Konzeption und Institutionalisierung des FDM

—

aus der Erfahrung eines Forschungsprojekts in den digitalen Geisteswissenschaften

Stephan Lücke (ITG), Martin Spenger (UB der LMU)

Dieser Beitrag zeigt am Beispiel des Forschungsdatenmanagements an der LMU, wie Wissenschafts- und Infrastrukturpartner erfolgreich zusammenarbeiten können. Im Rahmen des Modellprojekts „eHumanities – interdisziplinär" wird die Zusammenarbeit anhand des Pilotpojekts VerbaAlpina verdeutlicht. Der erste Teil des Aufsatzes beschreibt VerbaAlpina und die Besonderheiten aus der Perspektive der digitalen Geisteswissenschaften. Der zweite Teil bringt als Infrastrukturpartner die Universitätsbibliothek der LMU ins Spiel. Abschließend folgt ein kurzer Überblick über das Projekt „eHumanities – interdisziplinär".

1. Teil: Das Projekt VerbaAlpina

Das Projekt VerbaAlpina[1] (VA) ist ein von der DFG gefördertes Langfristvorhaben[2] mit einer Perspektive bis 2025 und befindet sich derzeit in der zweiten Förderphase, die noch bis zum Herbst 2020 andauert. Es handelt sich um ein interdisziplinäres Projekt im Umfeld der Digital Humanities. Der Schwerpunkt des Interesses liegt auf den Sprachwissenschaften, daneben erfolgt jedoch auch eine intensive und streckenweise exemplarische Auseinandersetzung mit den Herausforderungen der Informationstechnologie. Das von VA zusammengetragene und analysierte Datenmaterial kann darüber hinaus auch für andere Disziplinen wie etwa die Ethnographie, die Archäologie oder auch die Geschichtswissenschaften von Interesse sein. VA betreibt unter anderem eine intensive Methodenreflexion, die hauptsächlich in der Rubrik „Methodologie" auf der Projektwebseite dokumentiert ist. Zu einigen der im Folgenden thematisierten Punkte finden sie dort ausführlichere Darlegungen, auf die hier generell hingewiesen sei.

Initiatoren und Träger des Projekts sind Thomas Krefeld vom Romanistischen Seminar der LMU sowie Stephan Lücke von der IT-Gruppe Geisteswissenschaften der LMU. In der laufenden zweiten Förderphase verfügt VA über je zwei Doktorandenstellen im Bereich der Sprachwissenschaft (einer zuständig für die Romanistik/Slawistik, der andere für die Germanistik) und in der Informatik (für Datenbank- und Frontendentwicklung). Für die umfangreiche Koordinationsarbeit mit den zahlreichen Projektpartnern steht eine weitere Doktorandenstelle zur Verfügung, deren Inhaberin ihre Doktorarbeit im thematischen sprachwissenschaftlichen Umfeld des Projekts vorbereitet. Unterstützt wird das Team durch eine Reihe von Hilfskräften, die vor allem für die strukturierte Datenerfassung eingesetzt werden.

Vorrangiges Ziel von VA ist die systematische Erfassung und Analyse der im Alpenraum verbreiteten morpholexikalischen Typen, die zur Bezeichnung ausgewählter „Objekte" (Konzepte[3], Begriffe) Verwendung finden oder auch fanden. Vereinfacht gesagt, steht dabei die Frage im Zentrum, welche Konzepte an welchen Orten mit welchen Wörtern bezeichnet werden. So wird z.B. das Konzept BUTTER (VA verwendet zur Bezeichnung der Konzepte

[1] http://www.verba-alpina.gwi.uni-muenchen.de/
[2] http://gepris.dfg.de/gepris/projekt/253900505
[3] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=K#37

stets Versalien, um damit den Unterschied zu den Bezeichnungen/Wörtern klar zu machen) in unterschiedlichen Regionen des Alpenraums mit unterschiedlichen morpholexikalischen Typen bezeichnet: In Bayern und Österreich herrscht der Typ *Butter* vor, im Alemannischen ist *Anke* weit verbreitet, in Italien nennt man die BUTTER *burro*. VA ist fokussiert auf die jeweiligen morpholexikalischen Typen, das heißt, phonetische Variationen werden zwar vielfach dokumentiert, jedoch nicht systematisch erfasst oder gar analysiert. Eine vollständige Erfassung des morpholexikalischen Materials ist unmöglich. Daher beschränkt VA seine Dokumentation und Untersuchung auf typisch alpine Konzeptdomänen wie etwa die Almwirtschaft (Schwerpunkt der ersten Projektphase von 2014 bis 2017) oder auch Natur und Umwelt (laufende Projektphase) sowie Tourismus (geplant für die Projektphase ab 2020).

Der geographische Rahmen des Untersuchungsgebiets ist aus pragmatischen Gründen auf die Ausdehnung der sogenannten Alpenkonvention[4] festgelegt worden. Als Datengrundlage dienen in erster Linie traditionell in Buchform publizierte sogenannte Sprachatlanten[5], die konzeptorientiert („onomasiologisch“) in Kartenform die Verbreitung von morpholexikalischen Typen für die Bezeichnung vorgegebener Konzepte präsentieren. Die systematische Erfassung dieser Daten ist kaum automatisierbar und bedeutet einen erheblichen Aufwand an Handarbeit. Die Schwierigkeit liegt dabei weniger im (nicht eingesetzten) OCR-Verfahren, sondern vielmehr an der kartographischen Zuordnung bestimmter Bezeichnungen zu den einzelnen auf den Karten verzeichneten Punkten. Erschwert wird diese automatische Erfassung überdies durch die Verwendung von Symbolen auf der Karte, die die Orthographie des entsprechenden Typs also unterdrücken. Die Daten aus den Sprachatlanten werden ergänzt um Daten aus Wörterbüchern, die, anders als die Sprachatlanten, von den morpholexikalischen Typen ausgehen und die jeweiligen durch sie bezeichneten Konzepte dokumentieren. Es werden jedoch nur solche Wörterbücher berücksichtigt, die auch Aufschluss über die geographische Verbreitung der jeweiligen Bedeutungen geben. VA verfügt über eine große Anzahl nationaler und internationaler Partner[6], die vielfach über eigene Sprachdatensammlungen verfügen, die nach Möglichkeit und Eignung ebenfalls in den Datenbestand von VA übernommen werden. In der Summe ergibt sich ein kartierbares Netz der im Alpenraum verbreiteten Bezeichnungen der ausgewählten Konzepte, die auf einer interaktiven Online-Karte[7] präsentiert werden. Ein wesentlicher Mehrwert des Einsatzes der zeitgemäßen Online-Medien ist dabei der problemfreie und schnelle Wechsel zwischen der onomasiologischen und der semasiologischen Perspektive.

