WEBPLATTFORM FÜR DIE BEARBEITUNG, PUBLIKATION UND LANGZEITARCHIVIERUNG VON REGIONAL-WISSENSCHAFTLICHEN FORSCHUNGSDATEN (LAZAR)

Dr. Elguja Dadunashvilia[a] , Dr. Jakob Voß[b]

[a] Institut für Slawistik und Kaukasusstudien, Friedrich-Schiller Universität Jena / Staatliche Ilia-Universität Tiflis, Georgien, elguja.dadunashvili@uni-jena.de,
[b] Verbundzentrale des GBV, Deutschland, jakob.voss@gbv.de

KURZDARSTELLUNG: Der vorliegende Beitrag stellt die Ziele und Ergebnisse des seit Ende 2015 von der DFG geförderten Projekt LaZAR vor. Darin wird eine Plattform entwickelt um in regionalwissenschaftlicher Feldforschung erhobene Video-, Audio- und Bildmaterialien zugänglich und zitierfähig zu machen. Die Aufnahme der Forschungsdaten erfolgt über ein webgestütztes Erfassungssystem, das nicht nur die Möglichkeit bietet, die Mediendateien beim Hochladen zu bearbeiten und mit Metadaten zu versehen, sondern auch in der Lage ist, die Inhalte über standardisierte Schnittstellen zur Nachnutzung und Langzeitarchivierung zur Verfügung zu stellen. Erläutert werden die Anforderungen an die Plattform und die sich daraus ergebenen Komponenten.

1. EINFÜHRUNG

Die Erfassung, Publikation und Archivierung von Forschungsdaten, das heißt Daten die im Rahmen einer Forschungstätigkeit anfallen, wird in zunehmendem Maße als Bestandteil guter wissenschaftlichen Praxis vorausgesetzt. Das *Registry of Research Data Repositories* verzeichnet viereinhalb Jahre nach seiner Gründung mehr als 2000 institutionelle, fach-spezifische und -übergreifende Repositorien [1, 2]. Während Forschungsdaten zunächst vor allem in den Natur- und Lebenswissenschaften relevant waren, spielen sie mit der steigenden Verbreitung von Digital Humanities und reproduzierbarer Forschung zunehmend auch in den Geistes- und Sozialwissenschaften eine Rolle. Die Vielzahl von Repositorien zeigt, dass deren Aufbau auch von fachspezifischen Anforderungen bestimmt wird. Nicht zuletzt kommt es darauf an wie die Daten zustande

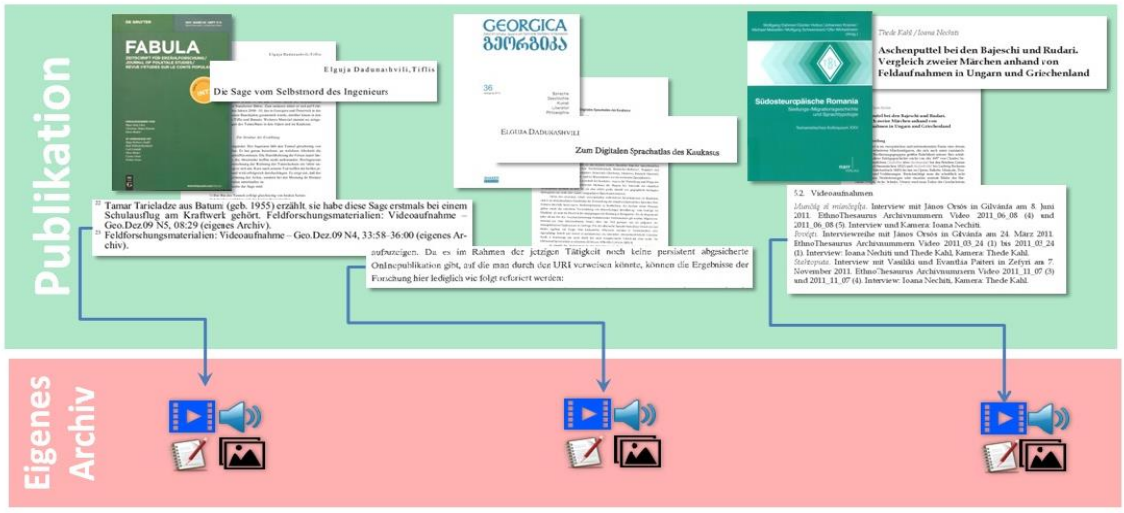

Abb. 1: *Multimedia Zitate in wissenschaftlichen Publikationen (gegenwärtige Praxis)*

kommen, bearbeitet und genutzt werden.

