Datenmanagementplan des Sonderforschungsbereichs Materiale Textkulturen (933) der Deutschen Forschungsgemeinschaft (DFG)

Ausarbeitung des Teilprojektes INF - Service-Projekt zu Informationsmanagement und Informationsinfrastruktur des Sonderforschungsbereichs

Stand: Heidelberg, 20. Dezember 2018

Verantwortlich für die Datenhaltung des Sonderforschungsbereichs Materiale Textkulturen ist sein Sprecher

Prof. Dr. Ludger Lieb
Germanistisches Seminar
Universität Heidelberg
Hauptstraße 207-209, D-69117 Heidelberg
Dienstraum: PB 128

Telefon: +49 (0)6221-54 3434
Telefax: +49 (0)6221-54 3378
E-Mail: ludger.lieb <_at_> gs.uni-heidelberg.de

Verantwortlich für den Betrieb der Serversysteme (Virtuelle Maschine) ist der Leiter des Universitätsrechenzentrums der Universität Heidelberg (URZ)

Prof. Dr. Vincent Heuveline
Universitätsrechenzentrum
Im Neuenheimer Feld 293, 69120 Heidelberg
Raum: 108

Telefon: +49 6221 54 20000
vincent.heuveline <_at_> uni-heidelberg.de

1 Einleitung

Der SFB Materiale Textkulturen, Materialität und Präsenz des Geschriebenen in non-typographischen Gesellschaften (933) – im Weiteren abgekürzt durch SFB 933 – untersucht Dinge, auf denen etwas geschrieben steht wie zum Beispiel Säulen, Stelen, Portale, Grabsteine, Tontafeln, Tonscherben, Amulette, Rollen, Papyri, Pergamentcodices und die aus Gesellschaften stammen, in denen es keine massenhafte Produktion von Geschriebenem gab. Das Interesse richtet sich dabei auf die spezifische materiale Beschaffenheit und die dadurch bewirkte Präsenz der beschrifteten Artefakte und des Geschriebenen selbst. Die beteiligten Forscherinnen und Forscher stellen an diese Artefakte eine Vielzahl von Fragen wie beispielsweise: Wie und unter welchen Bedingungen wurden sie hergestellt? In welchen räumlichen Arrangements befanden sie sich? Wer hatte Zugang zu ihnen? Was wurde mit ihnen gemacht bzw. welchen Handlungen wurden an ihnen vollzogen? Welche Handlungen lösten sie aus? Der Leitgedanke ist, dass Schrift, Schriftträger und darauf bezogene Praktiken eine unlösbare wechselseitige Verbindung eingehen, deren Berücksichtigung für das Verständnis der überlieferten Texte und ihrer kulturellen Umgebung hohe Erklärungskraft besitzt.

Am Sonderforschungsbereich sind unterschiedliche geisteswissenschaftliche Disziplinen beteiligt, darunter: Assyriologie, (B01), Ägyptologie (A03UP3), Papyrologie (A03UP2), Klassische Archäologie (A10), Alte Geschichte und Epigraphik (A01, A09), Bibel und Jüdische Bibelauslegung (B04), Byzantinische Archäologie und Kunstgeschichte (A01UP3), Sinologie (B09), Geschichte des Mittelalters (A01UP3, A06), Mittelalterliche Kunstgeschichte (A05), Kunstgeschichte Ostasiens, Klassische Philologie (Latinistik C03, Gräzistik C08), Lateinische Philologie des Mittelalters und der Neuzeit (A08), Germanistik (C05), Romanistik (C09), sowie die Ethnologie (C07).

Dass Fragestellungen der Geisteswissenschaften mit Hilfe von Informationstechnologie und mit Hilfe eigener und öffentlich verfügbarer Daten bearbeitet werden, ist heute gängige Praxis. Dabei ergeben sich aus der zunehmenden Verfügbarkeit digitaler Daten viele neue Forschungsansätze und -möglichkeiten. An der Schnittstelle zwischen Geisteswissenschaften und Informatik unterstützt das Service-Projekt zu Informationsmanagement und Informationsinfrastruktur (im Folgenden TP INF) die Forschenden des SFB bei der Datenerfassung und Aufbereitung, der Erstellung digitaler Editionen und Visualisierungen und sorgt für ihre langfristige Speicherung und Verfügbarkeit. Von vornherein spielen dabei Überlegungen zu einer Datenvernetzung mit anderen Instituten und Forschungseinrichtung sowie die spätere dauerhafte öffentliche Verfügbarkeit des angefallenen Datenbestands und eine potentielle wissenschaftliche und öffentliche Nachnutzung eine wichtige Rolle. Vor diesem Hintergrund sieht sich das TP INF als eine kuratierende Instanz, die die forschenden Teilprojekte bei ihrer datenbezogenen Arbeit unterstützt.

In diesem Dokument wird das Management der erhobenen, untersuchten und verarbeiteten Forschungsdaten des SFB 933 dokumentiert. Es dient als Empfehlung und Leitfaden für die einzelnen Teilprojekte zur strukturierten Datenhaltung und unterstützt die nachhaltige Aufbewahrung wertvoller Datensätze und ihrer langfristigen Nachnutzbarkeit.

Allgemeine Empfehlungen für die Konzipierung des Forschungsdatenmanagements wurden insbesondere in dem BMBF-geförderten Projekt WissGrid[1] erarbeitet, welches sich der Etablierung wissenschaftlicher Grids widmete. Für die Geisteswissenschaften ist die virtuelle Forschungsumgebung TextGrid[2] entstanden, welches wiederum im Rahmen der Digital Research Infrastructur for the Arts and Humanities (DARIAH-DE)[3] fortgeführt wird. Während WissGrid eine fach-übergreifende Checkliste für einen Datenmanagementplan zur

1 https://escience.aip.de/wissgrid/
2 https://textgrid.de/
3 https://de.dariah.eu/

Verfügung stellt, informiert DARIAH-DE bzgl. der Geisteswissenschaften im Speziellen zu den Fragen des Forschungsdatenmanagements[4]. Das Forschungsdatenzentrum für Archäologie und Altertumswissenschaften IANUS[5] hält ebenfalls Empfehlungen und eine Checkliste[6] für ein fach-spezifisch konzipiertes Forschungsdatenmanagement bereit. Weitere Informationen im Form eines Handbuches wurden durch nestor[7], dem deutschen Kompetenznetzwerk zur digitalen Langzeitarchivierung, Zusammenarbeit von Bibliotheken, Archiven, Museen und Experten der Langzeitarchivierung erarbeitet. Auf internationaler Ebene ist das Digital Curation Centre, DCC[8] zu nennen, welches ebenfalls eine Checkliste[9] und ein online Planungstool[10] zur Verfügung stellt; auf nationaler Ebene das Portal forschungsdaten.org[11], auf dem viele Informationen gebündelt worden sind. Die genannten Ressourcen bilden die Grundlage für das vorliegende Dokument.

