.8 Rollen- und Rechtemanagement

[30] Eine einfache aber wirkungsvolle Grafik zur Entscheidungshilfe, welche Lizenz am besten verwendet werden soll, die ursprünglich aus dem OpenEducationalRessources Bereich stammt aber auch für die hier besprochenen geisteswissenschaftlichen Inhalte relevant erscheint: http://www.web2rights.com/OERIPRSupport/charts/LICENSING%20DECISION%20TOOLv4.pdf

Auch für Software und Softwarebibliotheken existieren eine Reihe von Lizenzarten und frei nutzbaren Lizenzverträgen. Für beide Inhaltstypen gilt, dass ihre Verwendung vor allem von ihrer Regelung zur Veränderung von Inhalten / Code durch Dritte, der Nennung der Urheber / des Rechteinhabers sowie dem Verbot bzw. der dezidierten Erlaubnis zur kommerziellen Nachnutzung der Inhalte / des Codes abhängig ist..

In einer Infrastruktur kann es unter Umständen schwierig werden, Inhalte unterschiedlicher Lizenzart miteinander zu verknüpfen, wenn sie aufgrund ihrer Nutzungsrechte nicht miteinander kompatibel sind und sich daher in der gleichzeitigen Nutzung in einem System von Beginn an ausschließen. So kann Softwarecode, der beispielsweise unter der „Microsoft Shared Source Common Language Infrastructure" lizensiert wird, später auch in anderen Softwareprojekten genutzt und verändert aber nicht verkauft werden, was zum Teil erhebliche Einschränkungen mit sich bringt[31]. Eine erste Übersicht über empfohlene Lizenzen für Forschungsdaten wurde in DARIAH-DE erarbeitet[32], ob diese allerdings genügt, bleibt zu diskutieren.

Umsetzungsempfehlung

Aus der geschilderten Herausforderung ergeben sich folgende Anforderungen: Insgesamt wird im Fall der Lizensierung ein mehrstufiger Prozess vorgesehen, bei dem eingangs nur **frei nutzbare und damit unkompliziert zu publizierende** Daten in einer den Research Data Lifecycle unterstützenden Infrastruktur verarbeitet werden können.

Daneben bleiben auch projektinterne Empfehlungen abzuwarten, die ggf. für die technische Umsetzung einer etwaigen Lizenzunterstützung wertvolle Informationen liefern können. An dieser Stelle sei insbesondere auf die DARIAH-DE eigene Initiative zu **Lizensierungsempfehlungen für GeisteswissenschaftlerInnen** sowie Empfehlungen zu Lizenzierungstools verwiesen[33].

5.8 Rollen- und Rechtemanagement

Die diskutierten Fragen nach Rechten von Forschungsdaten lassen sich für alle Phasen des Researchdata-Lifecycles auch durch die Implementation eines Rechte- und Rollenmanagements wenn nicht lösen, so doch abfedern. Ein solches Rechte- und Rollenmanagements würde den Zugriff auf Daten mit unsicherem Status durch eine entsprechende Beschränkung soweit reglementieren, dass kein unberechtigter Zugriff entsteht und alle Rechte eingehalten werden.

Dabei können einerseits die verwendeten Daten selbst unter bestimmten Lizenzen stehen, die eine Nutzung nur für bestimmte Personen, Gruppen und/oder Zeiträume sowohl während des Forschungsprozesses als auch nach Veröffentlichung der Forschungsdaten selbst, ermöglicht – siehe Kapitel 5.7. Andererseits möchten Forschende selbst darüber entscheiden zu welchem Zeitpunkt wer auf die eigenen bzw. die im Projektverbund erhobenen und angereicherten Forschungsdaten zugreifen kann.

Dies gilt im Besonderen für die Phasen, in denen aktiv mit den Daten geforscht wird. Prinzipiell ist davon auszugehen, dass WissenschaftlerInnen die Möglichkeit haben wollen, den Zugriff auf

[31] Vgl. http://en.wikipedia.org/wiki/Shared_source#.

[32] Vgl. https://dev2.dariah.eu/wiki/display/DARIAHDE/Software+Licenses

[33] Vgl. https://de.dariah.eu/lizenzen

Forschungsdaten und der damit bewussten Zurverfügungstellung der Daten an Dritte vor einer eigentlichen Publikation steuern zu können. Um dies zu realisieren, ist ein umfangreiches Rollen- und Rechtemanagement notwendig. Hierbei sind zwei Felder von besonderer Bedeutung: 1. Die Authentifizierung der NutzerInnen und 2. die Autorisierung von NutzerInnen für Daten und Dienste.

