Frank Dickmann
Dr. Harry Enke
Patrick Harms

Technische Evaluation der Grid-Technologie für das Modellprojekt Kollaborative Datenauswertung und virtuelle Arbeitsumgebung – VirtAug

SOEB Arbeitspapier 2010-1

Forschungsverbund Berichterstattung zur sozioökonomischen Entwicklung der
Bundesrepublik Deutschland: Arbeit und Lebensweisen
Internet: http://www.soeb.de
Koordination: Soziologisches Forschungsinstitut (SOFI)
Friedländer Weg 31
D-37085 Göttingen
Projektleitung: Dr. Peter Bartelheimer

Überblick

In der quantitativ-empirischen sozialwissenschaftlichen Forschung sind IT-gestützte Datenanalysen schon lange ein wesentlicher Bestandteil des Forschungsprozesses. Von unterschiedlichsten Datenanbietern zur Verfügung gestellte, umfangreiche Datensätze werden mit Hilfe komplexer Verarbeitungsalgorithmen unter Verwendung von Statistiksoftware verarbeitet und ausgewertet. Mit steigendem Datenaufkommen stoßen jedoch sowohl die herkömmlichen Forschungsmethoden, die traditionellen Arbeitsweisen als auch die eingesetzten Rechnersysteme an ihre Grenzen. Die Arbeit in Forschungsverbünden, unterstützt durch die Nutzung moderner Kommunikationswege, durch eine den Forschungsprozess unterstützende Datenverwaltung und durch den Einsatz leistungsfähiger Rechnersysteme zur Entlastung lokaler Ressourcen, gewinnt immer mehr an Bedeutung.

Im vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Verbundprojekt „Sozioökonomische Berichterstattung“ (*soeb*) werden umfangreiche Erfahrungen in der kollaborativen sozialwissenschaftlichen Forschung gesammelt. Diese fließen ein in die Spezifikation und Entwicklung einer virtuellen Arbeitsumgebung für die empirischen Sozialwissenschaften. Ein solches System soll Forschungsverbünde in die Lage versetzen, Daten zu verwalten und abzulegen, gemeinsame Forschung, gegebenenfalls unter Einsatz von Hochleistungsrechnern, zu betreiben, und es soll den wissenschaftlichen Prozess bei jedem Schritt vom Zugriff auf Datenquellen bis hin zur Veröffentlichung von Ergebnissen begleiten und unterstützen. Außerdem sollen sowohl ein verbundinterner Wissens- und Datenaustausch, teilweise auf Basis multimedialer Kommunikation, als auch eine ausführliche Dokumentation des Forschungsprozesses, gegebenenfalls weit über ein Projektende hinaus, ermöglicht werden. Zu berücksichtigen sind hierbei vor allem die umfangreichen und restriktiven Sicherheitsbestimmungen, die sowohl durch den Gesetzgeber als auch durch die Datenanbieter vorgegeben sind.

Mit dem gleichfalls vom BMBF geförderten D-Grid ist eine leistungsfähige IT-Infrastruktur entstanden, die für den Aufbau von virtuellen Forschungsumgebungen für unterschiedlichste Wissenschaftsdisziplinen sehr gut nutzbare Komponenten bietet. Das Ziel der vorliegenden Expertise ist unter anderem eine Evaluation der D-Grid Infrastruktur hinsichtlich der Anwendung als Kernelement einer virtuellen Arbeitsumgebung für die quantitativ-empirische sozialwissenschaftliche Forschung. Dazu werden die Arbeitsabläufe der Forschungsprozesse analysiert, insbesondere deren kollaborative Komponenten, und Anforderungen an eine hierfür benötigte, softwarebasierte Infrastruktur abgeleitet. In einem nächsten Schritt werden existierende Systeme, unter anderem Komponenten des D-Grid, hinsichtlich ihres Einsatzes in der sozialwissenschaftlichen Verbundforschung untersucht. Aufgrund dieser Erkenntnisse sowie der gemeinsam mit Vertretern der Community und des *soeb* Projekts erarbeiteten Anforderungen werden eine Architektur sowie eine mögliche Implementierung einer virtuellen Arbeitsumgebung für die Sozialwissenschaften skizziert und wird grob deren Einsatz in Verbundprojekten dargestellt. Abschließend werden eine Klassifizierung der einzelnen Systemkomponenten für eine erste Umsetzungsphase sowie eine Abschätzung der zu erwartenden Implementierungsaufwände vorgenommen.

