Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 7
Geisteswissenschaften

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

nestor

Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Förderkennzeichen: 01 DL 001 B

Universitätsverlag Göttingen

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

D-GRID GmbH

Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme
hg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig
im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Lektorat) und Sonja Neweling (Redaktion) sowie Martina Kerzel und Lajos Herpay (Gestaltung und Montage) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte des Buches stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/bestandsaufnahme/)
sowie über den Göttinger Universitätskatalog
(http://www.sub.uni-goettingen.de) zur Verfügung.
Die URN lautet: http://nbn-resolving.de/urn:nbn:de:0008-2012031401.

Die digitale Version dieses Werkes steht unter einer Creative Commons Namensnennung-Nicht-kommerziell-Weitergabe unter gleichen Bedingungen 3.0 Unported Lizenz.

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2012
http://www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen.

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide
Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-86488-008-7

7 Geisteswissenschaften

Wolfgang Pempe

7.1 Einführung in den Forschungsbereich

Das räumliche, zeitliche, thematische und methodologische Spektrum der Disziplinen, die unter dem Sammelbegriff der Geisteswissenschaften zusammengefasst werden, ist so vielfältig[1], dass eine klare Eingrenzung des Forschungsbereichs schwerfällt, zumal die These der „zwei Kulturen"[2] und damit eine deutliche Trennung zwischen Geistes- und Naturwissenschaften längst nicht mehr haltbar ist.[3] In einigen Teilbereichen verschwimmen die Grenzen – und immer häufiger lässt sich eine, u.a. durch „digitale" Methoden beförderte, interdisziplinäre Zusammenarbeit beobachten, z.B. wenn es um die Altersbestimmung oder Analyse von Objekten oder die Rekonstruktion mit dem menschlichen Auge nicht (mehr) wahrnehmbarer Schriftzeugnisse geht. Aber auch naturwissenschaftliche Disziplinen bedienen sich geisteswissenschaftlicher Methoden, z.B. im Zusammenspiel von Linguistik und Bioinformatik[4] oder in den Neurowissenschaften[5]. Daneben ergeben sich durch die Anwendung aus Sicht mancher traditioneller Disziplinen neuartiger Methoden, wie z.B. quantitativer Analysen, neue Forschungsfelder. Bezogen auf den Bereich der Forschungsdaten bedeutet

---

1 Als ein Beispiel für viele sei hier die Definition des Wissenschaftsrats angeführt: „Philosophie, die Sprach- und Literaturwissenschaften, die Geschichtswissenschaften, die Regionalstudien, die Religionswissenschaften, die Ethnologie sowie die Medien-, Kunst-, Theater- und Musikwissenschaften" (Wissenschaftsrat (2006), S. 17).

2 Snow (1987).

3 Vgl. hierzu Behrens et al. (2010), Kap. 1.

4 „Wechselwirkungen zwischen linguistischen und bioinformatischen Verfahren, Methoden und Algorithmen: Modellierung und Abbildung von Varianz in Sprache und Genomen" (BMBF (2011g)).

5 Vgl. z.B. das Projekt Brain/Concept/Writing (humtec (2011)) oder etwas allgemeiner WissensLogs (2011).

dies, dass die ohnehin vorhandene, der Breite des Forschungsbereichs geschuldete Heterogenität, in Zukunft sowohl aufgrund der Erweiterung des Spektrums der Methoden als auch der anfallenden Daten noch erhöht wird – letztere z.B. aus naturwissenschaftlichen Analysen, für deren Bandbreite und Komplexität beispielhaft das im Kapitel zur Klimaforschung Gesagte gelten mag. Dies erfordert entsprechend ein modular aufgebautes System für die Langzeitarchivierung von Forschungsdaten, wenn überhaupt ein zentrales System für den gesamten Bereich der Geisteswissenschaften erreichbar oder sinnvoll ist.

In diesem Kapitel wird beispielhaft das Projekt TextGrid[6] betrachtet, das neben einer Virtuellen Forschungsumgebung eine eng damit verbundene, aber auch isoliert nutzbare Langzeitarchivierungskomponente entwickelt. TextGrid wird im Rahmen der D-Grid-Initiative vom BMBF[7] gefördert. Wie der Name bereits impliziert, lag der anfängliche Fokus – u.a. bedingt durch die Zusammensetzung der Partner[8] – auf den Textwissenschaften unter besonderer Berücksichtigung der Anforderungen der deutschen Philologie, Literatur- und Sprachwissenschaft. Die Priorisierung der im Rahmen des Projekts zu entwickelnden fachspezifischen Werkzeuge war deshalb zunächst davon bestimmt, die Erstellung von und den Umgang mit kritischen Editionen, Wörterbüchern und linguistischen Korpora zu unterstützen[9]. Gleichwohl war TextGrid von Anfang an darauf bedacht, eine Plattform anzubieten, die perspektivisch auch für andere geisteswissenschaftliche Disziplinen nutzbar ist. Die im Rahmen von TextGrid errichtete *Virtuelle Forschungsumgebung*[10] versucht der Heterogenität des fachlichen Umfelds der Geisteswissenschaften insofern Rechnung zu tragen, als sie auf eine weitgehend generische, offene, modulare und erweiterbare technische Infrastruktur aufbaut. Dies gilt gleichermaßen für die eingesetzten Werkzeuge wie für die unterstützten Datenmodelle. Auf diese Weise sollen neue Projekte und Nutzergruppen in die Lage versetzt werden, spezifische

6 Vgl. TextGrid (2011c).

7 Förderkennzeichen: 07TG01A–H (erste Projektphase 2006–2009), 01UG0901A–J (zweite Phase 2009–2012).

8 Die Liste der aktuellen Projektpartner findet sich unter TextGrid (2011a).

9 Vgl. hierzu TextGrid (2006).

10 Für eine aktuelle Definition des Begriffs vgl. Allianz-Initiative (2011a). Eine ausführlichere Diskussion des Begriffs „Virtuelle Forschungsumgebung“ bzw. „Virtual Research Environment“ findet sich bei Carusi; Reimer (2010), S. 13ff.

Anpassungen vorzunehmen und eigene Werkzeuge zu integrieren, um die jeweiligen Forschungsfragen und -prozesse zu unterstützen. In der aktuellen, zweiten Förderphase (2009–2012) wurde die fachwissenschaftliche Basis durch die Aufnahme von Partnern aus weiteren Disziplinen erweitert, so dass nun entsprechende Module für Teilbereiche der Musikwissenschaft, Kunstgeschichte, mittellateinischen Philologie (Glossografie) und der Judaistik[11] entwickelt werden. Die von Anfang an intendierte Möglichkeit der Anpassbarkeit an die Bedürfnisse weiterer geisteswissenschaftlicher Disziplinen hat dazu geführt, dass TextGrid auf politischer Ebene häufig als System „für die Geisteswissenschaften" diskutiert wird.

Das Projekt ist fachwissenschaftlich getrieben, d.h. die Projektpartner vermitteln die Anforderungen und Interessen ihrer jeweiligen Communities und pflegen beständig den Kontakt zur potentiellen fachwissenschaftlichen Nutzergruppe im Rahmen von Workshops, Schulungen und Konferenzen. Somit kann die technische Entwicklung auf spezielle Anforderungen der Nutzergemeinschaft eingehen und sich an deren Bedürfnissen orientieren.