Die **Authentifizierung** ist der Prozess, in dem festgestellt werden muss, ob es sich bei den NutzerInnen, die auf die Daten zugreifen möchte, tatsächlich um real existierende Personen (natürliche Einzelperson) handelt. Für solch einen Authentifizierungsprozess wurden bereits eine Vielzahl technischer Verfahren weltweit etabliert, da es sich hierbei um eine grundsätzliche Herausforderung aller webbasierten Dienste handelt. Dass jede(r) NutzerIn sich beispielsweise über seine / ihre **Mailadresse** bei Webdiensten authentifizieren muss, um für die Nutzung bestimmter Angebote frei geschaltet zu werden, ist mittlerweile zentraler Bestandteil des digitalen Alltagslebens. Prinzipiell kann es sich kein Serviceanbieter mehr leisten, nicht zu wissen, um wen es sich bei der Person genau handelt, die bestimmte Services nutzt oder Daten bereit stellt – auch um im Falle von Rechtsverletzungen Rechtsicherheit zu erhalten und anbieten zu können. Letztlich ist also eine Authentifizierungsschicht nur die technische Voraussetzung dafür, dass einzelne WissenschaftlerInnen die technische Möglichkeit erhalten, Daten und Dienste zu nutzen und mit einem Online-Dienst zu interagieren.

Ein **Autorisierungsdienst** geht über eine reine Nutzerkontenverwaltung hinaus, da über einen solchen auch der Zugriff auf unterschiedliche Angebote innerhalb eines Webdienstes **modular** fest gelegt werden kann. Die Wichtigkeit einer Autorisierungsschicht kann exemplarisch anhand der Datennutzung des Projektes European Holocaust Research Infrastructure (EHRI) aufgezeigt werden. Die in diesem Projekt zur Verfügung gestellten Daten sind hochsensibel, da sie aus dem Kontext des Holocaust stammen ggf. personenbezogene Informationen enthalten, sie unterliegen außerdem auch explizit **nutzungsspezifischen Zugriffsbeschränkungen**, da die Daten noch urheber-, verwertungs- oder persönlichkeitsrechtlichem Schutz unterliegen.

Kurzum: In einer hinreichend komplexen Forschungsinfrastruktur besteht der Bedarf nach einem Rollen- und Rechtemanagement, um auch angemeldeten (authentifizierten) NutzerInnen den Zugriff auf bestimmte Daten und Sammlungen zu ermöglichen oder explizit zu verwehren. Aus diesen Gründen wird kollaboratives Arbeiten überhaupt erst durch eine Authentifizierungs- und Autorisierungsschicht (im DARIAH-DE Kontext kurz: **AAI**) möglich.

Umsetzungsempfehlung

Im Rahmen von DARIAH-DE wurde in den vergangenen Jahren eine Autorisierungs- und Authentifizierungs-Infrastruktur aufgebaut, die die Vergabe von unterschiedlichen Rollen und Rechten an verschiedene Nutzergruppen – gerade auch für Nutzergruppen aus dem DARIAH-EU Kontext – ermöglicht.[34] Im Fokus des Research Data Lifecycle und auch im Rahmen eines ggf. festzulegenden Datenmodells (siehe Kapitel 6.3) wird zu klären sein, inwieweit die bereits durchgeführten und zukünftig geplanten Arbeiten an der DARIAH-DE AAI ausreichen, um die hieraus spezifischen Anforderungen zu erfüllen.

[34] Eine technische Beschreibung der DARIAH-DE AAI findet sich u.a. auf der DARIAH-DE Webseite: https://de.dariah.eu/aai

6. Entwurf für ein Referenzmodell

In diesem Kapitel soll ein für die digitalen Geisteswissenschaften prototypischer Arbeitsfluss diskutiert und erläutert werden:

Generell findet ein Forschungsvorhaben in einem spezifischen zeitlichen Rahmen mit spezifischen personellen und materiellen Ressourcen mit einem bestimmten Forschungs- / Erkenntnisziel statt. Dabei stützt sich ein geisteswissenschaftliches Forschungsvorhaben auf eine bestimmte Quellensituation, welche zu Beginn mehr oder minder bekannt ist. Ein Forschungsvorhaben kann außerdem in unterschiedliche, relativ generische Tätigkeiten unterteilt, werden, welche wiederum ihrerseits in spezifisch geisteswissenschaftliche Aktivitäten untergliedert werden können. Ein erster grober Überblick über einen in sich abgeschlossenen Forschungsprozess bietet die folgende Darstellung.

Dabei handelt es sich um eine Mischform aus datengetriebener und tätigkeitsorientierter Darstellung, d.h. es wird versucht, alle Tätigkeiten nach ihrem Veränderungspotential auf den zu Beginn des Forschungsvorhabens zu Grunde liegenden Daten und dazugehörigen Metadaten, darzustellen.