VA stellt sämtliche von ihm erzeugten Inhalte – wozu auch alle Softwareentwicklungen gehören – nach Möglichkeit unter einer offenen Lizenz[8] zur Verfügung (CC BY-SA). Ausnahmen bestehen nur für Daten, die anderweitig unter restriktiven Lizenzen geführt werden.

VA versteht sich als konsequent digitales[9] Online-Projekt und betrachtet die traditionellen Publikations- und Kommunikationsgepflogenheiten im Wissenschaftsbetrieb als überholt. Das Projekt verzichtet vollständig auf den Einsatz herkömmlicher Drucktechnologie und hält auch den Einsatz von PDF-Dokumenten als eine wenig geeignete Publikationsform, die gegenüber der Webtechnologie entscheidende Einschränkungen besitzt.

---

[4] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=A#103
[5] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=Q#48
[6] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=185&db=181
[7] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=133&db=xxx
[8] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=L#41
[9] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=182&letter=D#15

Die zentralen Projektdaten werden im relationalen Datenformat in einer MySQL-Datenbank verwaltet. Als Frontend dient eine generische WordPress-Installation, für die von den Projekt-Informatikern eine Reihe von spezifischen, auf die Projektbedürfnisse zugeschnittenen, Plugins entwickelt wurden, die über GitHub[10] der Allgemeinheit unter der CC BY-SA-Lizenz zur Verfügung gestellt werden. Das Projektportal ist multifunktional. Es dient gleichermaßen als Arbeitsinstrument für die Projektmitarbeiter wie auch als Publikations- und Dokumentationsplattform und schließlich zur wissenschaftlichen Kommunikation, wobei allerdings letztere Funktionalität noch nicht konsequent ausgebaut ist. Die Idee ist, dass sich Wissenschaftler und Laien auf dem Portal registrieren und das System für den wissenschaftlichen Austausch und auch als Instrument zur Verwaltung eigener Daten verwenden.

VA unterscheidet nicht zwischen „Forschungsdaten" einerseits und auswertenden Daten andererseits, eine Vorstellung, von der offenkundig auch die aktuelle Diskussion des Forschungsdatenmanagements geprägt ist. Sämtliche Projektdaten sind aufeinander bezogen und somit untrennbar miteinander verbunden. Aus diesem Grund verfolgt VA das Ziel, die Gesamtheit der Daten, also die gesammelten Sprachdaten sowie alle darauf bezogenen analytischen und erläuternden Texte, kartographischen Repräsentationen und sonstige Derivate als Ganzes dauerhaft und in zuverlässig zitierbarer Weise zu erhalten. Angesichts ständig ausgebauter Speicherungskapazitäten und dem im Vergleich mit mancher naturwissenschaftlichen Disziplin geringem Datenvolumen kann die Datenmenge nicht als Hinderungsgrund betrachtet werden, womit kein Grund erkennbar ist, nicht den kompletten Datenbestand dauerhaft zu bewahren.

Der aus den stets nur auf Teilbereiche des Alpenraums beschränkten Sprachatlanten und Wörterbüchern zusammengetragene gleichsam „historische" Datenbestand weist notwendigerweise eine ganze Reihe von Inkonsistenzen auf. So sind regionale Unterschiede hinsichtlich Belegdichte und Dokumentation der ausgewählten Konzepte festzustellen. Um diese Inkonsistenzen auszugleichen, wurde von VA ein Crowdsourcing-Tool entwickelt, das Nutzern im Internet erlaubt, regionale Bezeichnungen für die projektrelevanten Konzepte beizutragen. Aktuell sind auf diesem Weg exakt 11546 (Stand 30.4.19) morpholexikalische Einzelbelege in den Datenbestand von VA gelangt. Neben dem geographischen und konzeptbezogenen Ausgleich gelangt über das Crowdsourcing auch eine (zusätzliche, da auch die konventionellen Datenquellen unterschiedliche Zeiträume dokumentieren) diachrone Perspektive in den Datenbestand, die einen Einblick in den Sprachwandel und dessen Dynamik erlaubt.

VA steht im engen Verbund mit im Wesentlichen zwei an der LMU beheimateten Institutionen: Der IT-Gruppe Geisteswissenschaften[11] (ITG) sowie der Universitätsbibliothek[12] (UB). Die ITG besteht seit knapp 20 Jahren und ist ursprünglich aus einer Stelle für rechnergestützte Forschung an der Fakultät für Kulturwissenschaften der LMU hervorgegangen. Sie ist zuständig für und ist getragen von sämtlichen geisteswissenschaftlichen Fakultäten, wobei ihre zentralen Aufgaben in der Planung und Betreuung der IT-Infrastruktur, der Unterstützung bei und der Durchführung von digitaler Forschung und Lehre sowie im Management der von den Wissenschaftlern erzeugten Forschungsdaten liegen. Die bei VA beschäftigten Informatiker sind direkt an der ITG angesiedelt und haben dort die Möglichkeit zum fachlichen Austausch mit anderen Beschäftigten, die in einer ganzen Reihe von weiteren DH-Projekten mit vergleichbaren

[10] https://github.com/VerbaAlpina/
[11] https://www.itg.uni-muenchen.de/index.html
[12] https://www.ub.uni-muenchen.de/index.html

Aufgaben beschäftigt sind. Dieses strukturelle Konzept gewährleistet ein hohes Maß an Synergieeffekten, von denen auch die anderen an der ITG betriebenen Projekte profitieren können. VA ist bestrebt, die für das Projekt entwickelte Software soweit möglich und sinnvoll generisch und modular zu konzipieren, so dass sie mit möglichst geringem Aufwand auch in anderen Kontexten weiter- bzw. wiederverwendet werden kann.