Die von TextGrid entwickelte Virtuelle Forschungsumgebung besteht aus zwei Komponenten: Das *TextGrid Laboratory* oder *TextGrid Lab* bietet in einer einheitlich gestalteten Arbeitsumgebung (einer eclipse-basierten[12] Software) Zugang zu allen Werkzeugen, Services und zu den Daten im *TextGrid Repository*. Dieses Repository dient einerseits zur Ablage der mit den Werkzeugen des *TextGrid Laboratory* erzeugten bzw. zur weiteren Bearbeitung importierten Daten. Andererseits handelt es sich um ein digitales Langzeitarchiv für Forschungsdaten, das auch über ein Webportal sowie verschiedene offene Schnittstellen[13] nutzbar ist. Durch die Kombination von generischen Werkzeugen in einer offenen, nachhaltigen Umgebung mit einem digitalen Langzeitarchiv für Forschungsdaten, ermöglicht es TextGrid, die Nutzer von Beginn des Forschungsprozesses an so zu un-

11 Zur Liste der aktuellen Projektpartner siehe Fußnote 8; zur „Freien Fachsäule" (AP 6.7), vertreten durch das Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte (Duisburg), vgl. TextGrid (2008b), S. 42.

12 Etwas genauer: *Eclipse Rich Client Platform* (vgl. eclipse (2011)).

13 Neben dem *TextGrid Laboratory* wird mit Version 1.0 (Juni 2011) auch ein Online-Portal als weitere Nutzerschnittstelle zur Verfügung stehen. Darüber hinaus sind alle Middleware-Komponenten des *TextGrid Repository* über WebService-Schnittstellen (SOAP und/oder REST) adressierbar, vgl. TextGrid (2008a).

terstützen, dass – ganz im Sinne von Kapitel 1 – eine langfristige Verfügbarkeit und Nachnutzbarkeit der Forschungsdaten ermöglicht wird.

7.2 Kooperative Strukturen

Geisteswissenschaftliche Forschungsvorhaben werden nicht selten als institutionsübergreifende Projekte realisiert, insbesondere im Bereich der eHumanities[14]. Als aktuelles Beispiel aus dem TextGrid-Umfeld wäre hier die Faust-Edition[15] zu nennen, an der neben dem Freien Deutschen Hochstift Frankfurt die Klassik Stiftung Weimar und die Universität Würzburg beteiligt sind. Dieses Projekt verdient in diesem Kontext insofern besondere Erwähnung, als die Entwicklung eines Archivformats bereits Teil des Antrags[16] ist.

Für den Bereich der LZA von Forschungsdaten existieren bislang keine nennenswerten Kooperationen. Zwar haben Projekte mit internationaler Beteiligung wie kopal[17], Planets[18] oder SHAMAN[19] wesentlich zur Schaffung der technischen Grundlagen beigetragen, jedoch existieren bislang weder einheitliche Strategien noch standardisierte und übergreifende Lösungen, um anfallende Forschungsdaten langfristig zu archivieren. Den eingangs erwähnten (Kap. 1) Auflagen der DFG[20] bzgl. der Aufbewahrung von Forschungsdaten über einen Zeitraum von zehn Jahren kommen geisteswissenschaftliche Forschungsprojekte üblicherweise unter Rückgriff auf individuelle Lösungen nach (z.B. Ablage auf Projektrechner oder beim lokalen Rechenzentrum), wobei im günstigsten Fall allenfalls den Anforderungen der *bitstream preservation* Rechnung getragen wird. Da üblicherweise mit den zuständigen Mitarbeitern nach Projektende auch das für höherwertige LZA erforderliche inhaltliche Wis-

14 Der Begriff eHumanities wird analog zu eScience verwendet und beschreibt die durch digitale Methoden erweiterten („enhanced“ oder auch „enabled“) Geistes- und Kulturwissenschaften (vgl. auch die ausführliche Diskussion in TextGrid (2010), Abschnitt 2).

15 Vgl. Digitale Faustedition (2010).

16 Vgl. Digitale Faustedition (2008).

17 Vgl. kopal (2011).

18 Vgl. Planets (2011).

19 Vgl. SHAMAN (2011).

20 Vgl. DFG (1998), S. 12f.

sen um die jeweiligen Daten verlorengeht (s.u.), besteht hier dringender Handlungsbedarf. Grundsätzlich ist aber davon auszugehen, dass den Bibliotheken hier in Zukunft eine bedeutendere Rolle zukommt. So wurde kopal inhaltlich von der Niedersächsischen Staats- und Universitätsbibliothek Göttingen (SUB) und der Deutsche Nationalbibliothek (DNB) getragen.

Weniger kritisch stellt sich der Sachverhalt bezüglich digitaler Publikationen dar, deren langfristige Verfügbarkeit die Deutsche Nationalbibliothek als zuständige Archivbibliothek sicherzustellen hat.[21] Allerdings ist noch offen, wie mit (Forschungs-) Daten zu verfahren ist, welche die jeweilige Publikation ergänzen. Im Rahmen des EU-Projekts DRIVER[22] wurde exemplarisch ein Publikationsframework für solche sogenannten *enhanced publications*[23] im Open Access Bereich entwickelt. Jedoch liegt die Verantwortung für die Archivierung der publizierten Objekte bei den beteiligten Repositorien und Bibliotheken, so dass LZA-Aspekte im Rahmen dieses Projekts keine nennenswerte Rolle spielen.

Eine Sonderrolle nehmen in diesem Kontext digitale Editionen ein. Häufig handelt es sich bei dem abschließenden Deliverable eines solchen zeitlich befristeten Projekts um eine Website[24], bei der die zur Veröffentlichung bestimmten Forschungsdaten zumeist in einer Datenbank liegen, auf die dann die Web-Applikation zugreift. Zwar lassen sich solche Editionen zumindest theoretisch auch noch Jahre nach Projektende ggf. über entsprechende Emulationen erhalten, jedoch ist die Nachnutzbarkeit der zugrundeliegenden Forschungsdaten stark eingeschränkt, da diese durch die darüberliegende Anwendung gefiltert und für die Darstellung im Browser modifiziert werden.[25] Aus diesem Grunde ist es essenziell, wie beim bereits oben erwähnten Faust-Editionsprojekt bereits frühzeitig zwischen Präsentations- und Archivierungsschicht zu differenzieren, um eine möglichst weitgehende inhaltliche Nachnutzbarkeit der Forschungsergebnisse sicherzustellen.

---

21 Vgl. Deutsche Nationalbibliothek (2011).

22 Vgl. DRIVER (2011a).

23 Vgl. DRIVER (2011b).

24 Ein Thema, das ausführlich auf dem <philtag n="8"/> Workshop 2009 in Würzburg diskutiert wurde (vgl. o.V. (2009)).

25 Als Beispiel sei hier stellvertretend *The Electronic Text Corpus of Sumerian Literature (ETCSL)* genannt. Die Projektförderung endete 2006. Die ursprünglich in TEI XML ausgezeichneten Texte sind nur in einer HTML-Version verfügbar (vgl. ETCSL (2011)).

Wo die Aufgabe der LZA von Forschungsdaten in den Geisteswissenschaften letztlich zu verorten ist, ist noch offen. Eine klare Aufgabenteilung zwischen Wissenschaftlern, Bibliotheken und Rechenzentren hat sich noch nicht herausgebildet. Neben den bereits existierenden, fachspezifischen geisteswissenschaftlichen Zentren[26] entstehen derzeit an einigen Universitäten Digital Humanities Zentren[27], an denen alle für die LZA relevanten Institutionen beteiligt sind. Diese Zentren könnten Service und Support auf regionaler Ebene koordinieren.