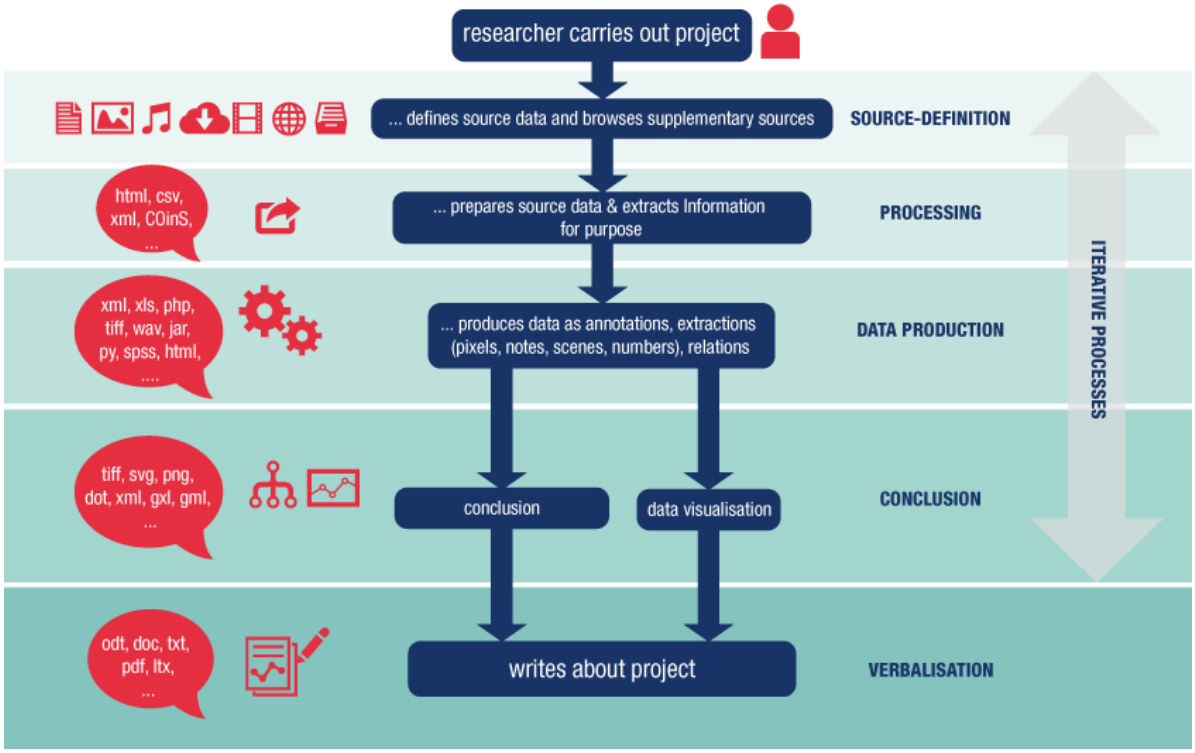

Abb. 6: Die klassische Arbeitsfolge in den (digitalen) Geisteswissenschaften

Dieser klassische geisteswissenschaftliche Arbeitsfluss birgt schon in sich selbst nicht-lineare Komponenten, wie ggf. das selbstkorrigierende Moment, wenn im Schritt der Prozessierung oder Datenproduktion festgestellt wird, dass das zugrunde liegende Quellenmaterial nicht ausreicht oder nicht hinreichend spezifisch ist und daher ggf. eingegrenzt oder erweitert werden muss um dann erneut die Prozessierung zu durchlaufen.

6.1 Ein generischer Workflow

Das soeben vorgestellte Modell für einen geisteswissenschaftlichen Arbeitsablauf lässt sich zu einem zyklischen Modell ausbauen, in welchem die Arbeitsschritte besser dokumentiert werden müssen und neben Bereitstellung einer entsprechenden textbasierten Publikation auch die zugrunde liegenden Daten selbst sowie Zwischenergebnisse zur Nachnutzung bereitgestellt werden sollten.