Vor allem im Hinblick auf die dauerhafte Bewahrung der Projektdaten spielt wiederum die UB die entscheidende Rolle für VA. Generell ist VA der Meinung, dass vor allem die Bibliotheken die natürlichen Ansprechpartner für alle Fragen der Datenbewahrung sind. Dies ist begründet zum einen durch die jahrhundertelange Tradition der entsprechenden Zuständigkeit, zum anderen durch die feste institutionelle Verankerung, die eine langfristige Bestandsgarantie verspricht, wie sie kaum eine andere Einrichtung in vergleichbarem Umfang besitzt. Hinzu kommt ein hohes Maß an Kompetenz, das an der UB sowohl im Hinblick auf die bibliothekarischen wie auch die informatischen Erfordernisse vorhanden ist. Drittmittelgeförderte Projekte mit begrenzter Existenzperspektive erscheinen problematisch, werden von VA jedoch zusätzlich genutzt. So wird aktuell ein Datenexport an das CLARIN-D Centre Leipzig[13] vorbereitet, mit dem VA eine Kooperationsvereinbarung abgeschlossen hat.

Über den Kontakt zur UB und ITG ist VA auch in das vom Bayerischen Staatsministerium für Wissenschaft und Kunst geförderte Projekt eHumanities – interdisziplinär[14] eingebunden. VA nimmt dort die Rolle eines Pilotprojekts ein, dessen Daten exemplarisch mit Metadaten angereichert und schließlich in das institutionelle Repositorium der UB (Open Data LMU[15]) gelangen, wo sie schließlich in versionierter Form dauerhaft gesichert und auch zitierbar sind. Die Entwicklung bzw. Anpassung der Metadatenmodelle an die spezifischen Projekterfordernisse erfolgt in enger Zusammenarbeit von VA-, ITG– und UB-Mitarbeitern. Dieser Prozess, der sich als ausgesprochen effizient erweist, ist in dieser Form nur durch die enge lokale Ansiedlung der beteiligten Institutionen möglich. Vor dem Hintergrund dieser Erfahrung steht VA allen Konzepten, die im Hinblick auf das Forschungsdatenmanagement Lösungen mit spezialisierten zentralen Institutionen favorisieren, die dann unter Umständen sehr weit vom Ort der Projekttätigkeit liegen, skeptisch gegenüber. Demgegenüber betrachtet VA die skizzierte enge Verzahnung der beteiligten Akteure als modellhaft. Zumindest theoretisch sollte eine Übertragung dieses Konzepts auf andere Universitätsstandorte angesichts der doch weitgehend flächendeckenden Verbreitung von Universitätsbibliotheken möglich sein.

VA steht seit Längerem auch in Kontakt mit dem vom BMBF geförderten Projekt GeRDI, das als Datenaggregator betrachtet werden kann, dessen Ziel es ist, Forschungsdaten der verschiedensten Disziplinen über einen zentralen Zugang unter Einsatz von Metadaten zugänglich zu machen. Durch die oben beschriebene Kooperation im Rahmen des Projekts eHumanities – interdisziplinär gelangen die VA-Daten über die UB auch in den Datenbestand von GeRDI[16].

Im Zusammenhang mit dem Forschungsdatenmanagement wird für die Aufbereitung von Forschungsdaten seit einiger Zeit auch die Erfüllung der sog. FAIR-Prinzipien propagiert bzw. bisweilen auch gefordert. Das Projekt VA hält die in diesem Akronym versammelten Postulate für durch und durch berechtigt und ist bestrebt, diesen Kriterien möglichst

[13] https://clarin.informatik.uni-leipzig.de/repo/
[14] https://www.fdm-bayern.org/
[15] https://data.ub.uni-muenchen.de/
[16] https://www.gerdi-project.eu/

weitgehend zu entsprechen. Die Auffindbarkeit (findable) und Zugänglichkeit (accessible) der VA-Daten ist durch die in Zusammenarbeit mit der UB erfolgte Anreicherung um Metadaten mit deren anschließender Einbindung in Kataloge (z.B. OPAC) und Aggregatoren-Dienste (GeRDI) sowie durch das angewendete offene Lizenzmodell hinreichend gewährleistet. Die Forderung der Interoperabilität (interoparable) und bis zu einem gewissen Grad auch der Nachnutzbarkeit (reusable) erscheint nur möglich, wenn das vom Projekt gesammelte Datenmaterial in möglichst feiner Granulierung vorliegt und auf die einzelnen Datensätze über eine URL eindeutig referenziert werden kann. Aus diesem Grund wird der Kerndatenbestand von VA nach Einzelbelegen, morpholexikalischen Typen, Konzepten und Gemeinden gruppiert, die jeweiligen Gruppen mit persistenten Identifikatoren versehen und in dieser Form, zusammen mit allen übrigen Projektdaten, versionsweise an die UB übertragen. Nach der Anreicherung um Metadaten und der Ablage im institutionellen Repositorium ist es sodann möglich, jede einzelne Instanz innerhalb der genannten Gruppierungen über eine DOI anzusprechen. Damit sind de facto projektspezifische Normdaten erzeugt, darüberhinaus ist eine der wesentlichen Forderungen erfüllt, die den Datenbestand von VA als „linked open data“ qualifizieren (die Existenz einer persistenten URL). Derzeit fehlen jedoch noch die ebenfalls erforderlichen RDF-Metadaten im XML-Format, deren Erzeugung jedoch geplant ist. Die UB erwägt außerdem, zusätzlich zu den DOIs eigene persistente URLs zu erzeugen, deren Vergabe und Betreuung in der alleinigen Verantwortung der UB liegen. VA begrüßt diese Perspektive, zumal die VA-Ressourcen zusätzlich zu den DOIs über ein weiteres, davon unabhängiges System persistenter Adressen erreichbar sein werden.