Das TextGrid-Projekt verfügt über die Zusammenarbeit in der D-Grid Initiative über Grid-Ressourcen, die zurzeit gemeinsam u.a. mit der Medizin und Physik im GoeGrid-Zentrum vom Göttinger Rechenzentrum, der GWDG[28], betrieben werden. Die Kooperation innerhalb des Verbundprojekts ermöglicht es auch, dass zukünftig für verteilte Daten und Dienste sowie Replika-Management weitere Grid-Knoten in Mannheim und Würzburg eingerichtet werden.

7.3 Daten und Metadaten

Entsprechend der primären Ausrichtung auf die Textwissenschaften fallen in erster Linie die für diesen Bereich typischen Daten an. Hierbei handelt es sich zum einen um Digitalisate beliebiger Textquellen wie Scans von Drucken, Manuskripten, Briefen etc., daneben aber auch Fotografien von Inschriften und dergleichen. Bei diesen Objekten handelt es sich in der Regel um TIFF-Dateien, aber auch andere Formate wie JPEG, PNG oder PDF kommen vor. Daneben existieren Rohdaten von Texten (d.h. ohne inhaltliches Markup), die manuell via *double keying*[29] oder mittels OCR erfasst wurden. Diese Daten werden häufig in Office-Formaten (MS Word, RTF) angeliefert und enthalten lediglich typografische Informationen. Eine dritte wichtige Gruppe stellt Textdaten dar, die im Rahmen der wissenschaftlichen Arbeit erstellt und/oder annotiert werden und über inhaltliches Markup verfügen. Diese liegen üblicher-

26 Vgl. BMBF (2011c) sowie die Karte bei BMBF (2011d).
27 Zum Beispiel das Göttingen Centre for Digital Humanities (vgl. GCDH (2011)).
28 Vgl. GWDG (2011a).
29 D.h. doppeltes, manuelles Erfassen mit anschließendem Abgleich und Fehlerkorrektur.

weise im XML-Format vor – im Fall von TextGrid bevorzugt TEI P5[30]. In TextGrid werden weiterhin technische Objekte wie XML-Schemata, XSLT-Stylesheets, Workflows etc. abgelegt und archiviert, aber auch hierbei handelt es sich in der Regel um Daten im XML-Format. Generell können im TextGrid Repository beliebige Objekte in beliebigen Dateiformaten abgelegt und zumindest der *bitstream preservation* zugeführt werden. Teile der Infrastruktur (z.B. Suche über XML-Daten) und vor allem das TextGrid Laboratory mit seinen Werkzeugen sind aber in erster Linie auf die Verarbeitung der eben genannten Datentypen und -Formate ausgelegt. Daneben ist jedoch langfristig mit einem immer höheren Aufkommen an Daten in Multimedia-Formaten, also Audio- oder Video-Aufzeichnungen zu rechnen, die insofern besondere Anforderungen an das Datenmanagement stellen, als eine Vielzahl von Codecs existiert, die bei Format-Validierungen und -Konvertierungen zu berücksichtigen sind. Aber auch Textdaten stellen hohe Anforderungen an das Datenmanagement: Da es sich bei digitalen Editionen um sehr komplexe Gebilde handeln kann, die fallweise aus vielen Einzelobjekten bestehen, gilt es, die Bezüge der Objekte untereinander – z.B. interne Querverweise – langfristig zu pflegen und Links/Referenzen durch Persistent Identifier zu ersetzen.

Wie oben erwähnt, existieren weder in den Geisteswissenschaften im Allgemeinen noch in den Textwissenschaften im Speziellen etablierte Datenzentren, die übergreifend Langzeitarchivierungsdienste anbieten würden. Was die Publikation von Forschungsdaten angeht, so existieren zwar verschiedene Repositorien und Portale, die aber in der Regel auf die Daten bestimmter Institutionen (z.B. Max Planck Digital Library oder der im Aufbau befindliche Wissensspeicher der BBAW) oder projektspezifische Inhalte (z.B. Deutsches Textarchiv) beschränkt sind. Der britische AHDS (Arts and Humanities Data Service)[31], bei dem es sich tatsächlich um eine übergreifende Publikationsplattform für Forschungsdaten handelte, wurde 2008 abgewickelt.

Da das *TextGrid Repository* technisch prinzipiell dazu geeignet wäre, diese Lücken zu füllen bzw. den einen oder anderen zentralen Baustein zu

30 Vgl. TEI (2011a).
31 Vgl. AHDS (2008).

einem Datenzentrum für die Geisteswissenschaften beizutragen, sollen hier einige Aspekte etwas ausführlicher dargestellt werden.

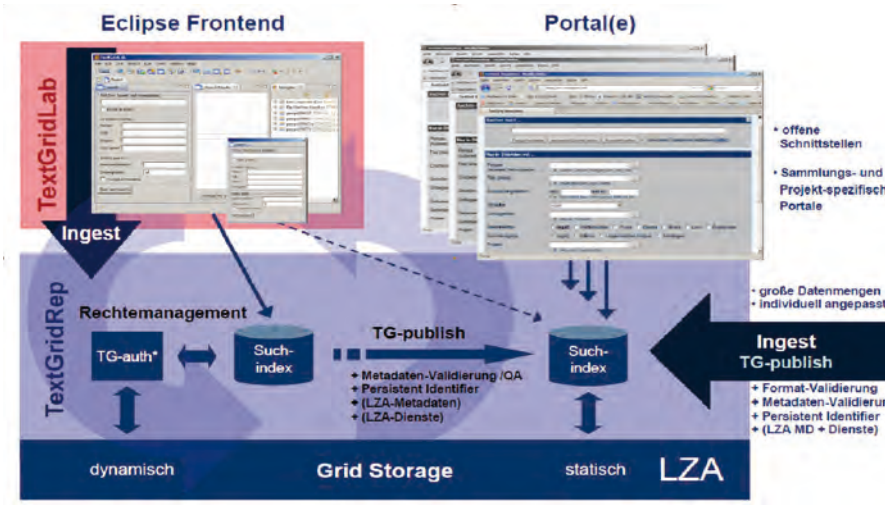

Abb. 1: Publikations-Workflow (TG-publish) in TextGrid (© TextGrid 2011)

Das *TextGrid Repository* erfüllt insbesondere zwei Anforderungen: Erstens müssen sowohl der/die einzelne Fachwissenschaftler/in als auch Datenanbieter mithilfe des Rechtemanagements jederzeit die Hoheit über ihre Daten behalten können, zweitens müssen die (in/mit TextGrid erstellten) Forschungsdaten (auch Zwischenergebnisse) langfristig zugänglich und zitierfähig sein. Daher sind im TextGrid Repository zwei Speicherbereiche mit zugehörigen Suchindizes für die Daten vorgesehen (vgl. Abbildung 1):

a. Ein dynamischer Speicherbereich, der für die Arbeit an eigenen, unpublizierten Daten mit dem TextGrid Laboratory vorgesehen ist. Wenn ein(e) Nutzer(in) in TextGrid Daten importiert oder neu erstellt, dann ist er/sie zunächst die einzige Person, die diese Daten einsehen und bearbeiten kann. Über das Nutzer- und Rechtemanagement kann er/sie anderen TextGrid-Nutzern Rollen zuweisen, die mit entsprechenden Rechten zum Lesen, Schreiben und Löschen von Objekten verbunden sind. Sichtbar sind immer nur die Objekte, auf die ein(e) Nutzer(in) mindestens Lesezugriff hat.