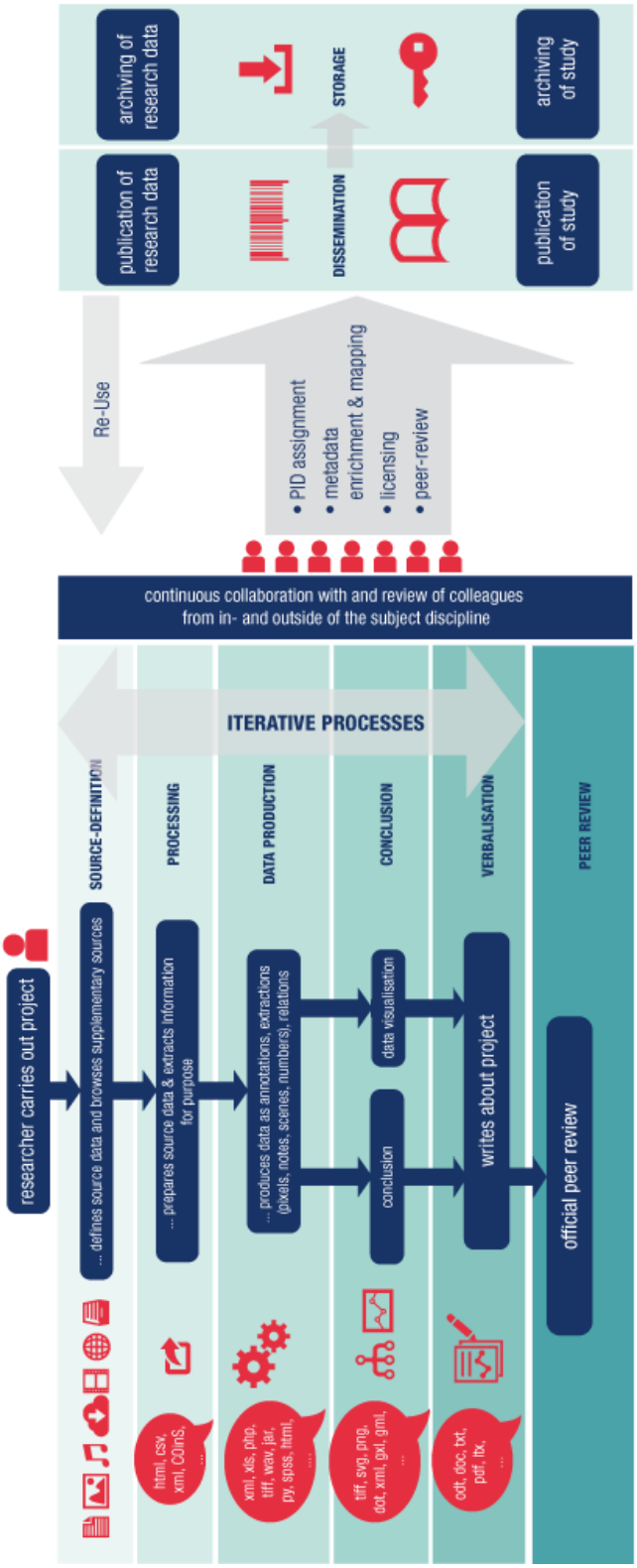

Abb. 7: Der Research Data LifeCycle unter Einbeziehung von Publikation, Archivierung und Nachnutzung

Abbildung 7 stellt dabei die Erweiterung eines traditionellen Forschungsworkflows dar, so dass neue Forschungsprozesse auf vorhergehenden Arbeiten aufbauen können und Zwischenschritte publiziert und archiviert werden können. Dazu gehört die Bereitstellung von Roh- und Enddaten, sowie die Rekonstruierbarkeit von Forschungsprozessen durch eine vollständige Dokumentation. Insgesamt können hier viele zyklische Prozesse beobachtet werden, welche die lineare Darstellung

von Forschungsaktivitäten schwierig gestalten und daher in einem zyklischen Verständnis besser untergebracht sind.

6.2 Aktivitäten in einem Basis-Forschungsdatenzyklus

Es lassen sich einige generische Aktivitäten im Forschungsdatenzyklus identifizieren, die – jeweils begleitet von einigen technischen Operationen – eine relativ umfangreiche Liste ergeben. Die folgende Tabelle listet diese Tätigkeiten auf und stellt in Spalte 3 und 4 Überlegungen zur Formalisierbarkeit der genannten Tätigkeit an:

Tabelle 4: Aktivitäten in einem generischen Forschungsdatenzklus

Schritt	Aktivität	Formalisierbar Ja/Nein/Vielleicht	Notiz
1	Formulierung Forschungsfrage, Benennung von Methoden	Vielleicht	Anhand einer Basis-Fragestellung lassen sich ggf. anschließende Tätigkeiten oder verwendbare Komponenten formalisieren und maschinell identifizieren.
2	Auswahl Forschungsdaten	Vielleicht	Die Wahl der Forschungsdaten ist eng an die Formalisierung der Forschungsfrage geknüpft: Wenn die Forschungsfrage maschinell auswertbar formalisiert werden kann, so können auch die zu verwendenden Forschungsdaten formell beschrieben werden.
3	Vorbereitung / Verwendung von Tools	Vielleicht	Ist ebenso an die Forschungsfrage und an die Homogenität der vorgefundenen Forschungsdaten geknüpft.
4	Generierung von (Zwischen-) Ergebnissen/ Verwendung von Tools	Ja	Wenn alle Vorbereitungen getroffen und spezifiziert sind: Ja
5	Visualisierung	Ja	"
6	Beschreibung der veröffentlichungs-würdigen Ergebnisse und Erkenntnisse	Nein	Kann sehr begrenzt formalisiert aber niemals maschinell ausführbar beschrieben werden.
7	Kuration / Vorbereitung der Archivierung	Ja	Im Rahmen einer einfachen Langzeitarchivierungsstrategie (Bitstreampreservation, ggf. Metadaten-Extraktion und Migration) lässt sich formalisiert ein „Preservation Plan" erstellen.

Aus der vorangestellten Tabelle und der Abbildung 7 ergibt sich, dass an mehreren Stellen in Research Data LifeCycle immer wieder die gleichen Funktionen fällig werden. Dies sind:

- Identifiervergabe
- Metadatenanreicherung /-abgleich
- Lizensierung
- Publikation
- Langzeitarchivierung
- Peer-Review

Diese – eher infrastrukturellen – Funktionen sollten immer dann aufgerufen werden, wenn neuen Zwischenergebnisse oder neue Beziehungen zwischen Daten produziert und Ergebnisse definiert werden. In der vorliegenden Tabelle sollten solcherlei generische Funktionen also bei allen Schritten bis auf Schritt 3 durchgeführt werden.