Die größte Herausforderung des Forschungsdatenmanagements besteht in der langfristigen Konservierung „lebender Systeme“, wie das Projektportal von VA eines ist. VA betrachtet dieses von ihm entwickelte Projektportal als eine zeitgemäße Publikationsform, die in ihrer primären Funktion – der Veröffentlichung – mit der traditionellen Buchpublikation vergleichbar ist, aber natürlich darüber hinausgehende Möglichkeiten bietet. Der Wunsch wäre, dieses Webportal möglichst ad infinitum online verfügbar zu halten, vergleichbar mit der Bewahrung eines Buches in einer Bibliothek. Leider stehen diesem Ideal technische Schwierigkeiten entgegen, die struktureller Natur und daher bislang nicht lösbar sind. Das Problem besteht hauptsächlich in der ständigen Weiterentwicklung der Software- konkret: Serverumgebung, innerhalb derer ein solches Projektportal läuft. Die projekt- bzw. portalspezifische Software muss in größeren Abständen immer wieder an die veränderte Umgebung angepasst werden, bedarf also mehr oder minder permanenter Pflege. VA ist zwar insofern strategisch gut aufgestellt, als das Projektportal von der ITG betreut wird, die über eine unbefristete Bestandsperspektive verfügt und im Rahmen ihrer personellen Möglichkeiten die Betreuung des VA-Webportals auch über das Projektende von VA hinaus übernehmen wird, jedoch kann nicht ausgeschlossen werden, dass in mittel- bis langfristiger Perspektive derart großer Aufwand für den Fortbetrieb des Portals geleistet werden müsste, der die Kapazitäten der ITG übersteigt. Versuchsweise wurde eine ältere Version des VA-Webportals auf einem sog. Docker-Image auf einem Server der UB abgelegt (https://verba-alpina-archiv.ub.uni-muenchen.de/), jedoch erscheint auch dies nicht als absolut zuverlässige Dauerlösung. Als derzeit einzig vernünftiges Konzept zur dauerhaften Bewahrung auch des Webportals erscheint nur die von VA betriebene möglichst ausführliche Dokumentation der Funktionalität der Webseite zusammen mit der Archivierung des entwickelten Softwarecodes auf GitHub sowie im institutionellen Repositorium der UB (Letzteres wird derzeit noch projektintern diskutiert). Späteren Generationen sollte es dann zumindest theoretisch möglich sein, das Gesamtsystem einschließlich all seiner Funktionen mit der dann verfügbaren Technik „nachzubauen“.

2. Teil: Die Perspektive der Bibliothek

Wie aufgezeigt spielen Bibliotheken eine entscheidende Rolle im Umgang mit Forschungsdaten. Dabei nimmt die bereits an den Einrichtungen vorhandene Expertise in der Erschließung und Zugänglichmachung von Informationen eine zentrale Rolle im Prozess des Forschungsdatenmanagements ein.

Neben der Erschließung der Forschungsdaten und der Verknüpfung mit Normdaten kann die generische und fachspezifische Anreicherung mit Metadaten als ein Kompetenzbereich der Bibliotheken betrachtet werden. Oft besteht zusätzlich eine entsprechende Infrastruktur an den Einrichtungen, die sich mit der Vergabe von persistenten Identifikationen (PID) befasst. Bibliotheken können in der Regel auf eine langjährige Erfahrung mit der Vergabe und dem Einsatz von Digital Object Identifiern (DOI) und Uniform Resource Names (URN) zurückgreifen.

Diese Aufgabenbereiche bilden die Grundlage, um Daten zugänglich und auffindbar zu machen. Neben der Beratung zum Forschungsdatenmanagement finden sich Bibliotheken zudem immer häufiger in der Position des „Data Publishers“, also des Datenveröffentlichers. Mit der Veröffentlichung von Forschungsdaten – beispielsweise auf institutionellen Repositorien – entstehen zusätzliche Aufgabenfelder für die Bibliotheken. In der Regel verfügen die Publikationsplattformen bereits über eine Infrastruktur, die es ermöglicht, die Metadaten über Schnittstellen an weitere Suchmaschinen oder Discovery-Systeme zu liefern. Bibliotheken kennen dabei auch die Recherche- und Nutzungs-Bedürfnisse der Forschenden und sorgen dafür, dass auch die Forschungsdaten über geeignete Plattformen für eine breite Nutzergruppe zugänglich sind.

Während mit der Vergabe von PIDs bereits eine dauerhafte Zitierbarkeit gegeben ist, müssen sich die „Data Publishers“ auch mit der Frage auseinandersetzen, wie Forschungsdaten langfristig verfügbar bleiben können. An vielen Bibliotheken bestehen bereits entsprechende Workflows für digitale Publikationen, die sich teilweise auch auf Forschungsdaten übertragen lassen. Gemäß den Regeln der guten wissenschaftlichen Praxis sollen Daten mindestens zehn Jahre aufbewahrt werden.[17] Dies erscheint jedoch verglichen mit „traditionellen“ Beständen von Bibliotheken sehr kurz. Während beispielsweise im Bereich „Altes Buch“ Medien bewahrt werden, die mehrere hundert Jahre alt sein können, ist es unklar, ob heute erstellte Forschungsdaten in zehn Jahren noch lesbar sind. Es wird daher an verschiedenen Lösungen gearbeitet, von Technologien wie *Bitstream Preservation* bis hin zur Langzeitarchivierung, damit auch Informationen aus digitalen Daten langfristig verfügbar sind.

Fallbeispiel Universitätsbibliothek der Ludwig-Maximilians-Universität München

Die Themen Open Access und Forschungsdaten sind an der Universitätsbibliothek der Ludwig-Maximilians-Universität München (UB der LMU) Teil des Alltagsgeschäfts. Neben elektronischen Publikationsplattformen für Zeitschriften (Open Journals LMU), Hochschulschriften (Elektronische Hochschulschriften) und weiteren wissenschaftlichen

17 https://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf

Publikationen (Open Access LMU), werden auch hybride Publikationsformen (z. B. Open Publishing LMU) angeboten.[18] Die Veröffentlichung von Forschungsdaten ist seit 2010 über das institutionelle Repositorium Open Data LMU möglich.