b. Ein statischer Speicherbereich, der für öffentliche, publizierte Daten vorgesehen ist. Wenn Nutzer in TextGrid publizieren, d.h. ihre Forschungsdaten öffentlich zugänglich machen wollen, dann werden bei diesem Publikationsvorgang die Daten eingefroren, d.h. sie sind inhaltlich nicht mehr änderbar, werden in einem separaten Speicherbereich abgelegt und mit Persistent Identifiern versehen, so dass sie unter einer festen Adresse langfristig zitierbar sind. Die auf diese Weise publizierten Objekte sind dann sowohl im TextGrid Lab von allen Nutzern auffindbar als auch über ein Internetportal einfach such- und abrufbar. Daneben existieren offene, REST[32]-basierte Schnittstellen, die die Implementierung spezifischer, z.B. projektgebundener Frontend-Applikationen ermöglichen. Damit diese Daten im wissenschaftlichen Sinne zitierfähig sind, können sie nicht mehr inhaltlich modifiziert werden. Wenn die Daten aktualisiert werden sollen, muss eine neue Version angelegt und diese ebenfalls publiziert werden. Der statische Speicherbereich ist der Bereich des TextGrid Repository, der für die Langzeitarchivierung vorgesehen ist. Daneben wird es eine Dark Archive Lösung geben, bei der die Daten zwar auch eingefroren werden, aber gewissen Zugangsbeschränkungen unterliegen (Lizenz- und Rechtemanagement).

Bei neu in TextGrid angelegten Objekten müssen einige wenige inhaltliche Metadaten vom Nutzer eingegeben werden, während die technischen und administrativen Metadaten weitestgehend automatisch generiert werden. Bei der Publikation eines Objekts im o.g. Sinne erfolgt eine Qualitätskontrolle (inklusive Validierung), bei der weitere (inhaltliche, bibliographische) Metadaten eingegeben werden müssen. Bei der Publikation von Forschungsdaten aus dem *TextGrid Laboratory* wird ein Prozess angestoßen, der automatisch Persistent Identifier vergibt, Format und Metadaten validiert, sowie die Daten im statischen, d.h. LZA-Speicherbereich ablegt und dort nachweist. Die Trennung in zwei Suchindizes gewährleistet, dass die Suche über öffentliche Daten performant ablaufen kann, da hier keine Zugriffsrechte abgefragt werden müssen.

32 Vgl. o.V. (2011c).

Daneben existiert eine Routine für den Masseningest in das *TextGrid Repository*. Hierzu wird die *kopal Library for Retrieval and Ingest* (koLibRI)[33] eingesetzt, die verschiedene Module zur Durchführung der eben genannten Arbeitsschritte in einem entsprechenden Workflow bündelt. Zur entsprechend den jeweiligen Gegebenheiten teil- oder vollautomatischen Vergabe von Metadaten werden Templates eingesetzt. Damit die angelieferten Daten von der Ingest-Routine korrekt als TextGrid-Objekte modelliert werden, müssen diese in einer bestimmten Struktur vorliegen, d.h. sie müssen einem sog. Ingest-Profil entsprechen. Standardmäßig werden zwei solcher Profile unterstützt um die zu importierenden Daten auf das TextGrid-Objektmodell[34] abzubilden: DFG-Viewer METS[35] und eine Verzeichnisstruktur, deren (Unter-) Verzeichnisse in TextGrid als Aggregationen umgesetzt werden. Gegebenenfalls müssen die angelieferten Daten entsprechend vorbehandelt werden.

Derzeit wird an der Erstellung von SLAs sowie Richtlinien für Datenlieferanten zur Aufnahme von Forschungsdaten in das *TextGrid Repository* gearbeitet. Letztere orientieren sich am nestor Leitfaden für die Informationsübernahme in das Langzeitarchiv[36] und unterscheiden zwischen den Rahmenbedingungen, die jeweils für *bitstream preservation*, *content preservation* und *data curation* maßgeblich sind.

Bezüglich der Erstellung von Service Level Agreements ist eine Task Force ins Leben gerufen worden, der Vertreter verschiedener an TextGrid beteiligter Institutionen sowie der GWDG als wichtigstem Ressourcen-Provider angehören. Neben der Identifikation der unterschiedlichen, von den SLAs zu berücksichtigenden Rollen (TextGrid-Konsortium, Ressourcen-Provider, Content-Provider, Einzelnutzer) und deren Beziehungen zueinander geht es darum, Nutzungsversprechen und Dienstgütevereinbarungen (= Service-Level) zu definieren.

33 Vgl. kopal (2011).

34 Zum TextGrid-Objektmodell siehe das aktuelle TextGrid Nutzerhandbuch (2011), Abschnitt 5 *TextGrid Objects* sowie die eher technisch orientierte Darstellung in Aschenbrenner; Panzer; Pempe (2010), Abschnitt 2 *TextGrid Basics – Concepts and Infrastructure*.

35 Vgl. DFG (2009b), S. 23.

36 Vgl. nestor (2011b).

Für die Langzeitarchivierung im TextGrid Repository werden vier Service-Level angestrebt[37]:

1. Einfache Langzeitarchivierung (*bitstream preservation*): Offline Speicherung, d.h. nicht notwendigerweise Just-In-Time-Zugriff, Redundante Speicherung, 10 Jahre Aufbewahrungsfrist gem. den Vorgaben der DFG

2. Einfache Langzeitarchivierung in der Virtuellen Forschungsumgebung: wie Level 1, aber zusätzlich online verfügbar, d.h. Just-In-Time-Zugriff

3. Langzeitarchivierung mit Bewahrung des digitalen Inhaltes (*content preservation* – im Sinne von technischer Nachnutzbarkeit): Wie Level 2, aber zusätzlich technisches Qualitätsmanagement, u.a. durch Formatvalidierung und Bewahrung der digitalen Inhalte durch Formatkonvertierungen

4. Langzeitarchivierung mit Datenpflege (*data curation*): Wie Level 3, aber mit zusätzlichen Maßnahmen, die die inhaltliche bzw. fachwissenschaftliche Nachnutzbarkeit der Daten sicherstellen, d.h. kontinuierliche Pflege der Metadaten und der inhaltlichen Bezüge (z.B. durch Verlinkung thematisch verwandter bzw. für das Hintergrundverständnis notwendiger Daten) sowie der Dissemination der Daten und Metadaten über eine Webschnittstelle.

Daneben gilt es, ggf. zusätzliche Vereinbarungen über die jeweils notwendigen/erwünschten technischen Rahmenbedingungen bzgl. der zugrunde liegenden *bitstream preservation* zu treffen, wie die Häufigkeit von Integritäts-Checks, Redundanz-Stufen, geografisch verteilte Speicherung der Daten etc.[38]

Unabhängig vom TextGrid-Kontext existieren in der Community bislang noch keine einheitlichen Mindestanforderungen für die Aufnahme von Forschungsdaten. Als richtungweisend könnten sich Ansätze wir das

37 Die folgende Aufzählung zitiert ein internes Dokument (TextGrid SLA (2011)), das voraussichtlich noch dieses Jahr veröffentlicht werden wird.

38 Vgl. Ludwig et al. (2011).

„Data Seal of Approval“[39] erweisen, das Richtlinien für *data provider* jeweils relativ zu den Anforderungen des gewählten Repositoriums vorgibt[40] – wobei es sich hier primär um eine Blaupause handelt.