In den Regionalwissenschaften ermöglicht die technische Entwicklung eine neue Forschungspraxis, die mit einer immensen Zunahme von Bild-, Video- und Audioaufnahmen einhergeht. Diese Medien können auch umfangreicher und besser in wissenschaftliche Publikationen eingebettet werden als dies bisher möglich war. Zu nennen sind hier vor allem Film und Audiomaterialien für deren Publikation das Druck- bzw. Textformat nicht brauchbar ist. Während Videoaufnahmen für die Feldforschung bis in die 80er Jahre des letzten Jahrhunderts als ein besonderes Privileg der Forschungseinrichtungen industriell hochentwickelter Länder galt, sind die technische Mittel für Bild- und Tonaufnahmen inzwischen fast alltägliche Werkzeuge. Unter diesen Umständen ist nicht verwunderlich, dass die Mehrheit von Vertretern regionalwissenschaftlicher Disziplinen ihre Forschungsdaten heute vor Allem in Form multimedialer Daten produzieren.

Die Zunahme multimedialer Forschungsdaten geht einher mit einer Zunahme ihrer Zitierung in wissenschaftlichen Publikationen. Wie in Abbildung 1 dargestellt, werden Video- und Audioaufnahmen häufig mittels Fußnoten referenziert, die auf Privatarchive verweisen. Der Zugriff auf diese Medien ist also praktisch nur per Kontaktaufnahme mit den Autoren möglich. Eine gelegentliche Alternative ist die Beifügung von Datenträgern zu gedruckten Büchern. Viel effektiver ist die Bereitstellung von Forschungsdaten im Internet. Die Publikation in Plattformen wie YouTube oder Dropbox ermöglichst zwar den momentanen Zugriff auf die Daten, erfüllt aber nicht die folgenden wesentlichen Anforderungen an veröffentlichte Forschungsdaten:

(a) einheitliche, detaillierte Erschließung

(b) Schutz vor nachträglichen Änderungen

(b) Vermeidung rechtlicher Probleme

(c) Dauerhafte Zugänglichkeit (Archivierung)

(d) Möglichkeit der Migration und Emulation (Langzeitarchivierung).

In Ermangelung eines Repositoriums für regionalwissenschaftliche Feldforschungsdaten wurde daher im Projekt LaZAR nach diesen Maßgaben ein entsprechende Plattform konzipiert und umgesetzt.

2. KONZEPT UND UMSETZUNG

Angesichts des Bedarfs für den gesicherten Umgang mit Forschungsdaten in den Regionalwissenschaften wurde 2014 damit begonnen, gemeinsam mit dem Institut für Slawistik und Kaukasusstudien der Friedrich-Schiller-Universität Jena, der Verbundzentrale des GBV (VZG) und der Technischen Informationsbibliothek Hannover (TIB) ein Konzept für regionalwissenschaftliche Forschungsdaten zu entwickeln. Die Vorarbeiten führten zu einem Förderantrag bei der Deutschen Forschungsgemeinschaft (DFG), der in zwei Phasen bewilligt wurde. Das Projekt *Webplattform für die Bearbeitung, Publikation und Langzeitarchivierung von regionalwissenschaftlichen Feldforschungsdaten* (LaZAR) läuft seit Dezember 2015 bis voraussichtlich Ende 2018 [3].