Primär richtet sich das Repositorium an Wissenschaftler/innen aller Fakultäten der LMU sowie kooperierender Institutionen. Die Ausrichtung ist interdisziplinär, und es wurden bereits Forschungsdaten aus über 15 verschiedenen Fachgebieten veröffentlicht. Nutzer/innen können nach erfolgreicher Registrierung ihre Daten eigenständig auf den Server hochladen. Eine einheitliche Forschungsdaten-Policy wurde bisher nicht eingeführt, es wird aber empfohlen, Daten im Sinne der Budapester Open Access Initiative und der Berliner Erklärung über offenen Zugang zu wissenschaftlichem Wissen der Allgemeinheit zur Verfügung zu stellen.[19]

Das Repositoriums Open Data LMU läuft unter der Open-Source-Software EPrints[20]. An der UB der LMU ist die Version 3.3.15 im Einsatz. EPrints wird in einer Linux-Umgebung aufgesetzt und benötigt Perl sowie eine MySQL-Datenbank. Eine OAI-Schnittstelle erlaubt den Export von Metadaten in vielen Standard-Formaten, darunter DataCite, Dublin Core oder RDF.

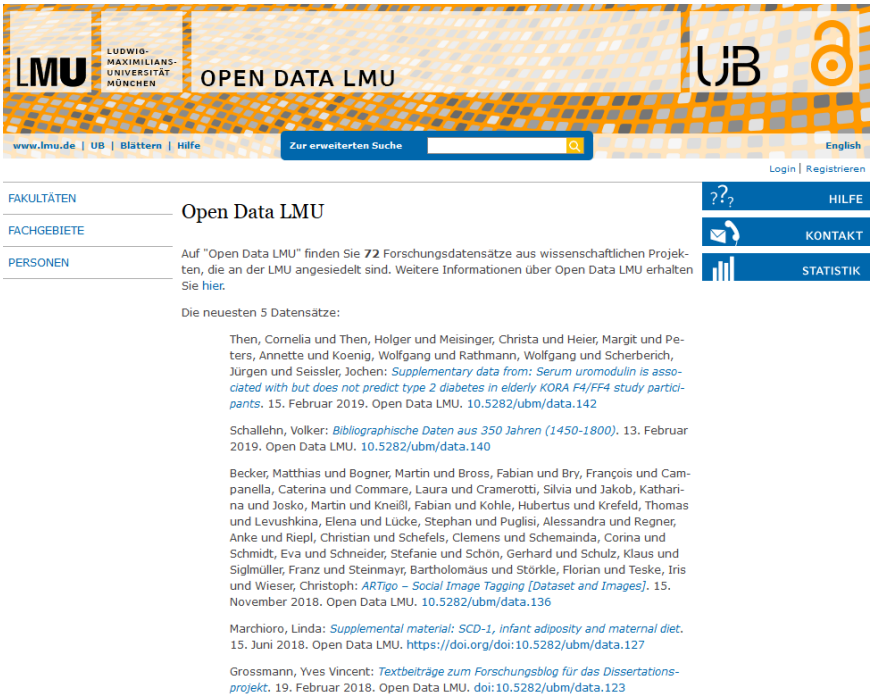

Abbildung 1: Das institutionelle Repositorium Open Data LMU

Durch die stetig wachsenden Anforderungen an das Forschungsdatenmanagement wurde an der UB der LMU alternative Repositorien-Software evaluiert. Eine Alternative sollte alle Funktionen, die EPrints bietet, beinhalten sowie einen komfortableren und vielseitigeren Umgang mit Forschungsdaten ermöglichen. Zentrale Anforderungen sind beispielsweise eine

[18] https://www.budapestopenaccessinitiative.org/
[19] https://openaccess.mpg.de/Berliner-Erklaerung
[20] https://www.eprints.org/

hohe Skalierbarkeit sowie die Anbindung an neue Technologien wie Linked Open Data und Semantic Web.

Die Wahl fiel schließlich auf Fedora Repositories[21]. Fedora ist ebenfalls ein Open Source-Produkt und wird von DuraSpace entwickelt. Hinter der Organisation steht eine große und aktive Community, die Lösungen für unterschiedliche Bedarfe anbietet. Das populärste Produkt von DuraSpace ist das Repositorium DSpace, welches mittlerweile zu den am häufigsten eingesetzten Datenrepositorien an deutschen Forschungseinrichtungen zählt. Während sich DSpace verhältnismäßig einfach installieren und einrichten lässt, fungiert Fedora als Middleware und bedarf tiefer greifender Entwicklungs- und Programmierarbeiten.

Wie genau sich Fedora als Middleware einsetzen lässt, zeigt folgende Skizze, die in der Entwicklungsphase an der UB der LMU erstellt wurde. Als Pilotprojekt diente dabei das in Teil 1 des Textes erwähnte Projekt VerbaAlpina (VA). Die Projektwebseite hat seit 2019 eine Schnittstelle, die Forschungsdaten in unterschiedlichen Formaten bereitstellt, wahlweise als csv, xml oder json[22].