Was das zu erwartende Volumen langfristig zu archivierender geisteswissenschaftlicher Forschungsdaten angeht, so dürfte selbst eine ungefähre Angabe aufgrund der mehrfach angesprochenen Breite der Forschungslandschaft kaum möglich sein. Die Eingrenzung auf einen „Kanon“ sowohl was die Inhalte, als auch was die Datentypen angeht, ist allenfalls ansatzweise möglich. Während das *TextGrid Repository* derzeit im Aufbau befindlich ist und sich das Datenvolumen noch in einem überschaubaren Bereich (< 50TB) bewegt, müsste ein Datenzentrum für die Geisteswissenschaften jedoch mit ganz erheblichen Datenmengen, auf jeden Fall im Petabyte-Bereich, operieren.

Derzeit fallen die größten Forschungsdaten-Volumina im Bereich der hochwertigen Digitalisierung an. Auch wenn davon auszugehen ist, dass nur ein Bruchteil der erfassten Digitalisate für geisteswissenschaftliche Forschungsfragen relevant ist, so liefern die Erfahrungen aus dem Göttinger Digitalisierungszentrum (GDZ) doch ein eindrucksvolles Bild: Im Jahr 2010 hat das GDZ 200.000 Seiten pro Monat digitalisiert. Jede Seite (color, 300 dpi, TIFF) benötigt 2×25 MB Speicherplatz (Master und optimiertes TIFF). Hinzu kommen Metadaten und Volltexte, jedoch sind diese aufgrund ihres geringen Volumens für diese Rechnung vernachlässigbar. Damit ergibt sich ein Datenvolumen von 10 TB/Monat oder 3,8 MB/s bei einem 24/7-Betrieb. Hochgerechnet auf die gesamte Digitalisierung in Deutschland kann von ungefähr 100 TB/Monat ausgegangen werden, also 1,2 Petabyte pro Jahr. Der gesamte europäische Bestand beträgt nach Schätzungen ca. 1,93 Milliarden Buchseiten[41], was nach dem Standard des Göttinger Digitalisierungszentrums 96 Petabyte entsprechen würde. Hinzu kommen theoretisch Archivalien, die europaweit (Bibliotheken und Nationalarchive) auf knapp 21 Milliarden Seiten geschätzt werden.[42] Allein das deutsche Literaturarchiv in Marbach be-

39 Vgl. Data Seal of Approval (2011b).
40 Vgl. ebenda.
41 Vgl. Niggemann; De Decker; Levy (2011); vgl. Poole (2010), S. 30.
42 Vgl. Poole (2010), S. 41 und 62. Für Deutschland existiert mit Kalliope eine zentrale Datenbank für Nachlässe und Autografen (vgl. Kalliope (2011)).

herbergt etwa 1200 Vor- und Nachlässe, die auf jeden Fall als geisteswissenschaftliche Forschungsdaten anzusehen sind.[43] Eindrucksvoll auch die Zahl der beim Altägyptisches Wörterbuch (Thesaurus Linguae Aegyptiae, BBAW)[44] digitalisierten Belegzettel: ca. 1,5 Millionen. Bei diesen Beispielen handelt es sich nur um (potentielle) Primärdaten aus der Digitalisierung, hinzu kämen noch alle Adaptionen und Neuausgaben der Daten.

Neben solchen Volumina fallen die eigentlichen Textdaten kaum ins Gewicht: Die von TextGrid erworbenen und nach XML/TEI aufbereiteten ca. 1200 Werke der *Digitalen Bibliothek* von zeno.org[45] umfassen gerade einmal knapp 2,5 GB. Weiterhin ist mit einer großen Anzahl hochauflösender Fotografien und Faksimiles aus der Kunstgeschichte und verwandten Disziplinen zu rechnen, sowie in Zukunft vermehrt mit Audio- und Videodateien, die neben einem nicht unerheblichen Speicherbedarf[46] auch mutmaßlich besonders hohe Anforderungen an *content preservation* und *data curation* stellen. Aus diesem Grund bietet TextGrid in der aktuellen Projektphase keinen gezielten Support für Multimedia-Formate an.

Um wieder zum engeren TextGrid-Rahmen zurück zu kommen, so ist die Standardisierung der Formate, insbesondere was die Textauszeichnung angeht, in einigen Bereichen schon weit fortgeschritten. Beispielhaft sei hier die Text Encoding Initiative[47] genannt, ein Standardisierungsgremium, das seit 1994 Richtlinien für die Textkodierung entwickelt. Der aktuelle, XML-basierte Standard TEI P5 findet weite Verbreitung in den Textwissenschaften. Auch in anderen Disziplinen wie z.B. der Musikwissenschaft[48] sind ähnliche Standards entwickelt worden. TextGrid schreibt keine Datenformate vor, bietet jedoch optimale Unterstützung für XML/TEI, sowohl was die über das TextGrid Lab verfügbaren Werkzeuge, als auch die Indexierung und Durchsuchbarkeit im Repository

43 Vgl. Deutsches Literaturarchiv Marbach (2011).
44 Vgl. Altägyptisches Wörterbuch (2011).
45 Vgl. TextGrid (2011b).
46 So umfasst beispielsweise alleine der derzeit digitalisierte Bestand des University of Southern California Shoah Foundation Institute ca. 135TB an Videoaufzeichnungen (Digitalisierung zu 67% abgeschlossen, Stand: 31.7.2011, vgl. http://dornsife.usc.edu/vhi/preservation/USC_SFI_preservation_details.pdf; USC Shoah Foundation (2011)).
47 Vgl. TEI (2011b).
48 Vgl. MEI (2011).

angeht. Letztgenanntem Zweck dient das sogenannte *Baseline Encoding* (auch als „Kernkodierung“ bezeichnet)[49], ein TEI-basiertes, Textsorten-spezifisches Markup, das die TextGrid-übergreifende Suche über XML-Daten ermöglichen bzw. erleichtern soll. Mittels sogenannter Adaptoren (i.d.R. XSLT-Stylesheets) kann so beim Ingest eine zweite, Baseline-Instanz einer XML-Datei im Suchindex angelegt werden, auf die die Standard-Suche in TextGrid zugreift. Das im Grid gespeicherte Original-Objekt bleibt davon jedoch völlig unberührt. Weiterhin werden die bereits oben erwähnten Grafikformate in ähnlicher Weise unterstützt (z.B. Text-Bild-Link Editor, digilib-Einbindung[50]).

Das auf Dublin Core aufbauende Metadatenschema von TextGrid lässt sich beliebig um projektspezifische Sub-Schemata erweitern und unterstützt beim Import und Export METS[51]/MODS[52] und Dublin Core. Um Eingabefehler zu vermeiden, die Suche zu erleichtern und um ein möglichst hohes Maß an Interoperabilität zu ermöglichen, setzt TextGrid verstärkt auf die Einbindung kontrollierter Vokabulare und Norm-/Referenzdaten, wie z.B. für Personen und Körperschaften (GND) oder Geodaten (Getty Thesaurus of Geographic Names). Diese Dienste werden zukünftig als Bestandteil der DARIAH-Infrastruktur gemeinsam mit DARIAH betrieben.