Inhaltsverzeichnis

1 Einleitung

Die vorliegende Expertise wurde im Rahmen des Modellprojekts „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung für die sozioökonomische Berichterstattung (VirtAug)“ für den Forschungsverbund Sozioökonomische Berichterstattung (*soeb*) erstellt. Für diese Aufgabe wurde zwischen dem Soziologischen Forschungsinstitut (SOFI) an der Georg-August-Universität Göttingen (http://www.sofi.de) und der D-Grid GmbH (http://www.d-grid-gmbh.de/) ein Forschungs- und Entwicklungsvertrag geschlossen. Die Bearbeitung lag bei Frank Dickmann (Abteilung Medizinische Informatik, Universitätsmedizin Göttingen), Dr. Harry Enke (Astrophysikalisches Institut Potsdam) und Patrick Harms (Abteilung Forschung & Entwicklung der Niedersächsischen Staats- und Universitätsbibliothek Göttingen).

Von 2000 bis 2004 und von 2005 bis 2009 förderte das Bundesministerium für Bildung und Forschung (BMBF) zwei Verbundvorhaben zur Erstellung eines ersten und zweiten „Berichts zur sozioökonomischen Entwicklung Deutschlands (*soeb* 1 und *soeb* 2, mehr Information auf der Projekt-Website http://www.soeb.de). Seit August 2009 führt das SOFI, das beide Forschungsverbünde koordiniert hatte, mit Förderung des BMBF eine fachöffentliche Konzeptphase durch, die in einen Vorschlag für Themen, Arbeitspakete und Akteure eines dritten Verbundvorhabens zur sozioökonomischen Berichterstattung münden soll.

Teil dieser Konzeptphase ist das Modellprojekt VirtAug, in dem das SOFI untersuchen und dokumentieren soll, „wie die datenbezogene Kooperation von Sozialwissenschaftler/innen an verschiedenen quantitativ-empirisch orientierten Forschungseinrichtungen und insbesondere eine kollaborative Auswertung der Mikrodaten von Forschungsdatenzentren künftig besser organisiert werden kann und welche IT-Verfahren eine solche Arbeitsweise unterstützen können“[1].

Die vorliegende Expertise evaluiert den Einsatz der Grid-Technologie als Basis für eine virtuelle sozialwissenschaftliche Arbeitsumgebung. Die ebenfalls vom BMBF geförderte D-Grid-Initiative (http://www.d-grid.de) entwickelt IT-Plattformen für den gemeinsamen Zugriff auf Rechnerkapazitäten, Daten und Programme. Das Projekt WissGrid (http://www.wissgrid.de) soll verschiedene Wissenschaftsdisziplinen dabei unterstützen, die im Rahmen von D-Grid entwickelten IT-Lösungen und -Werkzeuge für ihre fachspezifischen Arbeitsanforderungen zu nutzen.

Ziel der Expertise ist es, das SOFI und den Forschungsverbund Sozioökonomische Berichterstattung bei der Nutzung von Grid-Ressourcen und -Technologien zum Aufbau einer virtuellen Forschungsumgebung[2] zu beraten und zu unterstützen. Dazu wurden zunächst Arbeitsabläufe und Anwendungsszenarien aus der Sozioökonomischen Berichterstattung analysiert. Hieraus wurden entsprechende Systemanforderungen abgeleitet, und schließlich wurde untersucht, inwiefern die existierende D-Grid-Infrastruktur für eine mögliche Implementierung des Systems genutzt werden kann.