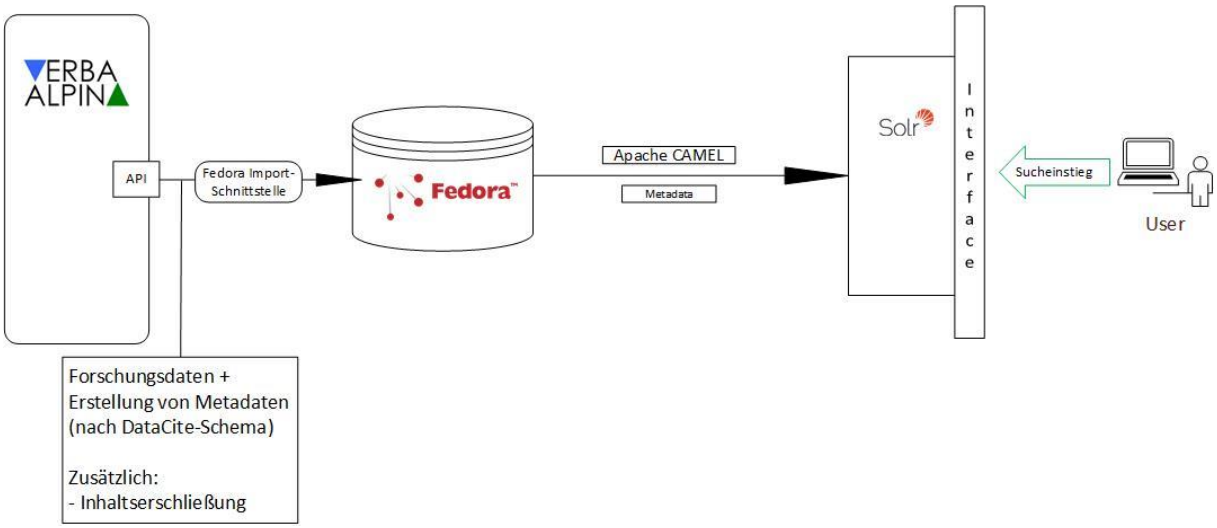

Abbildung 2: vereinfachte Skizze der Infrastruktur

Über die Schnittstelle gesammelte Daten werden anschließend um Metadaten angereichert. Dabei arbeiten Metadaten-Experten der UB der LMU mit Vertretern aus den Fachdisziplinen zusammen. Im Beispiel von VA entstand in Zusammenarbeit mit den Projektmitarbeiter/innen ein detailliertes Datenmodell, das anschließend in ein geeignetes Metadatenschema übertragen wurde.

Das Metadaten-Management ist die Kernaufgabe im frühen Stadium des Forschungsdatenmanagements und entscheidet darüber, wo und von wem die Daten aufgefunden werden können. Die UB der LMU verwendet dabei den Metadaten-Standard von DataCite.[23] Dieser Standard kann auch zur Registrierung von DOIs genutzt werden. Zusammen mit der ITG und dem Leibniz-Rechenzentrum (LRZ) der Bayerischen Akademie der Wissenschaften werden in einer Arbeitsgruppe Best-Practice-Empfehlungen für die

[21] https://duraspace.org/fedora/
[22] https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=8844
[23] https://schema.datacite.org/meta/kernel-4.2/

Verwendung des Metadaten-Schemas erarbeitet. Dies ist insofern sinnvoll, da den Forschenden eine Empfehlung gegeben werden kann, wie das Metadaten-Schema in ihrem Fachbereich am besten zu verwenden ist.

Das Thema Granularität spielt hierbei eine ebenso große Rolle. Je nach Disziplin kann die Granularität stark variieren. Da die Möglichkeit besteht, PIDs für Forschungsdaten zu vergeben, stellt sich unweigerlich die Frage, auf welcher Ebene dies geschehen soll. Im DOI-Handbook werden folgende Möglichkeiten genannt:

> A DOI name can be assigned to any object, regardless of the extent to which that object might be a component part of some larger entity. DOI names can be assigned at any desired degree of precision and granularity that a registrant deems to be appropriate.
>
> For example, for granularity in textual materials, separate DOI names can be assigned to a novel as an abstract work, a specific edition of that novel, a specific chapter within that edition of the novel, a single paragraph, a specific image, or a quotation, as well as to each specific manifestation in which any of those entities are published or otherwise made available.[24]

Im Falle von VA haben die Einzelbelege in der Forschung einen besonderen Status. Deshalb ist im Projektkontext die Überlegung, PIDs auf Einzeldatensatzebene zu vergeben. Bei VA handelt sich um ca. 81.000 Datensätze. Damit die Unterscheidung und Suchbarkeit der Einzelbelege sinnvoll gestaltet werden kann, fließen in die Metadaten auch Sacherschließung und geographische Informationen mit ein. Die in Teil 1 erwähnten Verbindungen der Einzelbelege unter den Typen „Gemeinden“, „morpholoexikalische Typen“ und „Konzept“ werden ebenfalls in den Metadaten berücksichtigt.

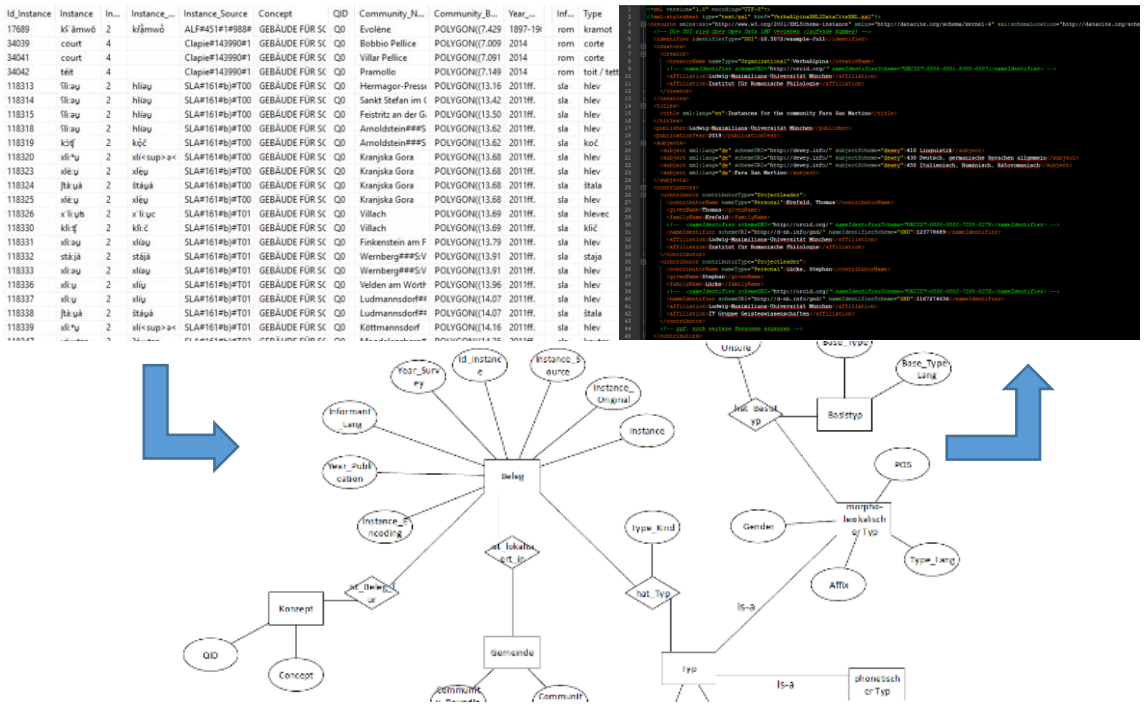

Abbildung 3: Transformation der Daten: Von den Roh-Daten aus einer CSV-Datei, über das Datenmodell hin zum DataCite-XML