Die genannten infrastrukturen Funktionen dienen vor allem der konsistenten Datenhaltung und der Gewährleistung von Wiederauffindbarkeit und Referenzierbarkeit. Außerdem dienen Funktionen wie Publikation und Peer-Review dem kollaborativen Austausch, da nur publizierte Daten von der Community begutachtet werden können. Daneben sind Übersichten mit unterstützenden Metadatenstandards oder aber mit einzelnen Feldern, die im Falle einer Metadaten-Transformation erhalten bleiben sollen, notwendig damit eine vollständigen Spezifikation eines Research Data LifeCycle möglich wird. Hier werden maßgebliche Empfehlungen direkt in die Weiterentwicklung des DARIAH Repositories sowie der Schema- und Crosswalk-Registry einfließen. Darüber hinaus besteht die langfristige Perspektive zur Formalisierung solcher Überlegungen in der Entwicklung eines Metadatenmodells, welches die Bereitstellung, Arbeit und Archivierung von Daten in einer Infrastruktur speziell für die Geisteswissenschaften abbilden kann.

Eine Liste mit empfohlenen Dateiformaten, Metadatenformaten oder Tools wird hingegen direkt als Ergebnis der Diskussionen publiziert und direkt im Web[35] angeboten.

6.3 Datenmodell

Die vorangegangenen Überlegungen ergeben immer stärker den Bedarf nach einem Datenmodell, welches für ein Forschungsprojekt und seine dazugehörigen Datensammlung alle Informationen abbildet, die bei der a) Verwaltung, b) Auffindbarkeit, c) Referenzierbarkeit, d) Nutzungsgeschichte etc. hilfreich sein können. Es ist sogar denkbar, dass durch ein entsprechendes Datenmodell eine so starke Formalisierung erreicht werden kann, dass selbst die Beantwortung von Forschungsfragen in Teilen automatisiert werden kann.

Zunächst ist es aber die Aufgabe eines solchen Datenmodells, alle in einem Kontext und zu einer Fragestellung erhobenen und erstellten Daten und Metadaten zusammen mit deskriptiven und administrativen Informationen, sowie ihren Prüfsummen und Identifiern in einer sinnvollen Sturktur abzubilden und einheitlich – möglichst generisch und nachnutzbar – in einer Infrastruktur zu administrierbar zu sein. Insbesondere für folgende Fragestellungen aus dem Bereich der Versionierung muss konsistent eine Lösung gefunden werden:

- Welche **Arbeitsschritte** in der DARIAH-DE Infrastruktur erfordern die Markierung einer Version als neue Version einer anderen bereits in anderem Kontext verwendeten Datei?

[35] Vgl: https://dev2.dariah.eu/wiki/x/cg9FAg

- Reicht der Download und erneuter Upload einer Datei in einer anderen „Collection" aus, um diese als **neue Version** zu definieren?
- Wie wird bei externen Dateien, in deren Metadaten nicht die **Provinienz** durch einen eindeutigen Identifier angegeben ist, überprüft, ob es sich dabei um eine Datei handelt, die bereits in einem anderen Forschungskontext verwendet worden ist?

Analog zu der Verwendung des OAIS Modells für den Bereich der Langzeitarchivierung *könnte* man alle Aspekte eines Forschungs-Datums als entsprechende Informationspakete in einem RDLC-Objekt anfügen. Hier bedarf es sowohl des bereits erwähnten Datenmodells als auch der Konzeption bzw. Nutzung entsprechender Metadatenfelder und -standards. Es müssen außerdem sowohl Prozesse als auch Personen Rollen zugewiesen bekommen, die Ihnen erlauben einzelne Metadatenfelder zu editieren. Das angestrebte Datenmodell stellt so die eigentliche Manifestation der intellektuellen **Workflowplanung** dar.

Wie bereits in Kapitel 5.3 Metadaten-Anreicherung angerissen, wird in einem Datenmodell sowohl beschrieben, mit welcher Detailtiefe und welcher Komplexität Dateien zu Objekten gebündelt werden, als auch welche Informationen zur Implementation des Forschungsdatenzyklus notwendig und / oder optional sind.

Denkbar als Metadatenstandard zur Beschreibung eins Forschungsdatenzyklus auf Objektebene sind hier sowohl der W3C PROV Standard (W3C 2013b) zur Bestimmung von Provenienz eines Datenobjekts als auch der PREMIS Standard der Library of Congress (PREMIS 2012) der sich großer Verbreitung und Bekanntheit im Bereich der Langzeitarchivierung erfreut. Ein weiterer aus den Sozialwissenschaften stammender Ansatz ist die Arbeit der Data Documentation Intitiative (Hoyle 2013), welche aktuell Version 3 ihres Datenmodells veröffentlicht hat. Dieser bezieht sich allerdings dezidiert auf sozialwissenschaftliche Forschungsdaten, die nicht nur eher sozialwissenschaftlicher Terminologie unterliegen sondern auch eher sozialwissenschaftlicher „Natur" sind. Die Rede ist hier von Stichproben, Studiendaten, Verfahren zur statistischen Auswertung u.ä.