Der Forschungsverbund Sozioökonomische Berichterstattung diente also als konkreter „Anwendungsfall“ für die Überlegungen zur Schaffung einer virtuellen Arbeitsumgebung. Auch die weitere Entwicklungsarbeit, für den die Expertise den Aufwand abzuschätzen versucht (vgl. Abschnitt 5.), bedarf

---

[1] Schmidt, Tanja/Bartelheimer, Peter (2010): Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung. Zwischenbericht. Berlin/Göttingen. S. 1.

[2] Die Begriffe „virtuelle Arbeitsumgebung“ und „virtuelle Forschungsumgebung“ werden im akademischen Kontext synonym verwendet.

des Bezugs auf diesen konkreten Anwendungsfall. Dabei gehen aber sowohl das SOFI als auch die Bearbeiter davon aus, dass die Ergebnisse der Expertise und das hier skizzierte System in den Sozialwissenschaften breitere Anwendung finden könnten. Zwar zeichnet sich der Forschungsverbund Sozioökonomische Berichterstattung gegenüber anderen Forschungsvorhaben durch Anforderungen an die Replikation von Indikatoren, durch Verflechtungsbeziehungen zwischen einer Vielzahl von Arbeitspaketen und Analyseebenen sowie durch die Nutzung einer relativ großen Zahl verschiedener Mikrodatensätze aus. Jedoch sind die grundlegenden Arbeitsabläufe, die für diese Expertise zu berücksichtigen waren, durchaus charakteristisch für die Nutzung von Mikrodaten in der quantitativ-empirischen Forschungspraxis in den Sozialwissenschaften, weshalb eine virtuelle Arbeitsumgebung für den Forschungsverbund zugleich in der Lage sein sollte, eine ganze Reihe anderer sozialwissenschaftlicher Vorhaben zu unterstützen.

Die Expertise definiert noch keine vollständige Architektur. Vielmehr betrachtet sie Anforderungen an eine virtuelle Arbeitsumgebung und Umsetzungsmöglichkeiten mit Hilfe der D-Grid-Infrastruktur. Sie stellt die Komponenten der virtuellen Forschungsumgebung dar, welche für deren Implementierung erforderlich sind, und kann als Grundlage für einen umfassenden Systementwurf und ein darauf folgendes detailliertes Systemdesign dienen.

2 Soziologische Forschung

2.1 Forschungsszenarien in der sozioökonomischen Berichterstattung

Die Forschungspraxis im Umfeld der sozioökonomischen Berichterstattung (*soeb*) ist in einem Zwischenbericht des Modellprojekts VirtAug (Schmidt/Bartelheimer 2010, vgl. Fußnote 1) am Beispiel der Nutzung des Sozio-oekonomischen Panels (SOEP) und des Mikrozensus (MZ) durch Bearbeiter/innen aus verschiedenen Forschungseinrichtungen beschrieben worden.

Gemeinsam ist den verschiedenen Anwendungsszenarien, welche von der virtuellen Arbeitsumgebung zu unterstützen sind, dass Mikrodatensätze aus Personen- und Haushaltssurveys, Verwaltungsregistern oder Unternehmensbefragungen durch Forschungsdatenzentren (FDZ) oder durch Daten haltende statistische Ämter, Forschungseinrichtungen oder Verwaltungen zur Verfügung gestellt und von mehreren Projektpartner/inne/n im Forschungsverbund genutzt werden. Die FDZ bieten die von ihnen bereitgestellten Daten zur Weiterverarbeitung in einer für die direkte Auswertung mit Hilfe von Statistik-Programmen (z.B. SPSS, Stata, R) aufbereiteten Form an. Je nach Nutzungsbedingungen, die von den jeweiligen FDZ festgelegt werden, darf und kann die Datenverarbeitung für einen gewissen Zeitraum beim Nutzer (Scientific Use Files) oder ausschließlich beim FDZ (Onsite-Nutzung, kontrollierte Datenfernverarbeitung) erfolgen. Die Ursache für diese teils sehr restriktiven Nutzungsbedingungen liegt in der datenschutzrechtlichen Sensibilität der Daten: Diese enthalten unter Umständen sehr detaillierte Informationen zu Personen oder Organisationen, deren Anonymität gewahrt werden muss.