2.1 ZIELE UND BESTANDTEILE

Das Hauptziel des LaZAR-Projekt besteht darin, einem weiten Forscherkreis die in regionalwissenschaftlichen Feldforschungen erhobenen Video-, Audio- und Bildmaterialien zugänglich und zitierfähig zu machen. Die Aufnahme der Forschungsdaten erfolgt über ein webgestütztes Erfassungssystem, das nicht nur die Möglichkeit bietet, die Mediendateien beim Hochladen zu bearbeiten und mit Metadaten zu versehen (Weblaboratorium), sondern auch in der Lage ist, die Inhalte über standardisierte Schnittstellen zur Nachnutzung zur Verfügung zu stellen (Repositorium). Insbesondere besteht eine direkte Verbindung zu einem Langzeitarchivierungsdienst (LZA), der durch die TIB bereitgestellt wird.

Bis Mitte 2017 konnte die zur Erreichung der Projektziele notwendige Infrastruktur mit folgenden Komponenten konzipiert und in wesentlichen Teilen umgesetzt werden:

- Datenmodell zur Erschließung regionalwissenschaftlicher Forschungsdaten
- Weblaboratorium für die Bearbeitung der Forschungsdaten

- Repositorium für die Speicherung und Recherche in den Forschungsdaten
- Zugriffs- und Rechtemodell für die Nutzerverwaltung
- Erarbeitung von Grundlagen für die rechtkonforme Erhebung, Veröffentlichung und Nachnutzung der Daten
- Exportschnittstellen für die Inhalte des Repositoriums (OAI-PMH und LOD)
- Langzeitarchiv für ausgewählte veröffentlichte Primärdaten
- Schnittstelle zum Ingest zwischen Repositorium und Langzeitarchiv
- Exit-Strategie für das Langzeitarchiv zum Export in andere Systeme

2.2 DATENMODELL

Das Datenmodell wurde zu Beginn des Projekts in Anlehnung an das DataCite Metadatenformat konzipiert und im Laufe der Umsetzung angepasst [4]. Zusätzlich weist das LaZAR-Datenmodell zwei Besonderheiten auf:

- Konsequenter Einsatz von Normdaten (GND, ORCID, Geonames…)
- Unterstützung hierarchischer Datensätze mit Kollektionen und Unterkollektionen

Der Einsatz von Normdaten zur Erschließung von Forschungsdaten ist zwar sehr sinnvoll aber bislang noch eher die Ausnahme. Notwendig sind Normdaten insbesondere zur Verknüpfung mit anderen Datenbeständen. Dazu müssen alle wesentliche Entitäten (Personen, Orte, Themen…) über einen eindeutigen Identifier referenzierbar sein.

Das hierarchische Datenmodell mit Kollektionen und Unterkollektionen (in LaZAR als *Konvolute* bezeichnet) lehnt sich an Dateisysteme mit Ordnern und Unterordnen an. Alle Medienobjekte und daraus abgeleitete Dateien (z.B. Transkriptionen) sind eigene dokumentarische Bezugseinheiten, die einzeln referenziert werden können. Jede Dateien ist allerdings einem Konvolut zugeordnet, das wiederum Bestandteil anderer Konvolute sein kann. Die oberste Einheit bilden die Wurzelkonvolute. Darüber hinaus können Objekte untereinander verknüpft werden, so dass auch multi- und nicht-hierarchische Beziehungen möglich sind. Zur Zitierung von in LaZAR veröffentlichten Objekten kommen somit zwei Varianten in Frage (Tabelle 1):

Wurzelkonvolut	Autor (Jahr): *Wurzelkonvolut*, LaZAR, Version, DOI
Datei oder Unterkonvolut	Autor (Jahr): Titel. In: Autor, *Wurzelkonvolut*, LaZAR, Version, DOI

Tabelle 1: *Zitierung von veröffentlichten Objekten*

2.3 TECHNISCHE UMSETZUNG

Das Weblaboratorium und das Repositorium zur Bearbeitung, Erschließung und Publikation der Forschungsdaten ist als eine Anwendung auf Basis der Software *easyDB 5* umgesetzt, die auch für andere Datenbanken an der VZG eingesetzt wird [5]. Die Unterstützung von Normdaten basiert auf dem Normdatendienst DANTE der VZG, für dessen API entsprechende Plugins entwickelt wurden [6].