Im *TextGrid Repository* werden nicht ausschließlich frei verfügbare Forschungsdaten abgelegt, sondern auch solche, die aufgrund Urheberrechts- oder vertraglicher Bestimmungen bestimmten Zugriffs-Restriktionen unterliegen. Dies trifft in besonderem Maße auf die meisten vom Institut für Deutsche Sprache in Mannheim (IDS)[53] vorgehaltenen Sprachkorpora zu. Die vom IDS in TextGrid zur Verfügung gestellten Ressourcen werden deshalb mit einer zusätzlichen Komponente für das Lizenz- und Rechtemanagement gekoppelt werden (TG-license, inkl. *Moving Wall* Funktionalität). Aufgrund vertraglicher Bestimmungen dürfen diese Inhalte ausschließlich auf einem vom IDS gehosteten (Text)Grid-Knoten in Mannheim vorgehalten werden. Replikation der Daten kann nur erfolgen,

49 Vgl. TextGrid (2009).
50 Vgl. digilib (2011).
51 Vgl. METS (2011).
52 Vgl. MODS (2011) hier als Erweiterung zu METS.
53 Vgl. IDS (2011).

wenn diese zuvor verschlüsselt werden. Die hierfür entwickelten Lösungen sind so generisch gehalten, dass sie sich auch für andere, ähnlich gelagerte Fälle nachnutzen lassen. Darüber hinaus können im Kontext einzelner Projekte (z.B. EHRI)[54] personenbezogene Daten anfallen, die aus Gründen des Datenschutzes ebenfalls nicht frei zugänglich gemacht werden dürfen.

Neben den eben erwähnten, spezifisch juristischen Zugriffsbeschränkungen besteht in den Geisteswissenschaften (und vermutlich nicht nur dort) aber auch generell das Problem, dass ältere Forschungsdaten zwar häufig in digitaler Form vorliegen, jedoch nicht (mehr) zugänglich sind und/oder in proprietären oder nicht mehr lesbaren Formaten vorliegen, geschweige denn mit deskriptiven Metadaten inhaltlich erschlossen sind. Dies betrifft häufig Daten aus Projekten, die seit mehreren Jahren abgeschlossen aber für die weitere Forschung durchaus noch von Interesse sind. Üblicherweise geht mit dem Projektpersonal das Wissen um die Daten verloren. Als Beispiel hierfür wurden bereits oben digitale Editionen genannt. Daneben existieren weitere Szenarien, in denen entweder nur die Daten oder die Metadaten in digitaler Form vorliegen, während der jeweils andere Part ausschließlich handschriftlich, gedruckt oder gar nicht mehr existiert. Um solche Daten nachträglich zu erschließen und für die weitere Forschung nutzbar zu machen, fehlen üblicherweise die Ressourcen – und entsprechende Förderlinien.

Ein weiterer kritischer Punkt, der aber nicht notwendigerweise spezifisch für die Geisteswissenschaften ist, ist der Umstand, dass die Rechte an (traditionell) publizierten Forschungsdaten häufig über längere Zeiträume bei Verlagen liegen, die den Zugang üblicherweise nur gegen Bezahlung gewähren – und dies dann nicht immer in einer Form bzw. einem Format, das eine optimale Nachnutzung ermöglicht.

Aufgrund der eingangs thematisierten Heterogenität geisteswissenschaftlicher Forschungsdaten existiert auch kein einheitliches Metadatenschema zur Beschreibung derselben. Gleichwohl hat sich mit Dublin Core[55] (DC) ein Standard etabliert, auf den, eher als auf explizit bibliothekarische Formate wie MARC[56] etc., häufig zurückgegriffen wird. Allerdings bietet DC so viele Freiheiten (was freilich auch eine große Stär-

54 Vgl. EHRI (2011).
55 Vgl. DCMI (2011).
56 Vgl. MARC (2011).

ke ist), dass für Zwecke der Interoperabilität mit detaillierten *Application Profiles* gearbeitet werden muss. Da sich DC weitgehend auf deskriptive Metadaten beschränkt, müssen auf DC basierende Metadaten-Schemata für Aspekte der LZA entsprechend erweitert werden, beispielsweise anhand von PREMIS[57]. Wie bereits oben erwähnt, baut das TextGrid Metadaten-Schema auf DC auf, unterscheidet jedoch strikt zwischen inhaltlichen bzw. deskriptiven und technischen Metadaten, wobei letztere weitestgehend automatisch generiert werden. Daneben existieren administrative Metadaten für Projektverwaltung und Rechtemanagement. Ein Metadatensatz in TextGrid hat folgende Struktur[58]:

- **generic** – dieser Bereich des Metadatenschemas ist für Objekttyp-unabhängige Metadaten bestimmt, die für jedes TextGrid-Objekt verpflichtend sind
  - **provided** – minimale deskriptive Metadaten. Um die Hürden für die Nutzer möglichst niedrig zu halten, müssen hier verpflichtend nur Angaben zum Titel und dem Typ des Objekts gemacht werden.
  - **generated** – technische und administrative Metadaten, die vom System angelegt und und laufend aktualisiert werden, wie Projekt, Dateigröße, letztes Änderungsdatum, Nutzer-Id, Objekt-Identifier (sog. TextGrid-URI) u.a.m.
- **item / edition / work / collection** – Objekttyp-spezifische Metadaten. Bei item handelt es sich entweder um die eigentlichen Inhaltsobjekte, z.B. eine Grafik, ein XML-Dokument, oder um eine sogenannte Aggregation, d.h. ein virtuelles Verzeichnis, das nur die Referenzen auf seine Inhaltsobjekte enthält. Einziges item-spezifisches Metadatum ist rightsHolder, das auch nur bei der Publikation verpflichtend wird. Träger der inhaltlichen / bibliografischen Metadaten sind edition und work. Werden items publiziert (s.o.), müssen diese entweder einem edition-Objekt (und dieses wiederum einem work) oder einer collection

57 Vgl. PREMIS (2011).

58 Vgl. das aktuelle TextGrid Metadatenschema unter http://www.textgrid.info/schemas/textgrid-metadata_2010.xsd; zum Objektmodell vgl. Fußnote 34.

zugeordnet werden, wobei es sich bei letzterem um einen generischen Container für nicht-bibliografische Metadaten/Objekte handelt.[59]

- **custom** – dieser Bereich des Metadatenschemas ist als Container für beliebige, nutzer-/projektspezifische Metadaten gedacht, die in einem separaten Namensraum liegen sollten.
- **relations** – hier lassen sich beliebige Beziehungen in RDF-XML einfügen, die beim Ingest in den TextGrid Triplestore eingetragen werden.

Die bei der Formatvalidierung („Publikation", s.o.) generierten technischen Metadaten werden als separate Objekte abgelegt.

Auch TextGrid nutzt einen über EPIC[60] institutionell abgesicherten Handle-Service zur Vergabe von Persistent Identifiern (vgl. das Kapitel zur Psycholinguistik). Der lokal in Anspruch genommene Service wird von der GWDG betrieben.[61] Bei der DNB archivierte Publikationen (s.o.), die fallweise ebenfalls als Forschungsdaten kategorisiert werden können, werden mit URNs versehen.