Für die Datenauswertung werden verschiedene Statistik-Programme, wie z.B. SPSS, Stata oder R, eingesetzt. Die Programme funktionieren nach ähnlichen Prinzipien: Sofern die Daten im programmspezifischen Format vorliegen, können Verarbeitungsalgorithmen auf den Daten ausgeführt werden. Diese Algorithmen werden als Syntax bezeichnet und sind dem Quellcode von Computerprogrammen sehr ähnlich. Eine Syntax definiert die logischen Schritte, die für die statistische Auswertung der Daten notwendig sind. Das Format der Syntax ist definiert durch das jeweils verwendete Statistik-Programm. Dieses interpretiert die Anweisungen der Syntax und verarbeitet die Daten entsprechend. Im Verlauf einer statistischen Auswertung können diverse Zwischenergebnisse entstehen, die wiederum Grundlage für einen darauf folgenden Auswertungsschritt sein können. So entstehen iterative Auswertungsvorgänge oder komplexe Workflows. Die Ergebnisse einer Auswertung werden abschließend in Berichten aufbereitet und zusammengefasst.

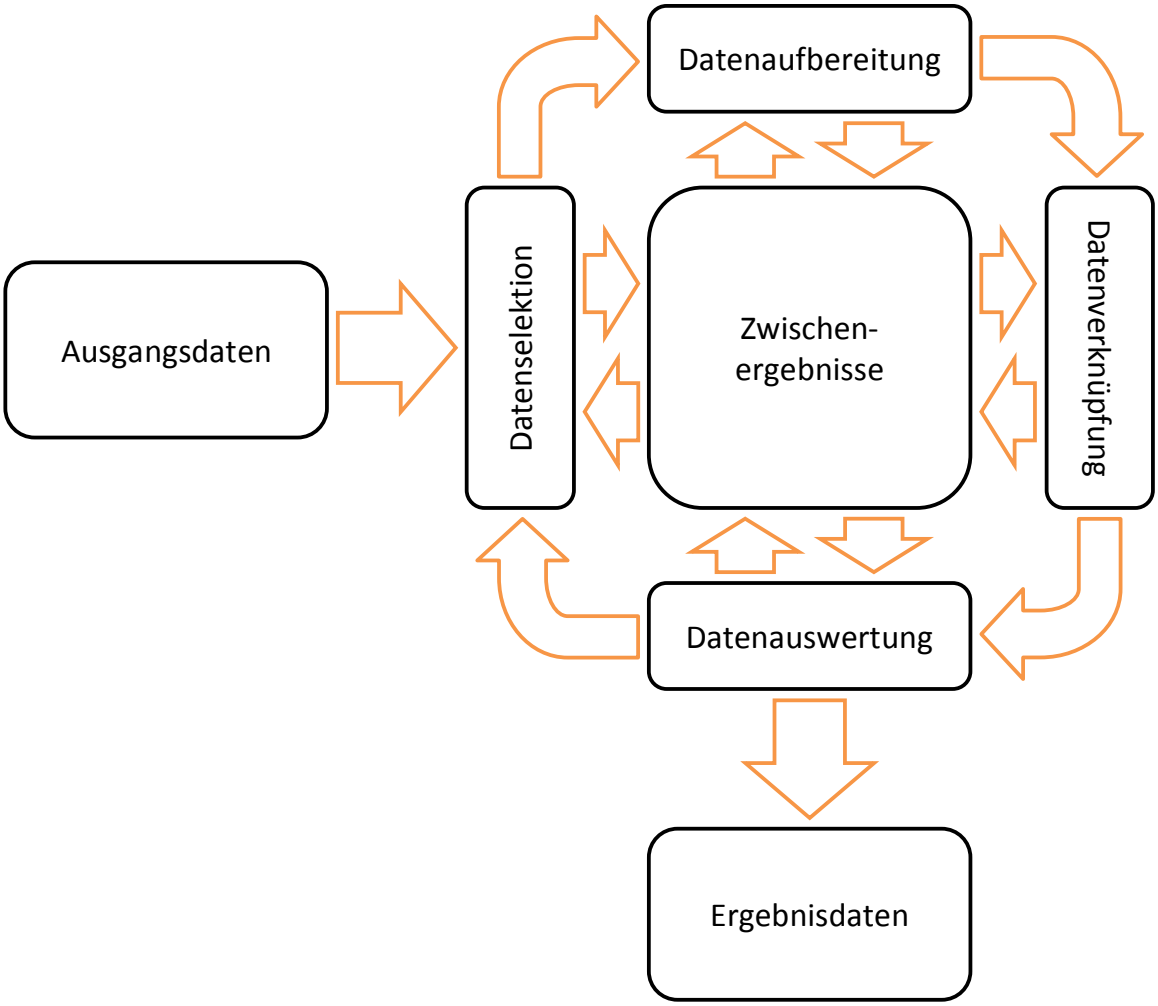

Abbildung 1 - Generische Darstellung des sozialwissenschaftlichen Forschungsprozesses