2 Struktureller Rahmen

2.1 Projektleiter und Mitarbeiter des INF-Projekts

Der SFB 933 wird seit Juli 2011 gefördert. Etwa 70 Wissenschaftlerinnen und Wissenschaftler aus achtzehn geisteswissenschaftlichen Disziplinen der Universität Heidelberg und der Hochschule für jüdische Studien Heidelberg sind an ihm beteiligt. Die DFG bewilligte im Mai 2015 eine zweite Förderperiode des SFB 933 bis Juni 2019.

Das Serviceprojekt INF steht seit Juli 2011 unter der Leitung von Prof. Dr. Christian Witschel und seit Juli 2015 zudem unter der Leitung von Prof. Dr. Vincent Heuveline (Universitätsrechenzentrum, URZ). Im INF-Projekt arbeiteten und arbeiten verschiedene Mitarbeiter mit einem maximalen Stellenumfang von insgesamt 137% einer vollen Stelle.

Mitarbeiter

Name	**Zeitraum**	**Institution**
Jeromin Fest	Februar 2017 bis Juni 2019	Universitätsrechenzentrum
Christoph Forster	Juli 2011 bis Juni 2019	extern, datalino PartG
Fabian Gebhart	Juli 2018 bis Juni 2019	Universitätsrechenzentrum
Frank Grieshaber	Juli 2011 bis Februar 2018	Seminar für Alte Geschichte und Epigraphik
Frank Krabbes	März 2018 bis Juni 2019	Universitätsbibliothek
Jonas Kratzke	Juli 2015 bis Mai 2018	Universitätsrechenzentrum
Martin Wlotzka	August 2017 bis Juni 2018	Universitätsrechenzentrum

4 https://wiki.de.dariah.eu/pages/viewpage.action?pageId=20058160
5 http://www.ianus-fdz.de/
6 http://www.ianus-fdz.de/it-empfehlungen/datenmanagement
7 http://www.langzeitarchivierung.de
8 http://www.dcc.ac.uk/
9 http://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf
10 https://dmponline.dcc.ac.uk/
11 http://www.forschungsdaten.org/index.php/Hauptseite

2.2 Zuständigkeiten

Die Mitarbeiter des INF-Projektes sind für die Bereitstellung der SFB-zentralen IT-Infrastruktur für eine gemeinsame Datenhaltung und -verarbeitung zuständig sowie für die Online-Präsentation ausgewählter Forschungsergebnisse und die Organisation der langfristigen Archivierung und Zugänglichkeit der Forschungsdaten. Die Datenerhebungen werden von den Teilprojekten größtenteils eigenständig übernommen. Zum Teil werden sie dabei durch Mitarbeiter, studentische/wissenschaftliche Hilfskräfte oder die Hilfskraft des TP-INF unterstützt.

Institutionelle Ansprechpartner für die Einbettung des Datenmanagements in das Forschungsdatenmanagement-Konzept der Universität Heidelberg sind durch Mitarbeiterinnen und Mitarbeiter des Kompetenzzentrums Forschungsdaten (KFD) gegeben. Das KFD ist eine gemeinsame Serviceeinrichtung der Universitätsbibliothek (UB) und des Universitätsrechenzentrums (URZ).

Ansprechpartner für die IT-Grundausstattung der einzelnen Arbeitsplätze sind die jeweiligen EDV-Beauftragten der beteiligten Forschungseinrichtungen.

2.3 Richtlinien, Vorgaben, Standards

Seitens der Universität Heidelberg gelten die Richtlinien für das Management von Forschungsdaten durch die „Research Data Policy“[12] sowie die Sicherung guter wissenschaftlicher Praxis[13]. Seitens der DFG gelten die „Leitlinien zum Umgang mit Forschungsdaten“[14].

Im Bezug auf Empfehlungen und Standards für das Datenmanagement in den Geisteswissenschaften sei die Dokumentation durch DARIAH-DE erwähnt.[15] Insbesondere wird die Verwendung etablierter kontrollierter Vokabulare beispielsweise der Gemeinsamen Normdatei (GND) der Deutschen Nationalbibliothek oder des Art & Architecture Thesaurus des Getty Research Institutes empfohlen.

3 Datenerhebung / Art der Daten

Die Forschungsdatengrundlage des SFB 933 wird sowohl durch die Sammlung von extern verfügbaren Digitalisaten als auch durch die eigene Erhebung von Daten gebildet. Grundsätzlich sind dabei die erhobenen Daten reproduzierbar, wenngleich unter teilweise sehr hohem Aufwand. Für die textbasierten Daten des SFB 933 sind umfangreiche händische Sichtungen und Kategorisierungen von Literatur- und Objektdatenbeständen erforderlich. Fotoserien von zum Beispiel archäologischen Stätten (z.B. Korykos, Türkei) oder Bauwerken (z.B. Hagia Sophia, Türkei, oder St. Michael in Hildesheim), zumal im Ausland, sind zugangs- und witterungsbedingt nicht identisch reproduzierbar.

Die Forschungsdaten, die im SFB 933 erhoben werden, sind in den allermeisten Fällen Kleindatenbestände die – ähnlich einem persönlichen Zettelkasten – für die Beantwortung von meist sehr spezifischer Forschungsfragen angelegt werden. Das INF-Projekt begleitet diese Datenerhebungen durch einen Kuratierungs-

12 http://www.uni-heidelberg.de/universitaet/profil/researchdata/, abgerufen am 28.04.2016
13 http://www.uni-heidelberg.de/universitaet/profil/wissenschaftliche_praxis/, abgerufen am 28.04.2016
14 http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten.pdf, abgerufen am 12.01.2018
15 https://wiki.de.dariah.eu/pages/viewpage.action?pageId=20058160, abgerufen am 28.04.2016

prozess, um die Daten für die Forschungsfrage bzw. für Folgenutzungen aufzubereiten.