Umsetzungsempfehlung

In der kommenden Arbeitsphase der AG gilt es, vorhandene Datenmodelle zu evaluieren und gegebenenfalls auch **Modellierungsmöglichkeiten** für geisteswissenschaftliche Workflows in Metadatenmodellen zu erforschen.

Ein Fokus sollte hier auf der **Implementierbarkeit** und **Detailtiefe** der vorgefundenen Standards liegen. Aber auch bereits an anderem Ort implementierte Standards, d.h. die Verbreitung und Erfahrungen andere Infrastrukturen mit solchen Technologien sollten eine zentrale Rolle bei der Wahl bzw. der Anpassung eines solchen Datenmodells gelten.

7. Fazit

Das vorliegende Dokument stellt einen großen Teil der notwendigen Überlegungen vor, welche bei der Planung und Implementation eines Forschungsdatenzyklus berücksichtigt werden sollten. Insbesondere liegt der Fokus auf der Empfehlung von Kriterien zur selbständigen Auswahl von Dateiformaten, Metadatenstandards, Tools und Langzeitarchivierungsstrategien für GeisteswissenschaftlerInnen.

Darüber hinaus konzentriert sich das Dokument auf die Beschreibung der Funktionalität und der Abläufe von Forschungsprozessen in den Geisteswissenschaften aus infrastruktureller Perspektive. Hier ist das Ziel eine möglichst generische Beschreibung von geisteswissenschaftlicher Forschung zu erreichen, welche durch das vorherige Kapitels abgedeckt wird.

Perspektivisch ist geplant im kommenden Jahr, Datenmodelle zur Abbildung von Provenienz und Workflows soweit zu evaluieren und ggf. anzupassen, dass auch hier eine praktikable Empfehlung für ein Modell oder zumindest ein Subset von Schemas ausgesprochen werden kann.

Hinweis: Die hier diskutierten Empfehlungen für Ressourcen in der Langzeitarchivierung wurden als tabellarische Übersichten zu Dateiformaten, Metadatenstandards, Lizenzen im Web zum Zwecke der Nachnutzung durch andere WissenschaftlerInnen veröffentlicht und werden dort redaktionell weiter betreut: https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=38080370.

8. Quellenverzeichnis

Hinweis: Der letzte Zugriff auf Webressourcen erfolgte am 09. April 2015.

[ADUK 2009] Archaeology Dataservice, UK. *Creating Texts and Documents*. Section 2: Creating Texts and Documents. 2009. http://guides.archaeologydataservice.ac.uk/g2gp/TextDocs_2.

[Andorfer 2015] Andorfer, Peter. *Forschen und Forschungsdaten in den Geisteswissenschaften*. DARIAH-DE Working Papers. Göttingen: DARIAH-DE, 2015.

[Archivematica 2014] Archivematica. *Format policies*. 2014. https://www.archivematica.org/mediawiki/index.php?title=Format_policies&oldid=9686

[Aurast 2014] Aurast, Anna; Beer, Nikolaos; Held, Marcus; Herold, Kristin; Kolbmann, Wibke; Kollatz, Thomas; Richts, Kristina; Schmunk, Stefan; Veit, Joachim: *Fachspezifische Empfehlungen für Daten und Metadaten*. Kapitel 6. Schlussfolgerungen. Göttingen: DARIAH-DE, 2014. https://dev2.dariah.eu/wiki/pages/viewpage.action?pageId=20058160

[Behrens et al. 2010] Behrens, Julia; Fischer, Lars; Minks, Karl-Heinz; Rösler, Lena. *Die Internationale Positionierung Der Geisteswissenschaften in Deutschland. Eine Empirische Untersuchung*. HIS:Projektbericht, 2010. https://www.bmbf.de/pubRD/internationale_positionierung_geisteswissenschaften.pdf.

[Boonstra et al. 2004] Boonstra, Onno; Breure, Lena; Doorn, Peter. *Past, Present and Future of Historical Information Science. Netherlands institute for scientific information royal netherlands academy of arts and sciences*. 2004. http://www.ahc.ac.uk/docs/pastpresentfuture.pdf.

[Cavanagh 2012] Cavanagh, Sheila. *Living in a Digital World: Rethinking Peer Review, Collaboration, and Open Access*. Journal of Digital Humanities, 2012. http://journalofdigitalhumanities.org/1-4/living-in-a-digital-world-by-sheila-cavanagh/.

[DCC 2015] Digital Curation Centre. *Metadata Standards*. Social Science & Humanities, 2015. http://www.dcc.ac.uk/resources/subject-areas/social-science-humanities.

[DDB 2012] Deutsche Digitale Bibliothek. *Glossar zu den technischen Spezifikationen der Deutschen Digitalen Bibliothek*. 2012. https://www.deutsche-digitale-bibliothek.de/static/de/sc_documents/Anlage_TS_Glossar.pdf.

[DDI 2011] Gregory, Arofan, und Open Data Foundation. *The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes*. 2011. http://odaf.org/papers/DDI_Intro_forNSIs.pdf.