[24] https://www.doi.org/doi_handbook/2_Numbering.html#2.3.2

Die Transformation erfolgt dabei über mehrere Schritte. Anhand eines detaillierten Datenmodells werden die Rohdaten der Schnittstelle bearbeitet und um entsprechende Metadaten angereichert. Dies erfolgt in Abstimmung mit den Forschenden. Anschließend wird eine DataCite-XML-Datei erstellt, die alle relevanten Metadaten beinhaltet. Diese XML-Datei wird von der UB der LMU auch für die Erstellung von DOIs verwendet.

Fedora bietet an dieser Stelle verschiedene Möglichkeiten für einen Ingest der Daten. Neben XML besteht auch die Möglichkeit, RDF-Dateien für das Anlegen der Datensätze zu verwenden. Im Testbetrieb wird an der Universitätsbibliothek der Import beider Varianten erprobt.

Der Fluss der Forschungs- und Metadaten ist daher sehr komplex. Ein daraus entwickeltes Aufgabenmodell wirft zudem einige Fragen auf, die im Rahmen des Projekts geklärt werden müssen:

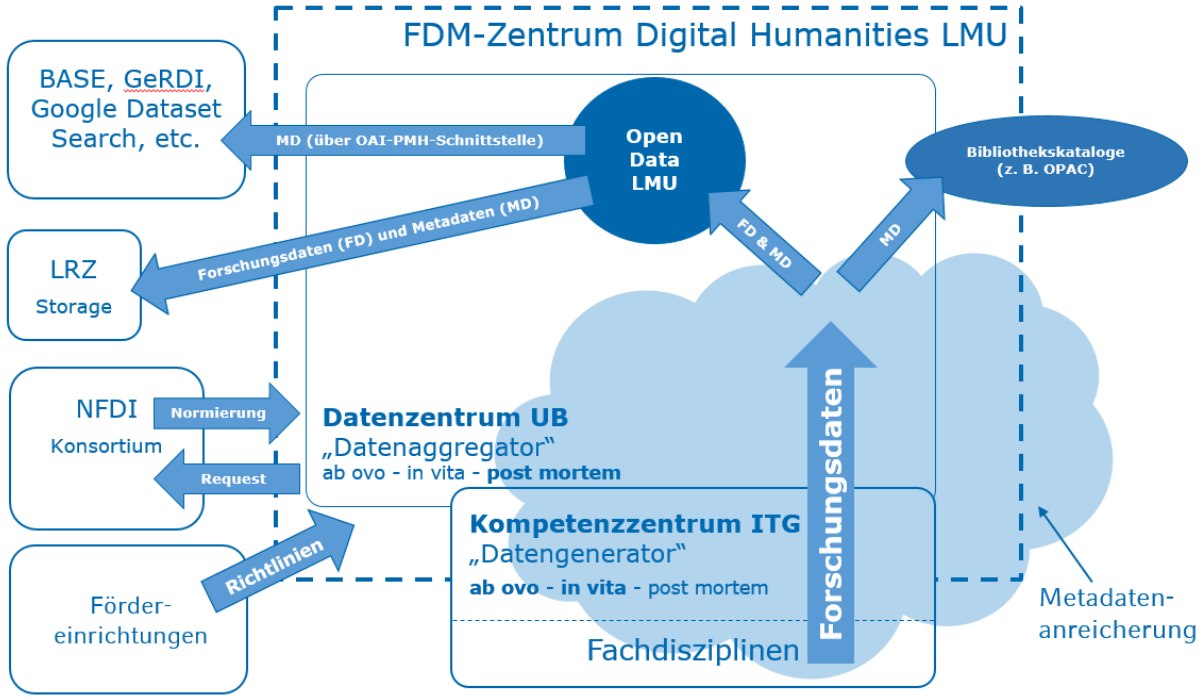

Abbildung 4: Erweitertes Aufgabenmodell zum FDM in den digitalen Geisteswissenschaften an der LMU

Das erweiterte Aufgabenmodell zeigt, wie die Aufgaben der ITG und der UB der LMU zusammenhängen. Die Anreicherung mit Metadaten erfolgt an der UB der LMU in enger Zusammenarbeit mit der ITG und den Forschenden aus den Fachdisziplinen. Sobald die Forschungsdaten um entsprechende Metadaten angereichert sind, werden Forschungs- und Metadatensätze in das Repositorium Open Data LMU übertragen. An diesem Punkt werden die Forschungsdaten zum ersten Mal über das institutionelle Repositorium durchsuch- und auffindbar. Um die Daten einer breiteren Nutzergruppe zur Verfügung zu stellen, werden über entsprechende Schnittstellen (z. B. OAI-PMH) Metadaten an weitere Dienste geliefert. Daten

von Open Data LMU werden momentan von BASE, Google Dataset Search oder GeRDI indexiert.

Zur Verfügbarkeit der Forschungsdaten bietet Open Data LMU eine sogenannte *Bitstream Preservation*. Die Daten werden gemäß den Regeln der guten wissenschaftlichen Praxis mindestens zehn Jahre aufbewahrt. Mit dem Umstieg auf eine neue Repositorien-Software soll gleichzeitig auch die Strategie der Datenarchivierung überarbeitet werden. Zukünftig soll ein Schwerpunkt nicht nur auf die Archivierung von Forschungs-, sondern auch der zugehörigen Metadaten gelegt werden. Eine den NESTOR-Richtlinien[25] folgende Langzeitarchivierung könnte dabei eingerichtet werden. Im Falle des Pilotprojekts VA bietet auch die Projektwebseite mit ihrer multimedialen Kartenansicht Informationen, die durch eine reine Ablage der Forschungsdaten verloren gehen könnte. Die UB der LMU arbeitet hier erneut im Tandem mit der ITG und dem VA-Projekt, um die Bedarfe der Wissenschaft zu erkennen und umzusetzen. In der Vergangenheit gab es des Öfteren Kooperationen mit dem LRZ, die sich mit der Archivierung von Daten beschäftigt haben. Auch für zukünftige Projekte soll in diesem Bereich kooperiert werden.