3.1 Digitalisierung

Im SFB 933 werden in geringfügigem Umfang Digitalisierungsaufgaben übernommen (z.B. Fotoserien der TPe A01UP2, A05). In den allermeisten Fällen wurden und werden bereits vorhandene Digitalisate genutzt.

3.2 Nutzung bereits bestehender Datenbestände

Über die Teilprojekte hinweg werden Daten aus verschiedenen Museen, Archiven, aus eigenen teilweise privaten Sammlungen, und bestehenden Datenbanken genutzt. Außerdem kann auf 3D-Modelle (TP A05 Hildesheimer Dom) zurückgegriffen werden. Die Zugriffsmöglichkeiten hängen von den jeweiligen Institutionen ab, die die Daten zur Verfügung stellen. Die Urheberrechte dieser bereits bestehenden Datenbestände sind entsprechend unterschiedlich geregelt. Darüber hinaus können in einigen Fällen bereits publizierte Datenbestände verwendet werden, für die eine freie Nutzungslizenz gilt (z.B. solche des British Museums unter CC BY-NC-SA 4.0). Solche Daten bleiben zunächst in ihrer vormaligen Struktur, werden aber so aufbereitet, dass sie seitens der Teilprojekte zusammen mit weiteren Daten in teilprojektspezifischen Datenbankansichten verarbeitet werden können.

Das Format und die Qualität der zusammengestellten Forschungsdaten sind sehr heterogen. Unter den Bilddatenformaten finden sich die meisten gängigen Typen. Datenbanken sind vor dem Ingest meist als Dateien in den Formaten FileMaker, Microsoft Excel oder Microsoft Word angelegt.

3.3 Rechtliche Aspekte

Für die vom SFB 933 nicht selbst erhobenen, jedoch verwendeten Daten gelten unterschiedliche Lizenzbedingungen. Diese reichen von sehr restriktiv gehandhabten Nutzungsrechten für zum Beispiel Bilddateien bis zu sehr freien Creative Commons Lizenzen (CC).

Für die in den Teilprojekten erhobenen Daten wird eine möglichst freie Anschlussnutzung angestrebt. Der SFB 933 ist bemüht, alle von ihm erfassten und generierten Daten mindestens unter der Creative Commons Lizenz CC-BY-NC-SA 4.0 (Namensnennung - Nicht kommerziell - Share Alike 4.0 International[16]) zu veröffentlichen. Daten, die für eine Qualifikationsarbeit erhoben worden sind, werden bis zum Abschluss des jeweiligen Verfahrens ausschließlich projektintern genutzt.

3.4 Dokumentation

Die Art der Datenerhebung wird bei der Datenübergabe an das TP-INF-Teilprojekt in einem Datenerhebungsbogen erfragt.

16 https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.de

4 Datenhaltung und Dienste

Der Kern des Forschungsdatenmanagements im SFB 933 ist MTK-Online. MTK-Online bezeichnet die zentrale IT-Infrastruktur für die gemeinsame Datenhaltung des SFB, bestehend aus einer virtualisierten Server-Umgebung, auf der die Dienste für den SFB gehostet werden (MTK-Server) und einem externen Speicherknoten (MTK-Speicher). Das Konzept für die zentrale Datenhaltung des SFB wird im folgenden Abschnitt erläutert. Die technische Umsetzung wird in den darauf folgenden Abschnitten beschrieben.

4.1 Konzeption der gemeinsamen Datenhaltung des SFB (MTK-Online)

Die Grundlage der gemeinsamen Datenhaltung der Teilprojekte bildet die vom TP INF verwaltete zentrale MySQL-Datenbank (Backend) sowie die auf dieser und dem gemeinsamen Speicher aufsetzenden webbasierten Oberflächen für die Datenmanipulation und -erfassung (Frontends). Alle Bestandteile zusammen bilden MTK-Online und werden nach den jeweiligen spezifischen Erfordernissen der Teilprojekte ständig erweitert. Dadurch wird es möglich, dass Datensammlungen nicht dezentral und unsichtbar in den einzelnen Teilprojekten entstehen, sondern möglichst gleich bei ihrer Erfassung in den gemeinsamen Datenbestand des SFB integriert werden. Dabei wird stets darauf geachtet, bereits vorhandene technische Lösungen (z.B. Tools wie den Shape-Editor für Bildannotationen) und – wo möglich – Standardisierungen einzusetzen, um die – auch spätere – Vernetzbarkeit und Interoperabilität der Datensammlungen zu ermöglichen (u. a. durch Nutzung kontrollierter Vokabulare wie beispielsweise die der GND, Geonames, PLEIADES und des Getty AAT oder Anwendung der TEI / MEI). Während in der ersten Förderperiode nur einige Teilprojekte die gemeinsame Plattform genutzt haben, konnten in der zweiten Förderperiode nun nahezu alle Datenbestände des SFB 933 in MTK-Online integriert und damit auch deren nachhaltige Sicherung und Verfügbarkeit vorbereitet werden. Die Mitarbeiter*innen des SFB erheben ihre Daten somit kaum noch in eigenen Listen und Datenbanken, sondern nutzen das zentrale System. Dieser gemeinsame Datenbestand umfasst zurzeit u. a. beispielsweise rund 30.000 Bilder, Daten zu 8.000 Texten, 2.600 Objekten, 500 Orten und 300 Personen.

Durch die Zusammenschau der erhobenen Daten zeigt sich aber auch deren (im Einrichtungsantrag des SFB 933 noch unterschätzte) Heterogenität und Spezifizität. Um die davon abhängige jeweilige Aussagekraft der Daten zu bewahren und weil es für die einzelnen Forschungsvorhaben darüber hinaus wichtig ist, nicht nur als Teil des SFB, sondern auch in ihren jeweiligen Fachgemeinschaften wahrgenommen zu werden, haben wir das ursprünglich zentrale Datenpräsentationskonzept zugunsten eines zunächst dezentralen substituiert. Diesem Konzept folgend, veröffentlichen wir nun nach und nach die Datenbestände der Teilprojekte als eigenständige, in der Regel einzelnen Forschungsarbeiten zuzuordnende Einheiten und konnten damit auch die Akzeptanz des TP INF im SFB steigern.[17] Nicht alle dieser Präsentationen sind dabei endgültig abgeschlossen; diejenigen der Teilprojekte A01UP3, A09 und C05/C09 sollen in der 3. Förderperiode noch weiter ergänzt werden. Auch im Sinne einer effizienten Ressourcennutzung werden die diesen Publikationen zugrunde liegenden Daten in dafür geeigneten Formaten (z.B. CSV, SQL, TIFF) und unter möglichst freien Lizenzen (Creative Commons) in langfristig von der Universität Heidelberg betriebene Repositorien überführt (heidICON, heiDATA, zukünftig heiMAP), in denen sie unter dauerhaften Identifikatoren (DOIs) einerseits in die SFB-eigenen Datenpublikationen eingebunden, andererseits aber unabhängig von diesen auch für weitere Forschungen nachhaltig zugänglich und nutzbar vorgehalten werden. Dafür ist es zunächst notwendig, die jeweiligen Datenbestände mit Metadaten zu versehen, so dass diese auch für die Auswertung von Maschinen zur Verfügung stehen (Dublin Core, DDI XML).