[DFG 2013] DFG. *Vorschläge Zur Sicherung Guter Wissenschaftlicher Praxis*. Ergänzte Auflage. Bonn, 2013. http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf.

[DPC 2008] Digital Preservation Coalition in collaboration. *Preservation Management of Digital Materials: The Handbook*. London, 2008.

http://www.dpconline.org/component/docman/doc_download/299-digital-preservation-handbook.

[DP4Lib 2014] DP4Lib Engaged. *Kernset technischer Metadaten für die Langzeitarchivierung digitaler Objekte.* Frankfurt: Deutsche Nationalbibliothek, 2014. https://wiki.dnb.de/download/attachments/31524273/Kernset-techMd_v1.1.pdf.

[FLVC 2012a] The Florida Virtual Campus. *Table of FDA-Supported File Formats*. FLVC – State University Library, 2012. http://fclaweb.fcla.edu/node/795.

[Gradmann, Meister 2008] Gradmann, Stefan; Meister, Jan Christoph. *Digital document and interpretation: re-thinking 'text' and scholarship in electronic settings*. 2008. Poiesis & Praxis, S.139-153. DOI 10.1007/s10202-007-0042-y

[Haynes 2004] Haynes, David. *Metadata for Information Management and Retrieval.* Become an Expert ... London: Facet Publ., 2004.

[Hoyle 2013] Hoyle, Larry; Corti, Louise; Gregory, Arofan; Martinez, Agustina; Wackerow, Joachim; Alvar, Eirik; Betancort Cabrera, Noemi; Gallagher, Damien; Gebel, Tobias; Hautamaki, Jani; Kuula, Arja; McEachern, Steve; Zuell, Cornelia. *A Qualitative Data Model for DDI. Data Documentation Inititative.* 2013. http://www.ddialliance.org/system/files/AQualitativeDataModelForDDI.pdf.

[IANUS 2012] IANUS. *MISSION STATEMENT / Leitbild.* 2012. http://www.paradigm.ac.uk/workbook/preservation-strategies/file-preservation.html.

[IANUS 2014] IANUS. *IT-Empfehlungen. Für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften. Dateiformate | IT-Empfehlungen*. 2014. http://www.ianus-fdz.de/it-empfehlungen/?q=node/34.

[Klimpel, Keiper 2013] Klimpel, Paul; Keiper, Jürgen. *Was Bleibt? Nachhaltigkeit Der Kultur in Der Digitalen Welt.* Berlin: Internet & Gesellschaft Collaboratory e. V., 2013. http://files.dnb.de/nestor/weitere/collab_was_bleibt.pdf.

[Kuipers, van der Hoeven 2009] Kuipers, Tom; van der Hoeven, Jeffrey. *Insight into Digital Preservation of Research Output in Europe (PARSE).* Survey Report, 2009. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf.

[Lauer 2013] Lauer, Gerhard. *Die digitale Vermessung der Kultur. Geisteswissenschaften als Digital Humanities.* In Big Data. Das neue Versprechen der Allwissenheit, 99–116. Berlin, 2013. http://gerhardlauer.de/files/8913/8217/2087/lauer_big-data.pdf.

[Lampe et al. 2010] Lampe, Karl-Heinz; Krause, Siegfried; Doerr, Martin. *Definition des CIDOC Conceptual Reference Model*. Version 5.0.1. ICOM Deutschland – Beiträge zur Museologie. ICOM Deutschland, 2010. http://www.icom-deutschland.de/client/media/380/cidoccrm_end.pdf.

[Lazorchak 2011] Lazorchak, Butch. *Digital Preservation, Digital Curation, Digital Stewardship: What's in (Some) Names?* The Signal, Digital Preservation, 2011. http://blogs.loc.gov/digitalpreservation/2011/08/digital-preservation-digital-curation-digital-stewardship-what%E2%80%99s-in-some-names/.

[LeFurgy 2013] LeFurgy, Bill. *Is JPEG-2000 a Preservation Risk?* The Signal: Digital Preservation, 2013. http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/.

[Lormant et al. 2005] Lormant, Nicolas; Huc, Claude; Boucon, Danièle; Miquel, Christine. *How to Evaluate the Ability of a File Format to Ensure Long-Term Preservation for Digital Information?* Edinburgh: ukoln, 2005. http://www.ukoln.ac.uk/events/pv-2005/pv-2005-final-papers/003.pdf.

[metadatadeluxe 2015] PB Metadata Deluxe. *VRA Embedded Metadata Working Group*. 2015. http://metadatadeluxe.pbworks.com/w/page/20792294/VRA%20Embedded%20Metadata%20Working%20Group.

[Miller 2011] Miller, Steven J. Metadata for Digital Collections: *A How-to-Do-It Manual. How-to-Do-It Manuals for Libraries*. New York, NY 2011: Neal-Schuman Publ, 1989- 179. New York.