Abbildung 1 beschreibt generisch einen solchen Forschungsprozess. Ausgehend von den Ausgangsdaten erfolgt eine Datenselektion durch die Forscher bzw. Arbeitsgruppe. Die selektierten Daten werden anschließend als Zwischenergebnisse gespeichert und aufbereitet. Die Aufbereitung umfasst zum Beispiel das Anpassen von Datenformaten oder Metriken. Daraufhin erfolgt ggf. eine Verknüpfung mit weiteren Zwischenergebnissen, die wiederum als Zwischenergebnis gespeichert wird. Im Idealfall folgt dann die Datenauswertung, woraus die Ergebnisdaten gewonnen werden.

Die Arbeit des Forschungsverbunds Sozioökonomische Berichterstattung hat einen ausgeprägt explorativen Charakter, weshalb der Forschungsprozess nicht mit einer einzigen Datenauswertung beendet ist und ggf. weitere Iterationen der vorangegangenen Schritte durchlaufen werden (siehe Abbildung 1). Am Anfang eines Projekts ist nicht immer abzusehen, wie und in welcher Art die verschiedensten Datenquellen und Zwischenergebnisse mit einander kombiniert werden müssen und können. Außerdem ist noch nicht festgelegt, welche Datenquellen überhaupt Einsatz finden. Schließlich werden statistische Berechnungen unter Umständen mehrfach, jedoch mit anderen Parametern, durchgeführt, um ein optimales oder aussagekräftiges Ergebnis zu erzielen. Es ergeben sich daher sehr viele unterschiedliche Varianten in der Arbeitsweise und den Abläufen. Diese folgen jedoch ähnlichen Schemata und lassen sich in immer wiederkehrende Einzelschritte aufteilen.