17 https://www.materiale-textkulturen.de/daten.php

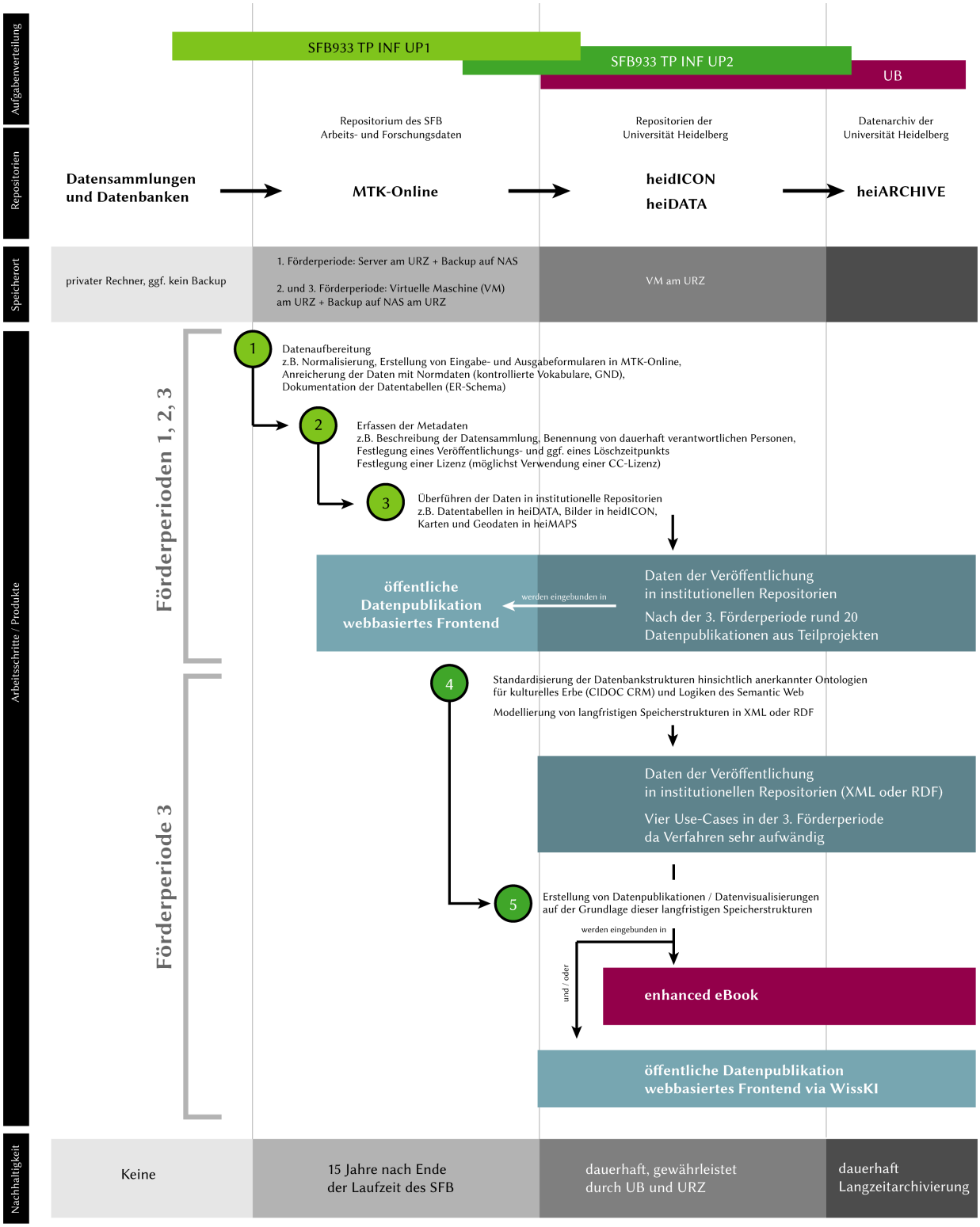

Abbildung 1: Aufgaben des TP INF im Bereich des Datenmanagements in Zusammenarbeit mit dem Universitätsrechenzentrum und der Universitätsbibliothek Heidelberg

Darüber hinaus beschreibt das TP INF den Inhalt der jeweiligen Tabellenspalten durch deren Referenzierung im CIDOC CRM, das als ISO-Norm 21127 die erweiterbare Ontologie für Begriffe und Informationen im Bereich des Kulturerbes bildet und einen allgemeinen Rahmen für die formale Semantik der Informationen

in diesem Bereich setzt. Dadurch werden die Voraussetzungen für die Informationsintegration geschaffen (z.B. Konvertierung der Daten in RDF für die Bereitstellung von Linked Open Data). Durch die Überführung der Forschungsdatenpublikationen in heiDATA, werden diese schließlich auch national und international sichtbar, da der in Heidelberg ansässige Fachinformationsdienst Altertumswissenschaften (Propylaeum) über die in heiDATA angelegten Bestände informiert.

Das TP INF trägt so Sorge für die elektronische Erfassung von Daten, deren Qualitätssicherung, Präsentation und Publikation sowie schließlich ihre langfristige Verfügbarkeit und nimmt damit eine Mittlerrolle zwischen den Forscher*innen und den bereits zur Verfügung stehenden oder noch auszubauenden Archivierungsdiensten und Repositorien der Universität Heidelberg ein. Das TP versteht sich in diesem Sinne als kuratierende Instanz für die Arbeits- und Forschungsdaten des SFB und arbeitet diesbezüglich eng mit dem gemeinsam von URZ und UB betriebenen Kompetenzzentrum Forschungsdaten der Universität Heidelberg (KFD) zusammen, um die relevanten Datenbestände gemäß der FAIR-Prinzipien zu veröffentlichen, nach denen diese auffindbar, zugänglich, interoperabel und wiederverwendbar sein sollen.