7.4 Interne Organisation

Grundsätzlich steht die LZA geistes- bzw. textwissenschaftlicher Forschungsdaten noch am Anfang. Nachdem bezüglich der technischen Lösungen weitestgehend Klarheit besteht, gilt es nunmehr, die organisatorischen, rechtlichen und finanziellen Rahmenbedingungen zu klären. Neben einem Geschäfts- und Fördermodell, das den langfristigen Betrieb sicherstellt, muss auch eine Organisation oder Körperschaft existieren, die als Vertragspartner und Betreiber des Archivs auftreten kann. Weiterhin müssen die rechtlichen Rahmenbedingungen geklärt

59 Diese zugegeben etwas komplexen Beziehungen dienen der Vermeidung von Redundanzen (und unterstützen somit die Konsistenz der Metadaten innerhalb des TextGrid Repository) und sollen andererseits sicherstellen, dass die Objekte bei der Publikation ausreichend und mit qualitativ hochwertigen Metadaten beschrieben sind. Die Nutzeroberfläche des TextGrid Lab mindert diese Komplexität jedoch deutlich, so dass vom Nutzer stets nur ein möglichst geringes Maß an Interaktion erbracht werden muss.

60 Vgl. EPIC (2011).

61 Vgl. GWDG (2011b).

werden, d.h. es müssen allgemeine Nutzungsbedingungen existieren, in denen die Rechte und Pflichten aller Beteiligten festgelegt sind, sowie Service Level Agreements (SLAs) für die Inanspruchnahme bestimmter Dienstleistungen, Redundanzstufen und Archivierungsoptionen (*bitstream preservation*, *content preservation* und *data curation*) formuliert werden (s.o.). Diese Punkte müssen in allernächster Zukunft zumindest vorläufig gelöst werden. Hierbei kann auf Ergebnisse und Erfahrungen anderer Projekte und Initiativen, wie nestor und DP4lib[62] zurückgegriffen werden.

Sowohl die Entwicklung des TextGrid Repository als auch der von TextGrid zu adaptierenden Entwicklungen aus WissGrid[63] sind jeweils Bestandteil der Projektanträge und somit finanziell voll von der jeweiligen Projektförderung abgedeckt. Die derzeit genutzten Hardwareressourcen stammen aus den D-Grid Sonderinvestitionen des BMBF und decken mit 275 TB (Plattenspeicher und Tape) voraussichtlich den Bedarf bis Ende 2011. Der Betrieb der Ressourcen wird von der GWDG bis auf weiteres aus Eigenmitteln finanziert.

Um die mittel- und langfristige Finanzierung des Betriebs von LZA-Einrichtungen bzw. Infrastrukturen sicherzustellen, muss das Problem von verschiedenen Seiten angegangen werden. Einerseits muss eine Grundfinanzierung existieren, die langfristig den Betrieb und die Verfügbarkeit der Basisdienste als solche sicherstellt. Andererseits müssen, bspw. wenn es um *content preservation* und *data curation* geht, die damit einhergehenden zusätzlichen Aufwände von den jeweiligen „Kunden", d.h. aus dem jeweiligen Projektkontext getragen werden. Denkbar wäre im letztgenannten Fall die Einführung einer Datenpauschale für DFG-Projekte – analog zur Publikationspauschale. Damit könnten Wissenschaftler bzw. Projekte selbst passende Anbieter auswählen – ggf. abhängig von der fachlichen Ausrichtung des jeweiligen Langzeitarchivs (siehe dazu weiter unten). Universitäten könnten z.B. eigene Zentren betreiben oder sich bei anderen Anbietern „einkaufen". Hierzu sollte jedoch Klarheit bezüglich der zu erwartenden, mittel- und langfristig anfallenden Kosten bestehen.

62 Vgl. DP4lib (2011).

63 WissGrid Dienste Framework (LZA-Dienste) und Grid-Repository (vgl. zu den jeweils aktuellen Spezifikationen WissGrid (2011a)).

Was die konkreten Kosten der Langzeitarchivierung in TextGrid angeht, so lassen sich diese derzeit noch schwer spezifizieren. Zunächst fehlen hier Erfahrungswerte über den laufenden Betrieb und das durchschnittliche Datenaufkommen (wie etwa in der Klimaforschung oder der Psycholinguistik)[64], außerdem müssen die jeweiligen Service Level genau spezifiziert werden, damit die beteiligten Ressourcen-Provider die Kosten für den technischen Betrieb kalkulieren können. Bei den höherwertigen Diensten wie *data curation* ist zudem die Aufgabeteilung zwischen dem TextGrid Konsortium, das die Dienste nach außen hin anbietet, und den Ressourcen-Providern (GWDG) zu klären. Weiterhin müssen in diesem Kontext etwaige Entschädigungen für Datenverlust berücksichtigt werden, wobei die Versicherungsbeiträge auf die Betriebskosten aufgeschlagen werden müssten. Dies zeigt, wie dringend hier konkrete Vereinbarungen und Spezifikationen bezüglich der angebotenen Dienste und Service Level benötigt werden.

Neben der organisatorischen, finanziellen und technischen Nachhaltigkeit gilt es auch, die fachwissenschaftliche Nachhaltigkeit bzgl. *content preservation* und insbesondere *data curation* sicherzustellen. Hier geht es in erster Linie darum, das mit den zu archivierenden Daten vertraute Personal der Datenlieferanten zu beraten und gemeinsam sicherzustellen, dass dieses Material auch langfristig fachwissenschaftlichen Anforderungen genügt, d.h. es von den jeweiligen Communities über einen längeren Zeitraum nachgenutzt werden kann. Um diese Dienstleistungen nachhaltig und zuverlässig anbieten zu können, müssen an den beteiligten Institutionen auch langfristig Stellen für technisch und wissenschaftlich qualifiziertes Personal vorgehalten werden – ein Problem, das auch der Wissenschaftsrat in seinen *Empfehlungen zu Informationsinfrastrukturen* thematisiert: „Forschungsprimärdaten, die teilweise mit großem finanziellem und personellem Aufwand erhoben werden, werden der wissenschaftlichen Öffentlichkeit oft nicht zugänglich gemacht, weil keine Mittel zur Finanzierung von wissenschaftlichem Per-

64 Für Langzeitarchivierung von Publikationen im engeren Sinne („E-literature") wurden im Rahmen des LIFE³ Projekts (LIFE (2010)), ausgehend von verschiedenen Fallstudien und basierend auf einem Lifecycle-Modell, konkrete Kostenpläne erstellt. Auch wenn sich die Ergebnisse nicht ohne weiteres verallgemeinern lassen, so bieten doch die zugrunde gelegten Parameter wertvolle Anhaltspunkte für eigene Kalkulationen (vgl. Hole et al. (2010)).

sonal für die hierzu erforderlichen Tätigkeiten der Datenaufbereitung, -anonymisierung, -übermittlung und -pflege zur Verfügung stehen. Ähnliches gilt für Digitalisierungsprojekte."[65]

Grundsätzlich besteht immer das Problem, dass bei Forschungsdaten, die aus dem Kontext eines zeitlich befristeten Projekts in die LZA gegeben werden, *data curation* nur sehr schwer über einen längeren Zeitraum sichergestellt werden kann, da die Projektmitarbeiter, die über das hierfür erforderliche inhaltliche Wissen verfügen, üblicherweise nach Projektende nicht mehr zur Verfügung stehen. Die Weitergabe des fachwissenschaftlichen Hintergrundwissens an die zuständigen Mitarbeiter der archivierenden Institution ist hierfür unerlässlich, kann sich aber je nach Komplexität des Forschungsgegenstands und der Daten als beliebig problematisch erweisen. In jedem Fall ist eine ausführliche Dokumentation vorzulegen.