Eine virtuelle Arbeitsumgebung kann wesentliche Teilschritte dieser Arbeitsabläufe bzw. Einzelschritte unterstützen und somit zu einer Steigerung der Effizienz in den Arbeitsabläufen und somit im For-

schungsprozess beitragen. Damit die dafür notwendigen Funktionalitäten jedoch erarbeitet werden können, muss der sozialwissenschaftliche Arbeitsprozess des Forschungsverbunds in grundlegende, frei kombinierbare Teilschritte zerlegt werden. Die virtuelle Arbeitsumgebung muss dann gezielt die für die Teilschritte notwendigen Funktionalitäten zur Verfügung stellen und implementieren. So sollte es z.B. generell und zu jedem Zeitpunkt möglich sein, aus der virtuellen Arbeitsumgebung heraus die Ausführung einer Syntax auf einen Datensatz mit Hilfe eines Statistik-Programms zu starten. Dafür sollten nur minimale Vorbedingungen zu erfüllen sein, um den frei kombinierbaren Charakter dieser Funktionalität zu gewährleisten. Gleiches gilt z.B. für die Ablage von Datensätzen, Syntaxdateien oder auch Ergebnissen.

Die virtuelle Arbeitsumgebung sollte außerdem infrastrukturelle Angebote enthalten, um die allgemeine Arbeitsweise des Forschungsverbunds effizienter und übersichtlicher gestalten zu können. So sollte sie z.B. Funktionalitäten bieten, mit denen eine kollaborative Arbeitsweise, z.B. zur gemeinsamen Erstellung und Bearbeitung einer Syntaxdatei, vereinfacht wird.

2.2 Begriffsdefinitionen

Die Datensätze, die Forschungsdatenzentren (FDZ) oder andere Daten haltende Stellen verwalten und zur Verfügung stellen, werden als Originaldatensätze bezeichnet. Allerdings können diese Daten nicht immer als statisch vorausgesetzt werden, da die FDZ gelegentlich Änderungen an den Daten vornehmen – zum Beispiel nachträgliche Fehlerkorrekturen, Formatänderungen oder Ergänzungen. Korrekturen werden etwa notwendig, wenn im Nachhinein festgestellt wird, dass eine Kodierung oder eine generierte Variable fehlerhaft sind. Formatänderungen sind wiederum erforderlich, wenn ein Datensatz z.B. durch neue Erhebungen ergänzt wird oder eine Kodierung für einen bestimmten Wert nicht mehr zeitgemäß ist (z.B. wenn sich in einer Wiederholungsbefragung über lange Zeiträume bei Einkommensgrößen die Währung von DM zu Euro ändert).

Ein besonderes Ausgangsdatenformat bilden die so genannten Scientific-Use-Files (SUF). Hierbei handelt es sich um Datensätze, die von einem FDZ für ein konkretes Forschungsvorhaben zur weiteren Verarbeitung außerhalb des FDZ herausgegeben werden. Diese Daten dürfen je nach Nutzungsbedingungen z.B. nur von bestimmten Personen und in einem bestimmten Kontext analysiert werden. Ob es sich bei den SUF um vollständige oder reduzierte Datensätze handelt, hängt von den Nutzungsregelungen, der datenschutzrechtlichen Relevanz und dem herausgebenden FDZ ab. Die Nutzungsrechte an Scientific-Use-Files erlöschen zu einem jeweils definierten Zeitpunkt. Die Nutzer sind dann verpflichtet, die Scientific-Use-Files inklusive aller Kopien zu löschen. Außerdem können die Regelungen auch das Löschen von auf Basis der Scientific-Use-Files entstandenen Dateien verlangen.

In einem ersten Auswertungsschritt werden Teildatensätze aus den Originaldatensätzen gebildet, auf denen dann die eigentliche Aufbereitung, Verknüpfung und Auswertung erfolgt. Diese werden als Arbeitsdatensätze bezeichnet. Sie sind die Grundlage für eine bestimmte Forschungsarbeit. Arbeitsdatensätze können aus einer für die jeweilige Analyse relevanten Auswahl von Fällen und Variablen bestehen; in ihnen können aber auch Informationen aus verschiedenen Ausgangsdateien (z.B. Personen- und Haushaltsinformationen, Querschnitt- und Längsschnittinformationen) zusammengeführt sein. Sie können in unterschiedlichsten Formaten vorliegen. Für die Verarbeitung in Statistik-Programmen werden meist die programmspezifischen Formate verwendet. Zum Austausch der Daten in einem kollaborativen Forschungsprozess kann es sinnvoll sein, die Daten in gebräuchlichere Formate umzuwandeln.