[NDIIPP 2013] National Digital Information Infrastructure and Preservation Program (NDIIPP). *Sustainability of Digital Formats*. Planning for Library of Congress Collections, 2013. http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

[Neuroth 2010] Neuroth, Heike; Huth, Karsten; Oßwald, Achim; Scheffel, Regine; Strathmann, Stefan (Hg.). *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Version 2.3. nestor, 2010. http://www.nestor.sub.uni-goettingen.de/handbuch/index.php.

[NISO 2004] NISO (Hg.). *Understanding Metadata*. 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

[OAIS 2012] The Consultative Committee for Space Data Systems (CCSDS). *REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS)*. MAGENTA BOOK. Washington, DC, USA, 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

[Paradigm 2008] Paradigm. *File Formats. Selecting File Formats for Preservation.* Paradigm, 2008. http://www.paradigm.ac.uk/workbook/preservation-strategies/file-preservation.html.

[Paradis et al. 2013] Paradis, Jim; Fendt, Kurt; Kelley, Wyn; Folsom, Jamie; Pankow, Julia; Graham, Elyse; Subbaraj, Lakshmi. *Whitepaper: 'Annotation Studio: Bringing a Time-Honored Learning Practice into the Digital Age.'* Comparative Studies | Media Writing, 2013. http://cmsw.mit.edu/annotation-studio-whitepaper.

[PREMIS 2012] PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata. V.2.2.* Library of Congress, 2012. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

[Reiche et al. 2014] Reiche, Ruth; Becker, Rainer; Bender, Michael; Schmunk, Stefan; Schöch, Christof. *Verfahren der Digital Humanities in den Geistes- und Kulturwissenschaften*. DARIAH-DE Working Papers. Göttingen: DARIAH-DE, 2014. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2014-4.pdf.

[ReMind 2005] ReMind, HAW Hamburg. *Metadaten.* ReMind: Metadaten, 2005. http://www.bui.haw-hamburg.de/pers/ulrike.spree/remind/metadaten.htm.

[Riley 2010] Riley, Jenn. *Seeing Standards: A Visualization of the Metadata Universe.* Indiana University Libraries Digital Projects & Services, 2010. http://www.dlib.indiana.edu/~jenlrile/metadatamap/.

[Sahle, Kronenwett 2013] Sahle, Patrick; Kronenwett, Simone. *Jenseits der Daten. Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center for the Humanities'*. Libreas 23 (2013): 76–96. http://edoc.hu-berlin.de/libreas/23/sahle-patrick-1/PDF/sahle.pdf.

[Sanderson, Van de Sompel 2010] Sanderson, Robert; Van de Sompel, Herbert. *Making Web Annotations Persistent over Time.* In JCDL '10 Proceedings of the 10th Annual Joint Conference on Digital Libraries, 1–10. JCDL '10. New York: ACM - Digital Library, 2010. doi:10.1145/1816123.1816125.

[Schmidt 2012] Schmidt, Desmond. *The Role of Markup in the Digital Humanities.* In Historical Social Research / Historische Sozialforschung, 37:, No. 3 (141):, Controversies around the Digital Humanities:125–46. Köln: GESIS - Leibniz-Institute for the Social Sciences, Center for Historical Social Research, 2012. http://www.jstor.org/stable/41636601.

[Stiller et al. 2015] Stiller, Julianne; Thoden, Klaus; Leganovic, Oona; Heise, Christian; Höckendorff, Mareike; Gnadt, Timo. *Nutzungsverhalten in den Digital Humanities*. Göttingen: DARIAH-DE, 2015. https://dev2.dariah.eu/wiki/download/attachments/14651583/Report1.2.1-final3.pdf?version=1&modificationDate=1426154224304&api=v2.

[Succeed 2014] Succeed. D4.1. *Reccomendations for Metadata and Data Formats for Online Availability and Long-Term Preservation*. Succeed. Alicante: Universidad de Alicante, 2014. http://www.succeed-project.eu/sites/default/files/deliverables/Succeed_600555_WP4_D4.1_RecommendationsOnFormatsAndStandards_v1.1.pdf.

[Tonkin 2008] Tonkin, Emma. *Persistent Identifiers: Considering the Options.* Ariadne, no. 56 (2008). http://www.ariadne.ac.uk/issue56/tonkin.

[UMN 2015] University of Minnesota Libraries. *Data Documentation and Metadata.* 2015. https://www.lib.umn.edu/datamanagement/metadata#meta.

[Unsworth 2000] Unsworth, John. *Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?.* 2000. http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

[W3C 2013a] W3C. *Open Annotation Data Model.* 2013. http://www.openannotation.org/spec/core/.

[W3C 2013b] W3C. *PROV Model Primer.* 2013. http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/.

[TEI 2013] Text Encoding Initiative. *Text Encoding Initiative.* 2013. http://www.tei-c.org/index.xml.

[Wissenschaftsrat 2006] Wissenschaftsrat. *Empfehlungen zur Entwicklung und Förderung der Geisteswissenschaften in Deutschland*. Berlin: Wissenschaftsrat, 2006. http://www.wissenschaftsrat.de/download/archiv/7068-06.pdf.