An der gemeinsamen Datenhaltung haben sich – nach der Empfehlung des TP INF, die auch vom Vorstand des SFB geteilt wurde – nur diejenigen TPe beteiligt, die das ausdrücklich gewünscht haben. Eine Verpflichtung zur Integration der teilprojektspezifischen Daten in MTK-Online bestand und besteht nicht. Aufgrund der personellen Ausgestaltung des Teilprojekts INF wäre ein solcher Anspruch in der ersten Förderperiode auch nicht umsetzbar gewesen. Ab der zweiten Förderperiode des SFB konnten – bis auf wenige Ausnahmen – alle Forschungsdaten der Teilprojekte in MTK-Online erfasst bzw. verwaltet werden.

4.2 Der MTK-Server

Der MTK-Server wird als virtuelle Maschine in der universitätseigenen Cloudinfrastruktur heiCLOUD betrieben. Die heiCLOUD befindet sich mitsamt der zugehörigen Hardwareumgebung in den Maschinenräumen des URZ und wird von diesem verwaltet. Dadurch ist sichergestellt, dass alle Daten und Dienste des SFB 933, die sich in der heiCLOUD befinden, unter der Kontrolle des TP INF an der Universität Heidelberg verbleiben. Durch den Einsatz der virtualisierten Infrastruktur kann die Provisionierung (also die technische Ausgestaltung der eingesetzten IT-Ressourcen) jederzeit bedarfsgerecht angepasst werden.

Eingesetzte Software

Für MTK-Online wird, wo immer möglich, freie und Open Source Software (FOSS) verwendet. Die wichtigste eingesetzte Software wird nachfolgend vorgestellt.

- Als Betriebssystem verwendet MTK-Online die Linux Distrubution Ubuntu in der Server-Variante, die in der jeweils aktuellen Version betrieben wird.
- Die Administration von MTK-Online durch das TP INF erfolgt über SSH-Zugriff auf die virtuelle Maschine. Hier kommt ein OpenSSH Server unter Verwendung von OpenSSL zum Einsatz. Die Nutzerauthentifizierung ist ausschließlich per public-key-Verfahren möglich.
- Als Webserver wird Apache verwendet. Die Webseiten und Arbeitsoberflächen von MTK-Online sind in PHP v5 implementiert. Eine Migration auf PHP v7.2 ist für das Frühjahr 2019 geplant.
- Der größte Teil der Datenbestände des SFB933 wird in MySQL-Datenbanken verwaltet. Dafür wur-

de ein MySQL Server eingerichtet, der das Backend für MTK-Online und die Datenpublikationen der Teilprojekte bildet. Der Zugriff auf die Datenbanken erfolgt zumeist durch die Arbeitsoberflächen in MTK-Online, die die Nutzeranfragen im Hintergrund an den MySQL-Server stellen. Zusätzlich wurde für das TP B06 ein eigener, passwortgesicherter MySQL-Account erstellt, über den ausschließlich lesend auf die entsprechenden Tabellen zugegriffen werden kann.

- Für Teilprojekte, die ein Versionierungssystem einsetzen, wurde ein zentraler SVN Server in der Virtuellen Maschine eingerichtet. Die SVN-Daten liegen dabei auf dem MTK-Speicher, und die SVN-Accounts sind mit Passwörtern gesichert. Durch entsprechende Rechtevergabe ist sichergestellt, dass keine unberechtigten Zugriffe auf SVN-Projekte erfolgen können.
- Manche Teilprojekte benötigen neben den Arbeitsoberflächen in MTK-Online auch Speicherplatz für Daten, auf die sie direkt zugreifen können. Dafür wurde ein eigener Bereich auf dem MTK-Speicher eingerichtet, der über MTK-Online per SFTP erreicht werden kann. Dafür wird der interne SFTP-Server des SSH-Servers (s.o.) verwendet. Die SFTP-Accounts sind mit Passwörtern gesichert und so konfiguriert, dass sie ausschließlich SFTP-Zugriff auf die eigens dafür angelegten Verzeichnisse erlauben (sog. „chrooted sftp“). Für die einzelnen Teilprojekte wurden getrennte Verzeichnisse angelegt. Dadurch wird verhindert, dass Nutzer unerlaubte Befehle ausführen oder auf die Daten anderer Teilprojekte zugreifen.
- MTK-Online wird mit Nagios Core überwacht. Der Nagios-Server läuft außerhalb der heiCLOUD-Umgebung des SFB am Universitätsrechenzentrum und überwacht MTK-Online per NRPE Plugin. Es werden CPU-Last, Belegung der internen Festplatte der VM, Belegung des MTK-Speichers, und die Anzahl der Prozesse überwacht.

Update-Policy

Die zum eingesetzten Linux Distribution Ubuntu-Server empfohlenen Updates (insbesondere sicherheitsrelevante Updates) werden durch die Administratoren des MTK-Servers jeweils zeitnah eingespielt. Durch den Einsatz einer virtualisierten Server-Umgebung besteht die Möglichkeit, die Funktionsfähigkeit der Dienste nach einem Update in einer Prüf-Instanz zunächst zu testen, bevor ein Update im produktiven Kontext eingesetzt wird. Sofern trotz initialer Tests im Produktivbetrieb dennoch Probleme infolge von Updates auftreten sollten, besteht die Möglichkeit, die zuletzt verwendete Konfiguration (vor Einspielen des Updates) wieder zu reaktivieren und temporär weiter zu verwenden. Mit diesem Prozess ist sichergestellt, dass das System zeitnah auf dem jeweils aktuellen Stand der Technik ist und die Verfügbarkeit des Dienstes möglichst hoch ist.