Das Langzeitarchiv ist eine unabhängige Anwendung, die von LaZAR durch den Export ausgewählter publizierter Daten gespeist wird. Für die Präsentation ist eine zusätzliche Sicht auf publizierte Daten und ihre Bereitstellung als Linked Open Data geplant. Die gesamte Infrastruktur von LaZAR besteht somit aus den vier miteinander gekoppelten Anwendungen: easyDB, DANTE, LOD-Sicht und LZA.

2.4 AKTEURE UND RECHTEMODELL

Bei der Erschließung der Forschungsdaten wird zwischen den Urhebern und Herausgebern eines Konvoluts und den Produzenten einer Datei unterschieden. Letztere stellen konkrete Forschungsdaten her, indem sie beispielsweise im Feld Foto-, Video- und Audioaufnahmen machen, Protokolle und Berichte verfassen, Skizzen und Pläne anfertigen sowie andere Daten annotieren. Der Urheber eines Konvoluts agiert gleichzeitig als Herausgeber indem er oder sie Dateien strukturiert und mit Metadaten versieht. Bei Unterkonvoluten können weitere Herausgeber als Co-Autor des Oberkonvoluts zusammenkommen.

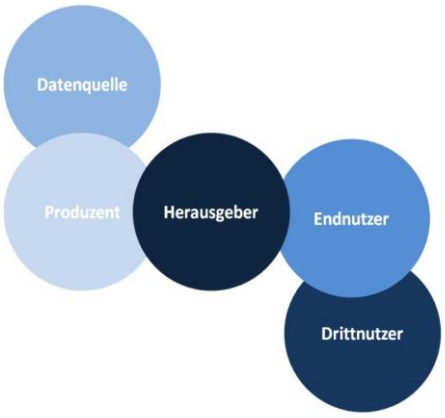

Abb. 2: *Beziehungen zwischen dem Herausgeber und anderen Beteiligten*

Hinsichtlich des Rechtemanagements dienen Herausgeber gleichzeitig als nicht-technische Administration um neue Benutzer freizuschalten und ihnen Bearbeitungsrechte an eigenen Konvoluten zuzuweisen. Dabei können sich Wissenschaftler selbst als Beitragende registrieren während die Recherche in publizierten Dokumenten ohne Registrierung möglich ist.

2.5 RECHTLICHE ASPEKTE

Neben der Gewährleistung der Authentizität kommen insbesondere bei regionalwissenschaftlichen Forschungsdaten weitere Rechtliche Aspekte hinzu, deren Bearbeitung ebenfalls Bestandteil des LaZAR-Projekt ist. Zu berücksichtigen sind mit der Freigabe der Forschungsdaten verbundene urheberrechtliche, ethische und Datenschutzrechtliche Fragen. Wie in Abb. 2 ersichtlich muss keine direkte Beziehung zwischen dem Herausgeber und den nicht selten Personenbezogenen Datenquellen bestehen. Ebensowenig kann eine direkte Beziehung zwischen Herausgeber und Nutzern der Daten vorausgesetzt werden. Aus diesem Grund sind klare Vereinbarungen in Form von Lizenzverträge notwendig, die Herausgeber mit den Produzenten und Endnutzern eingehen. Soweit möglich werden dafür Creative-Commons-Lizenzen (CC) eingesetzt, insbesondere für die Metadaten. Hinzu kommen Empfehlungen zur im Rahmen von Feldforschungsarbeiten durchzuführenden Datenerhebungen.

Im Laufe der Projekt wurden folgende Vorschriften und Empfehlungen erarbeitet:

- Allgemeine Geschäftsbedingungen mit Hinweisen auf Urheberrecht, Rechte und Pflichte der Beteiligten sowie Nutzungs- und Datenschutzbestimmungen
- Workflow zu Erfassung, Anonymisierung bzw. Pseudonymisierung, Speicherung und Pflege personenbezogener Daten soweit für die Sicherung von Rechten der Produzenten, Informanten und Mitwirkenden sowie den bibliographischen Nachweis bzw. Publikationen notwendig
- Vertrag und Workflow zur Übermittlung der für den öffentlichen Gebrauch gesperrten Forschungsdaten
- Empfehlungen für Vorbereitung und Upload von Ursprungsdateien und Derivaten von Video- und Audiodateien