Für die langfristige Planung wird ebenfalls erwogen, Forschungsdaten über bestehende Recherche-Systeme an Bibliotheken auffindbar zu machen. Dies könnte über Bibliothekskataloge erfolgen. Dabei ist angedacht, analog zu den oben genannten Discovery-Diensten nur die Metadaten weiterzugeben. Sobald die Infrastruktur um das Fedora-Repositorium zuverlässig im Produktivbetrieb läuft, wird diese Möglichkeit evaluiert.

Die Zusammenarbeit zwischen UB und Wissenschaftler/innen der LMU hat gezeigt, dass es von großem Vorteil sein kann, auf bestehende Strukturen aufzubauen. Sowohl die Universitätsbibliothek, als auch die geisteswissenschaftlichen Institute sind ein fester Bestandteil der Universität und können auch längerfristige Vorhaben besser umsetzen, als beispielsweise Projekte mit begrenzter Laufzeit. Die Zusammenarbeit hat zudem auf beiden Seiten die Kompetenz der Forschungsdaten-Ansprechpartner vergrößert. Dabei spielt auch die Beratung auf verschiedenen Stufen des Forschungsprozesses eine zentrale Rolle, die sowohl von Infrastruktur-, als auch von Wissenschaftspartnern übernommen werden kann. Wenn die Universitätsbibliothek schon frühzeitig in die Planung und Durchführung von Forschungsdatenvorhaben involviert wird, erzeugt dies Synergien, die ein erfolgreiches Vorhaben leichter realisierbar machen. Mit der ITG und der UB der LMU besteht zudem ein institutioneller „Verbund“ für Forschungsdaten in den Digitalen Geisteswissenschaften an der LMU. Die Zusammenarbeit kann perspektivisch neue Kooperationen und Finanzierungsmöglichkeiten entstehen lassen.

Das Modellprojekt „eHumanities – interdisziplinär“

Die in Teil 1 und Teil 2 genannten Infrastruktur- und Wissenschaftspartner arbeiten nicht nur innerhalb der LMU zusammen, sondern sind auch Teil des Projekts: „eHumanities – interdisziplinär“. Dort arbeiten ITG und die UB der LMU unter der Federführung der UB der Friedrich-Alexander-Universität Erlangen-Nürnberg (UB der FAU) an Fragestellungen zu Forschungsdaten in den digitalen Geistes- und Sozialwissenschaften. Das Projekt wird vom

[25] https://www.langzeitarchivierung.de/

Bayerischen Staatsministerium für Wissenschaft und Kunst gefördert und hat eine Laufzeit von drei Jahren (März 2018 – März 2021).

Im Projekt beschäftigen sich die Mitarbeiter/innen mit der Konzeption und Evaluierung neuer Hilfsmittel und der Erarbeitung von Best-Practice-Empfehlungen. Die Verbindung von digitaler Bibliotheksexpertise mit informatischen und fachmethodischen Schnittstellenkompetenzen steht dabei im Vordergrund.

Projektergebnisse werden über die Projektwebseite[26] veröffentlicht. Der Zwischenbericht über das erste Projektjahr[27] wurde bereits veröffentlicht und zukünftige Berichte und Ergebnisse werden ebenfalls mit der Community geteilt. Ziele sind dabei auch, Erfahrungsberichte zu Fedora zu veröffentlichen und Programmierarbeiten über die Plattform GitHub zur Verfügung zu stellen. Durch die Veröffentlichung von Quellcode in Kombination mit den Berichten soll eine Transferierbarkeit der Projektergebnisse auf weitere Vorhaben und/oder Disziplinen möglich sein.

Bereits nach einem Drittel der Projektlaufzeit hat sich gezeigt, dass die Zusammenarbeit von Infrastruktur- und Wissenschaftspartnern neue Sichtweisen auf altbekannte Probleme ermöglicht, wie z. B. die Fragen nach der Langzeitarchivierung, Granularität und Auffindbarkeit. Durch eine Kooperation wird zudem vermieden, dass Bibliotheken und Wissenschaftler/innen Insellösungen für ihre Projekte aufbauen. Die Zusammenarbeit sorgt dafür, dass sich an der LMU aus Erfahrung mit Pilotprojekten der digitalen Geisteswissenschaften feste Workflows etablieren, von denen alle Teilnehmenden profitieren und durch die wiederkehrende Themen schneller bearbeitet werden können.

Die Bereitstellung von Datenmodellen und das Erarbeiten von Best-Practice-Empfehlungen für Metadatenschemata können zudem in Datenmanagementpläne übernommen werden. Dies ist ein weiteres Arbeitspaket des Modellprojekts, mit dem sich die UB der FAU beschäftigt. Dort wird im Frühjahr 2019 eine RDMO-Instanz eingeführt, die anschließend auch auf Server der UB der LMU übertragen werden soll. Um die Wissenschaftler/innen für das Thema „Forschungsdaten“ zu sensibilisieren, wird an der UB der FAU in einem weiteren Arbeitspaket ein digitales Lern- und Informationsangebot erstellt. Die Schulungs- und Videomaterialien werden anschließend als *Open Educational Resources* auch anderen Interessenten zur Verfügung gestellt und in den Digitalen Campus Bayern integriert. Dort können sie beispielsweise zum Bestandteil von Curricula im Bereich Digital Humanities werden. Mit leichten Veränderungen sind diese Materialien auch auf andere Institutionen und verwandte Disziplinen transferierbar.

[26] https://www.fdm-bayern.org
[27] https://zenodo.org/record/2645935