4.3 MTK-Speicher

Als zentrale Speicher-Ressource hat der SFB 933 eine Synology Rackstation RS2416RP beschafft. Diese wird in den Maschinenräumen des URZ betrieben. Da Datenverluste oder Ausfälle des zentralen MTK-Speichers gravierende Auswirkungen auf die Arbeit der Teilprojekte hätten, ist er besonders gesichert. In der Synology Rackstation sind zwölf Festplatten mit einer Speicherkapazität von jeweils 5,5 Tbyte mit RAID6-Sicherung verbaut, so dass beim Ausfall einzelner Festplatten kein Datenverlust entsteht. Das Rechenzentrum hat aus eigenen Mitteln eine zweite, identische Rackstation beschafft, die in einem anderen Brandabschnitt des URZbetrieben wird. Gemeinsam mit dem ersten System ist diese zu einem Hochverfügbar-

keitsspeicher verbunden, sodass alle Daten stets auf beide Speicherknoten gespiegelt werden. Dadurch werden selbst beim Ausfall mehrerer Festplatten oder einer ganzen Rackstation Datenverluste verhindert und die Zeiten der Nichtverfügbarkeit aufgrund von Wartungen oder Ersatz defekter Festplatten auf ein Minimum reduziert. Jede der Rackstations ist redundant an Strom und Netzwerk angebunden.

In der Zukunft ist geplant, die beiden Systeme durch den zentralen Speicherdienstes SDS@hd – Scientific Data Storage[18] zu ersetzen. Dieser Dienst ist optimiert für die Speicherung wissenschaftlicher Daten, auf die häufig Zugriffe erfolgen (hot data). Er zeichnet sich durch eine sehr hohe Datenverfügbarkeit und -sicherheit aus, insbesondere ist eine Kopie aller Daten auf einem getrennten System ebenfalls in einem zweiten Brandabschnitt sichergestellt. Durch die Möglichkeit der Verschlüsselung und den Einsatz sicherer Zugriffsprotokolle können auch Datenschutzstandards eingehalten werden. Darüber hinaus wird der Dienst zentral durch das Universitätsrechenzentrum betrieben, sodass für den SFB kein Administrationsaufwand mehr notwendig sein wird.

4.4 Datenschutz und Sicherheit

Datenschutz

Der Fokus der Forschung des SFB 933 liegt auf historischen schrifttragenden Artefakten. Bei diesen Forschungsdaten handelt es sich demnach nicht um personenbezogene Daten im Sinne des Datenschutzes, da dieser einen Bezug auf lebende Personen erfordert.

Dennoch sollen die Daten, die im Rahmen des SFB 933 erhoben und verarbeitet werden, mit geeigneten Maßnahmen und nach aktuellem Stand der Technik geschützt werden.

Zugriffskontrolle

Die Zugriffe auf sämtliche Systeme werden durch Accounts mit entsprechenden Berechtigungen umgesetzt, wobei das Minimalprinzip verwendet wird (nur die benötigten Berechtigungen werden vergeben). Nur die Administratoren, die den technischen Betrieb sicherstellen, verfügen über erweiterte Berechtigungen.

Der Zugriff auf die virtuelle Maschine ist ausschließlich über das verschlüsselte SSH-Protokoll mit Public-Key-Authentifizierung möglich. Ein direkter Zugriff auf den MTK-Speicher ist nicht möglich.

Die Webseiten des SFB sind weltweit zugänglich, die übrigen Dienste der VM nur aus den Netzen der Universität Heidelberg. Bis auf die öffentlichen Webseiten sind alle Dienste passwortgeschützt. Alle mit MTK-Online arbeitenden Forscher*innen des SFB 933 und ggfs. ihre Hilfskräfte erhalten ein personalisiertes Passwort, mit dem sie Zugang zur Datenhaltung ihres Teilprojekts bekommen.

Firewall

MTK-Online ist mit einer Firewall geschützt, um unberechtigte Zugriffe abzublocken. Folgende Regeln sind eingerichtet:

18 http://sds-hd.urz.uni-heidelberg.de/

Richtung	Protokoll	Port	IP
ausgehend	Alle	alle	0.0.0.0/0
eingehend	ICMP	alle	129.206.0.0/16
eingehend	TCP	22 (SSH)	0.0.0.0/0
eingehend	TCP	80 (HTTP)	0.0.0.0/0
eingehend	TCP	443 (HTTPS)	0.0.0.0/0
eingehend	TCP	1501 (TSM)	129.206.119.161/32
eingehend	TCP	3306 (MySQL)	129.206.0.0/16 und 147.142.0.0/16
eingehend	TCP	3690 (SVN)	129.206.0.0/16 und 147.142.0.0/16
eingehend	TCP	5666 (Nagios)	129.206.0.0/16

Backup

Für MTK-Online wird ein Backup am URZ gesichert. Das Backup wird täglich erstellt und umfasst den gesamten MTK-Speicher sowie sowie die Verzeichnisse

/boot
/etc
/home
/opt
/var

des MTK-Servers.

4.5 Datenmenge, Typen und Formate

Insgesamt sind auf dem MTK-Speicher derzeit ca. 388.000 Dateien mit einem Gesamtvolumen von 1,9 Tbyte gespeichert. Die wesentlichen Typen und Formate sind in folgender Tabelle aufgeschlüsselt:

Datentyp	Formate	Anzahl Dateien	Volumen (Kbyte)
Video	avi	6	517.296
	mov	10	1.005.344
	mp4	66	54.202.744
Geo-Referenzen	shp	75	155.080
	geojson	42	128.164
	xyz	7	129.668.480
	asc	1	566.580
	ply	4	72.903.712
Webseiten	flv	27	2.909.976
	html	12.883	894.948
	js	19.094	370.596
	swf	28	452.116
	webm	47	10.831.612
	ogv	42	4805.188
Dokumente	rtf	176	901.000
	pdf	588	2.394.100

Sonstige	xml	1.011	179.888
	sql	1.419	41.671.396
	zip	21	26.259.808
	gz	8	8.534.004
	rar	1	8.295.116

4.6 5 Nachhaltigkeit

4.7 Zugänglichkeit und Anschlussnutzung

Die Universität Heidelberg betreibt die beiden institutionellen Repositorien heidICON[19], die Bild- und Multimediadatenbank der Universitätsbibliothek, und heiDATA[20], das Forschungsdatenrepositorium des Kompetenzzentrums Forschungsdaten (KFD). Beide Repositorien erlauben es, Datensätze im Sinne von Open Data für die freie Nutzung zugänglich zu machen. Während heidICON auf Ton-, Bild- und Videodaten spezialisiert ist und entsprechende Benutzeroberflächen zu deren Wiedergabe bietet, ist heiDATA für Forschungsdaten aller Art geeignet.

Beide Repositorien stehen dem SFB 933 für die Ablage und Präsentation der Daten zur Verfügung. Datensätze werden mit den einschlägigen administrativen, technischen und deskriptiven Metadaten versehen, erhalten eine DOI und werden in der Heidelberger Hochschulbibliographie verzeichnet, wodurch sie recherchierbar und referenzierbar sind. In diesem Sinne sind nachhaltige Datenpublikationen möglich. Derzeit werden den Nutzern von heidICON und heiDATA keine Kosten in Rechnung gestellt.