Langfristig kann *data curation* nur von eigens geschultem Personal in enger Kooperation mit den Datenlieferanten bzw. -Produzenten durchgeführt werden. Möglicherweise ist es sinnvoll, hierfür auf bestimmte Forschungsschwerpunkte und -daten spezialisierte Datenzentren einzurichten, für die dann, gemäß den Empfehlungen des Wissenschaftsrats, eine entsprechende Grundfinanzierung sichergestellt sein muss.[66]

7.5 Perspektiven und Visionen

Um nochmals auf den Ausgangspunkt, die Heterogenität geisteswissenschaftlicher Forschungsdaten, zurückzukommen, sei hier nochmals auf die Probleme der Langzeitarchivierung insbesondere in Hinblick auf *data curation* hingewiesen. Neben den technischen, organisatorisch-administrativen und finanziellen Herausforderungen gilt es auch die fachwissenschaftliche Nachnutzbarkeit von Forschungsdaten nachhaltig sicherzustellen. Wie im vorangehenden Abschnitt anhand der Frage des für die LZA zuständigen Personals entwickelt, könnte eine dezentrale Struktur föderierter, jeweils auf bestimmte Aspekte spezialisierter Repositorien, die organisatorisch und finanziell in einem Verbund zusammengeschlos-

65 Wissenschaftsrat (2011b), S. 50.
66 Vgl. ebenda, S. 50f.

sen sind, diesen Rahmenbedingungen am ehesten Rechnung tragen. Eine solche Lösung kann aber nur gemeinsam mit allen Beteiligten, sowohl den Förderern als auch den Infrastrukturanbietern und Datenlieferanten, d.h. den Forschenden etabliert werden. Die für eine solche Föderation erforderlichen technischen Grundlagen sind im Prinzip vorhanden[67] und wurden bereits im Rahmen der DARIAH *preparation phase* am Beispiel von TextGrid experimentell umgesetzt[68].

Generell bietet die Kooperation mit DARIAH[69] große Chancen für TextGrid. DARIAH errichtet im Rahmen der ESFRI-Roadmap[70] eine europaweite Forschungsinfrastruktur für die Geisteswissenschaften und wird nach der erfolgreich abgeschlossenen *preparatory phase* seit kurzem in verschiedenen Teilnehmerstaaten auf nationaler Ebene gefördert – in Deutschland als DARIAH-DE[71] vom BMBF. Parallel hierzu existiert mit CLARIN (Common Language Resources and Technology Infrastructure[72]) ein weiteres ESFRI-Projekt, dessen Infrastruktur-Planungen dezidiert auf den Umgang mit Sprachressourcen ausgerichtet sind. Beide Projekte kooperieren bereits seit längerem miteinander, für die laufende Förderphase sind gemeinsame Arbeitsgruppen zwischen DARIAH-DE und CLARIN vorgesehen. In diese Konstellation ist TextGrid alleine schon über die Zusammensetzung der Partner mit eingebunden: Viele der TextGrid-Partner sind ebenfalls in DARIAH und CLARIN involviert. Da DARIAH bestrebt ist, bereits bestehende Infrastrukturdienste und Werkzeuge mit einzubinden bzw. auf diese aufzubauen, bieten sich hier konkrete Kooperationsszenarien mit TextGrid. Neben dem bereits erwähnten gemeinsamen Betrieb von Referenzdaten-Diensten ist eine enge Zusammenarbeit im Bereich Storage-Infrastruktur sowie u.a. die Nutzung übergreifender Registries für Content und Metadaten geplant. Auf diese Weise werden Teile der TextGrid Infrastruktur auf eine breitere Basis gestellt und somit eine gewisse technische Nachhaltigkeit gesichert.

---

67 Vgl. Aschenbrenner; Pempe (2010).
68 Das sog. OAI-ORE / ATOM Experiment (vgl. DARIAH (2010), S. 18).
69 Vgl. DARIAH (2011).
70 Vgl. ESFRI (2011a), zur Roadmap siehe ESFRI (2011b).
71 Vgl. DARIAH Homepage: http://de.dariah.eu/.
72 Vgl. CLARIN (2011).

Was die finanzielle Nachhaltigkeit anlangt, so sei ergänzend zu den Ausführungen im Abschnitt *Interne Organisation* (s.o.) auf die Bund-Länder-Problematik hingewiesen. Jede Finanzierung, bei der Infrastruktur-Ressourcen aus Landesmitteln bezahlt werden, birgt die Gefahr, dass Nutzer aus anderen Bundesländern – oder gar anderen EU-Staaten – keinen Zugriff auf diese Ressourcen erhalten. Der WissGrid-Report *Konzept eines Betriebsmodells für Virtuelle Forschungsumgebungen* formuliert das Problem wie folgt: „Landeseigene Investitionen können i.d.R. nicht bundesweit genutzt werden, da sie durch Landesmittel finanziert werden. [...] Dementsprechend sind landesgeförderte Ressourcen nur eingeschränkt für Virtuelle Forschungsumgebungen geeignet."[73] D.h., dass für langfristige Finanzierungsmodelle Lösungen gefunden werden müssen, die die (internationale) Zusammenarbeit in verteilten Forschungsinfrastrukturen und die übergreifende Nutzung digitaler Langzeitarchive nicht aufgrund formaljuristischer Regelungen unnötig einschränken.

Doch was nützen selbst nachhaltig betriebene Langzeitarchive, wenn sie nicht konsequent zur Aufbewahrung und Publikation von Forschungsdaten genutzt werden? Ein grundsätzliches Problem besteht nach wie vor darin, bei den Forschenden das individuelle Bewusstsein für die Notwendigkeit zu schaffen, dass die „eigenen" Forschungsdaten langfristig zur Nachnutzung zur Verfügung gestellt werden. Hier muss einerseits der Mehrwert sichtbar gemacht werden, der aus der Nachnutzung auch der Daten anderer erwächst, sowie die Motivation geschaffen und entsprechende Belohnungssysteme etabliert werden, die den damit einhergehenden Mehraufwand, z.B. die Daten standardkonform zu beschreiben, sowie den „Herrschaftsverlust", wenn die Daten publiziert sind, entsprechend honorieren.

Generell lässt sich jedoch die Tendenz feststellen, dass Wissenschaftlerinnen und Wissenschaftler in den Geistes- und Kulturwissenschaften vermehrt in interdisziplinären Forschungsprojekten zusammenarbeiten. Durch die zunehmende elektronische Verfügbarkeit von Forschungsdaten, spezialisierten Werkzeugen und durch die Vernetzung von Repositorien eröffnen sich auch neue Möglichkeiten geisteswissenschaftlicher Forschungsarbeit, gerade im Bereich quantitativer Analysen, Data-Dri-

73 WissGrid (2010), S. 16.

ven Research, Text Mining etc. Mit der Einbindung von LZA-Diensten und -Repositorien in die Arbeitsabläufe, die sich mittlerweile in den eHumanities etablieren, kann bei den Forschenden auch die Akzeptanz neuer Arbeitsweisen und Publikationsformen sowie das Bewusstsein für die Notwendigkeit, Forschungsdaten zu veröffentlichen und langfristig nachnutzbar zu halten, geschaffen werden. Die Integration einer Virtuellen Forschungsumgebung mit einem digitalen Langzeitarchiv im Fall von TextGrid könnte hier, ganz im Sinne der im Einleitungskapitel formulierten Empfehlung, „eine intuitiv nutzbare Infrastruktur zur Verfügung zu stellen […] und damit die Entwicklung einer neuen Publikationskultur von Forschungsdaten zu unterstützen“[74], Vorbildcharakter haben.

74 Vgl. Kap. 1, S. 8.