Für die Aufnahme der Daten ins Repositorium setzt der Herausgeber die Überprüfungsmöglichkeiten der Authentizität der Mediendaten voraus. Falls eine Datei für die öffentliche Freigabe nicht geeignete Segmente enthält, die gesperrt, oder anonymisiert werden müssen, müssen von dieser Masterdatei mehrere Segmente ausgeschnitten werden, für die jeweils eigene Zugriffsrechte angegeben werden können. Die einzelnen Segmente werden dabei durch entsprechende Relationen (part-of, derived-from…) mit der Masterdatei in Verbindung gesetzt.

2.6 ZUGRIFFSRECHTE

Während die Metadaten veröffentlichter Daten immer frei sind, sind für die Mediendateien folgende Zugriffsrechte möglich:

(a) Freier Zugang zu den Daten (CC-Lizenzen)

(b) Zugang nach ausdrücklichen Zustimmung

(c) Betrachtungsmöglichkeit der Daten an einer im Archiv eingerichteten Stelle [7]

Generell räumt der Produzent bzw. Urheber Nutzungsrechte über die von ihm für die Publikation und Archivierung übergebenen digitalen Objekte und Metadaten durch einen Vertrag ein, der beim Beitrag eigener Forschungsdaten in LaZAR als Bestandteil der allgemeinen Geschäftsbedingungen in Kraft tritt. Im Regelfall sind die weiteren Nutzungsrechte durch die verwendeten CC-Lizenzen geregelt. Der oder die Herausgeber verpflichten sich ihrerseites zur

(a) Gewährleistung eines Arbeitsortes beim Verleger für die Betrachtung der digitalen Objekten die nur vor Ort einsehbar sind

(b) Aushandlung von Lizenzverträgen zur Einräumung der Nutzungsrechte über die digitalen Objekte mit vertraulichen Inhalten [8]

LaZAR macht den Abruf der nicht für den offenen Gebrauch geeigneten Daten (Fall b) nur nach Zustimmung eines entsprechenden Lizenzvertrags möglich. Der wichtigste Punkt eines solchen Vertrages ist eine Verschwiegenheitsklausel. Durch das Einfügen dieser Klausel wird der jeweiliger Endnutzer dazu verpflichtet, die erlangten Forschungsdaten nicht bzw. nur innerhalb seines Forschungsprojektes weiterzugeben.

2.7 EXPORT UND VERSIONIERUNG

Der Export von Datensätzen ist über die Weboberfläche des Repositoriums, per OAI-PMH und Linked Open Data möglich (letzteres ist für 2018 geplant). Als weitere Funktion wird derzeit eine Versionsverwaltung umgesetzt, so dass Bearbeitungen an veröffentlichten Datensätzen als zusätzliche Version veröffentlicht werden können.

5. LITERATURHINWEISE

[1] Pampel, H. et.al. (2015). Stand und Perspektive des globalen Verzeichnisses von Forschungsdaten-Repositorien re3data.org. In P. Müller et. al. (Hrgs.), 8. DFN-Forum Kommunikationstechnologien Bonn: GI http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1169893:3

[2] https://www.re3data.org/metrics/types

(Stand 2017-10-08).

[3] Eintrag in der DFG-Projektdatenbank: http://gepris.dfg.de/gepris/projekt/277236902

[4] DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.0 September 2016. http://doi.org/10.5438/0012

[5] https://www.programmfabrik.de/easydb/ (Stand 2017-10-08)

[6] Dokumentation der Schnittstelle zu DANTE. http://api.dante.gbv.de/ (Stand 2017-10-08)

[7] Upmeier, Arne: Rechtliche Aspekte. In: Hrsg.: H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, H. Karsten, *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, nestor, Göttingen, 2010, Version 2.3, Kap. 16.2, S. 9

[8] Hillegeist, Tobias: Rechtliche Probleme der elektronischen Langzeitarchivierung wissenschaftlicher Primärdaten, Universitätsverlag Göttingen, 2012, S. 78-82