Über die Datenpublikation in heidICON bzw. heiDATA hinaus bieten beide Repositorien die Möglichkeit, von weiteren Präsentationsoberflächen aus auf die hinterlegten Daten zuzugreifen. Dadurch ist es möglich, Datensätze dauerhaft im Repositorium zu publizieren und gleichzeitig auf aktuelle Anforderungen zugeschnittene Präsentationsoberflächen mit individueller Funktionalität bereitzustellen. Das Konzept für die langfristige Speicherung und Präsentation von Forschungsdaten des SFB 933 ist entsprechend in mehreren Stufen aufgebaut und sieht den Einsatz von mehreren Diensten vor, siehe Abbildung 2.

Abbildung 2: Übersicht über die Dienste und Merkmale der Archivierung und Präsentation von Daten.

19 https://heidicon.ub.uni-heidelberg.de
20 https://heidata.uni-heidelberg.de

4.8 Generische Präsentation und Publikation

Ein möglichst großer Teil der Forschungsdaten des SFB 933 soll für einen längeren Zeitraum aufbewahrt und öffentlich bereitgestellt werden (Archivierung, Präsentation und Publikation von Daten). Für eine solche Bereitstellung müssen die Daten zuvor aufbereitet werden. Die folgenden Kriterien werden bei der Auswahl der Forschungsdaten herangezogen, für die eine Publikation/Präsentation angestrebt wird:

- Die Datensammlung/Datenbank wird in Projektpublikationen verwendet und/oder zitiert.
- Die Datensammlung/Datenbank ist für die Veröffentlichung geeignet (ausreichende Qualität der erhobenen Daten, (Bild-)Rechte sind gesichert).
- Die Datensammlung/Datenbank bildet einen in sich abgeschlossenen und sinnvollen Wissensbestand ab (das Gegenteil könnte der Fall sein, wenn die Datenbank eine zufällige, fragmentarische oder unfertige Datensammlung wäre).
- Die Datensammlung/Datenbank enthält (zumindest auch) Daten, die vom Teilprojekt selbst erzeugt wurden, und nicht nur übernommene Datenbestände.
- Es ist davon auszugehen, dass in der Forschung außerhalb des SFB933 künftig an den Daten Interesse bestehen wird, bzw. dass sie Ausgangspunkt für neue Forschungen sein können.

Für die Publikation und Präsentation von Daten stehen in Heidelberg die Dienste heiDATA und heidICON zur Verfügung, die eine dauerhafte Datenablage vorsehen. Diese Dienste kommen für die Publikation und Präsentation der Forschungsdaten des SFB933 zum Einsatz. Aufgrund ihrer einfachen Handhabbarkeit und leichten Umsetzbarkeit haben wir diese regionale Lösung einer überregionalen Lösung vorgezogen und daher auch auf die noch im vergangenen Antrag vorgesehene Kooperation mit DARIAH verzichtet, zumal diese langfristig mit höheren Kosten verbunden gewesen wäre. Im Zuge der Datenpublikationen in heiDATA bzw. heidICON werden die zugrunde liegenden Datenbestände in das zurzeit im Aufbau befindliche und vom URZ und der UB gemeinsam entwickelte, OAIS-kompatible Langzeitarchivsystem heiARCHIVE überführt werden.

4.9 Überführung der Daten in heiDATA

Das Forschungsdatenrepositorium der Universität Heidelberg heiDATA basiert auf der Open-Source Forschungsdatenrepositorium Software dataverse[21]. Für die Präsentation der Forschungsdaten in heiDATA werden eine Reihe von Metadaten erhoben. Die folgende Tabelle zeigt einen Ausschnitt der erhobenen Metadaten beispielhaft am Datensatz „Late Antique Statue Bases of Lepcis Magna“:

Bezeichnung	Erklärung	Beispiel
Title	Full title by which the Dataset is known.	Late Antique Statue Bases of Lepcis Magna
Author	The person(s), corporate body(ies), or agency(ies) responsible for creating the work.	Bigi, Francesca (Heidelberg University, Seminar for Ancient History and Epigraphy and SFB 933)
Contact	The contact(s) for this dataset.	Witschel, Christian (Heidelberg University, Seminar for Ancient History and Epigraphy and SFB 933)
Description	A summary describing the purpose, nature, and scope of the Dataset.	Interactive and searchable map of the city of Lepcis Magna, displaying the geographical distribution within the city scape of the late antique honorary and

21 https://dataverse.org/

		building inscriptions. [...]
Subject	Domain-specific Subject Categories that are topically relevant to the Dataset.	Arts and Humanities
Keyword	Key terms that describe important aspects of the Dataset.	Leptis Magna (GND) http://d-nb.info/gnd/4111264-7 [...]
Language	Language of the Dataset	English
Producer	Person or organization with the financial or administrative responsibility over this Dataset	Bigi, Francesca (Indipendent Researcher)
Production Date	Date when the data collection or other materials were produced (not distributed, published or archived).	2016-11-01

Neben diesen deskriptiven Metadaten (in heiDATA „Citation Metadata“) stehen weitere Formulare für das Eingeben von fachspezifischen Metadaten zur Verfügung. Im Sinne der optimalen Nutzbarkeit werden die Daten möglichst ausführlich mit Metadaten beschrieben.

Darüber hinaus muss bei der Veröffentlichung von Forschungsdaten auch immer eine Lizenz mit angegeben werden, welche die Weiternutzung der Daten regelt. Hierbei orientiert sich das TP INF an den Creative-Common-Lizenzen (siehe Abschnitt 3.3).

4.10 Spezifische Datenpräsentationen

Für einen großen Teil der Forschungsdaten, die in heiDATA bzw. heidICON gespeichert werden, werden im Rahmen von MTK-Online zusätzlich interaktive dynamische Datenbank-Frontends erstellt. Durch die im Rahmen des SFB 933 entwickelte Plattform MTK-Online ist ein sehr hohes Maß an Spezialisierung bei der Präsentation und Publikation von Daten möglich. Diese Datenpublikationen bzw. Webpräsentation werden für einen Zeitraum von mindestens fünfzehn Jahren durch das URZ gehostet. Eine entsprechende Vereinbarung wurde mit dem Leiter des URZ und ehemaligen TPL des TP INF, Herrn Prof. Vincent Heuveline, getroffen.

4.11 Archivierung (heiARCHIVE)

Derzeit entwickelt das KFD ein digitales Langzeitarchiv, das dem SFB 933 gegen Ende der aktuellen Förderperiode zur Verfügung stehen wird. In diesem Archiv können Forschungs- und andere Daten langfristig, also über Zeiträume von 10 Jahren und mehr, ggf. auch unbefristet, eingelagert werden. Auf konzeptioneller Ebene ist es oberstes Ziel des Archivs, die Nutzbarkeit der in den Daten enthaltenen Informationen für den Menschen zu erhalten. Dazu führt das Archiv eine Datenkuratierung durch, die u.a. folgende Maßnahmen umfasst:

- Auswahl geeigneter Dateiformate für die Archivierung, wobei insbesondere die Dokumentation, Standardisierung und freie Verfügbarkeit der Formate wichtig sind.
- Formatvalidierung und ggf. Konvertierung in Standardformate.
- Ausführliche Dokumentation der Datensätze mit Metadaten zu Provenienz, Kontext, Inhalt, Lizenzen u.a.
- Datenhaltung in einem geschützten Bereich des Archivs ohne direkten Zugriff von Nutzern.

- Redundante Speicherung auf sicheren Speicherinfrastrukturen.
- Migration der eingelagerten Daten bei Wechsel der Speicherinfrastrukturen oder Veralten von Dateiformaten.

Wie in Abschnitt 4.2 beschrieben, werden alle Forschungsdaten des SFB 933 während seiner Laufzeit zentral in MTK-Online gespeichert. Spätestens am Ende der zweiten Förderperiode wird der gesamte Datenbestand auf seine Archivierungsfähigkeit und Aufbewahrungspflicht hin geprüft. Gemäß den Richtlinien der DFG sind grundsätzlich alle Forschungsdaten für mindestens 10 Jahre aufbewahrungspflichtig, die einer wissenschaftlichen Veröffentlichung bzw. Abschlussarbeit zu Grunde liegen (sofern sie nicht schon in einem Repositorium hinterlegt wurden (siehe Abschnitt 4.1 Konzeption der gemeinsamen Datenhaltung des SFB). In Zusammenarbeit mit den Wissenschaftler*innen des SFB identifiziert das TP INF alle Datensätze, die der Aufbewahrungspflicht unterliegen, und überführt sie in heiARCHIVE. Des Weiteren werden alle Datensätze identifiziert, die aus anderen Gründen archiviert werden sollen, z.B. nur unter erheblichem Aufwand oder gar nicht reproduzierbare Daten, oder anderweitig wertvolle Daten. Darüber hinaus unterstützt das TP INF die Mitglieder des SFB 933 bei individuellen Archivierungsvorhaben.

4.12 Data Governance

Die Teilprojekte legen für deren Forschungsdaten die Regelungen für den Zugriff und langfristige Erhaltung der Forschungsdaten fest, wobei das Teilprojekt INF ggf. unter Hinzunahme des Kompetenzzentrums Forschungsdaten (KFD) beratend zur Seite steht, und unterstützt die technische Umsetzung.

Es können während der Projektlaufzeit oder danach Fragen zum Zugriff und Erhaltung der Forschungsdaten entstehen, die sich nicht durch das jeweilige Teilprojekt beantwortet lassen, etwa weil die zuständigen Personen nicht mehr verfügbar sind. Ähnliche Fragestellungen beziehen sich auf die Entscheidung inwiefern spezialisierte Präsentationsoberflächen in MTK-Online erhalten bzw. wiederhergestellt werden sollen, falls auf Grund von IT-Sicherheitsmaßnahmen oder anderen Gründen ein großer personeller oder finanzieller Aufwand notwendig wäre. Solche Fragen erfordern eine fachbezogene Beurteilung und werden daher nicht durch das Teilprojekt INF selbst beantwortet. Für die Entscheidung ist daher während der Projektlaufzeit des SFB 933 Prof. Dr. Ludger Lieb zuständig. Nach Ende der Laufzeit geht diese Verantwortung über an die Leitung des Heidelberg Centre for Cultural Heritage (HCCH), derzeit Prof. Dr. Christian Witschel.

Heidelberg Zentrum Kulturelles Erbe
Geschäftsführender Direktor: Prof. Dr. Christian Witschel
Marstallhof 4, 69117 Heidelberg
Telefon: 06221/54 2231
E-Mail: christian.witschel<_at_>zaw.uni-heidelberg.de

Anhang 1

Tabelle: In der zweiten Förderperiode fertig gestellte Datenpublikationen

A01 UP 2	Bildpublikation der Kapitelle der Hagia Sophia, interaktiver Plan dazu: Druckvorlage zu drei Verteilungskarten	https://hagiasophia.materiale-textkulturen.de
A01 UP 2	Verteilung der spätantiken Statuenbasen in Lepcis Magna auf interaktivem Plan, Anbindung an die Epigraphische Datenbank Heidelberg (EDH)	https://lepcismagna.materiale-textkulturen.de
A01 UP 2	Interaktive Pläne zu den Mosaikinschriften auf den Fußböden von 13 Kirchenräumen in der spätantiken Provinz *Venetia et Histria*, Bildpublikation	https://mosaikinschrifen.materiale-texkulturen.de
A01 UP 3	Interaktiver Stadtplan von Pompeji mit Kartierung der Graffti und Dipinti, Bildpublikation	https://pompeji.materiale-textkulturen.de
B06	Bildpublikation zum *Welschen Gast*	https://wgd.materiale-textkulturen.de
B11	Annotationen musiktheoretischer Schriften inklusive Notenbeispielen und Diagrammen	https://musiktheorie.materiale-textkulturen.de
Ö	Onlinepublikation zum Sammelband „5300 Jahre Schrift“	https://5300jahreschrift.materiale-textkulturen.de

Tabelle: Datenpublikationen, die in der zweiten Förderperiode fertiggestellt, aber in der dritten Förderperiode noch ergänzt werden sollen

A01 UP 3	Interaktiver Stadtplan des mittelalterliche Rom: Inschriften im öffentlichen Raum	https://rom.materiale-textkulturen.de
A09	Ostraka-Datenbank	https://ostraka.materiale-textkulturen.de
C05	Datensammlung zu erzählten Inschriften	https://inschriftlichkeit.materiale-textkulturen.de