Die Unterscheidung zwischen Mikro- und Makrodatensätzen bezieht sich auf die Analyseeinheit, für die Informationen bereitgestellt werden. Bedeutsam ist sie wegen der unterschiedlich strengen Daten-

schutzanforderungen: befragte Personen, Institutionen oder Unternehmen dürfen in Mikrodatensätzen nicht re-identifizierbar sein. Daher unterliegen Mikrodaten, die meist personenbezogene Informationen enthalten, oft eingeschränkteren Nutzungsbedingungen als höher aggregierte Makrodaten, deren Informationsgehalt für eine Re-Identifizierung von Personen nicht mehr ausreichend ist.

Von Ausgangsdaten und Arbeitsdatensätzen sind die Ergebnisdaten zu unterscheiden, die als Grundlage für Publikationen dienen. Diese werden i.d.R. in spezifischen Ausgabeformaten der Statistik-Programme abgelegt. Zum Austausch der Daten in einem kollaborativen Forschungsprozess ist es sinnvoll und erforderlich, die Daten in gebräuchliche Formate wie PDF, CSV, TXT oder JPEG zu speichern. Die Ergebnisdaten werden insbesondere im Kontext von Arbeitsgruppen des Forschungsverbunds verwendet, um daraus Publikationen zusammenzustellen. Ergebnisdaten weisen in der Regel keine Einzeldaten mehr aus, sondern aggregierte Ergebnisse der jeweiligen Forschungsarbeit. In Ergebnistabellen werden aus methodischen Gründen (Repräsentativität) oder aufgrund entsprechender Datenschutzanforderungen der FDZ keine Informationen ausgewiesen, die nur für einzelne oder wenige Fälle zutreffen. Eine Verpflichtung zum Löschen dieser Daten, wie sie bei Scientific-Use-Files oder Arbeitsdatensätzen gelten kann, besteht daher nicht.

Originaldatensätze, Arbeitsdatensätze und Ergebnisdaten sind sich strukturell sehr ähnlich. Sie können z.B. alle in statistikprogrammspezifischen Formaten vorliegen oder in Forschungsprozessen Anwendung finden. Daher werden diese Begriffe im Rahmen der vorliegenden Expertise unter der Bezeichnung Forschungsdaten zusammengefasst.

Für die Verarbeitung der Forschungsdaten werden im arbeitsteiligen Forschungsprozess, wie bereits in der Einleitung beschrieben, verschiedene Statistik-Programme eingesetzt. Die Verarbeitungsschritte werden in Form von Programmanweisungen kodiert, die als Syntax bezeichnet werden. Eine Syntax beschreibt

- auf welchen Datensätzen (Originaldatensätze, Arbeitsdatensätze, etc.) und
- in welcher Reihenfolge bestimmte Berechnungen (statistische Algorithmen, Umkodierungen, Verknüpfungen, etc.) erfolgen sollen und
- in welcher Form die Ergebnisse dieser Berechnungen (Arbeitsdatensätze, Ergebnisdaten, etc.) ausgegeben werden.

Außerdem kann eine Syntax mit Kommentaren versehen werden, die vom Statistik-Programm bei der Verarbeitung ignoriert werden. Die Syntax ist daher das wichtigste Format, in dem wissenschaftliche Erkenntnisse zur Methodik für den weiteren Arbeitsprozess dokumentiert werden. Sie dient zur Durchführung einer Forschungsarbeit und kann gleichzeitig zur Dokumentation des Forschungsprozesses (sowohl durch die Kommentare als auch durch die kodierten logischen Abläufe) eingesetzt werden.

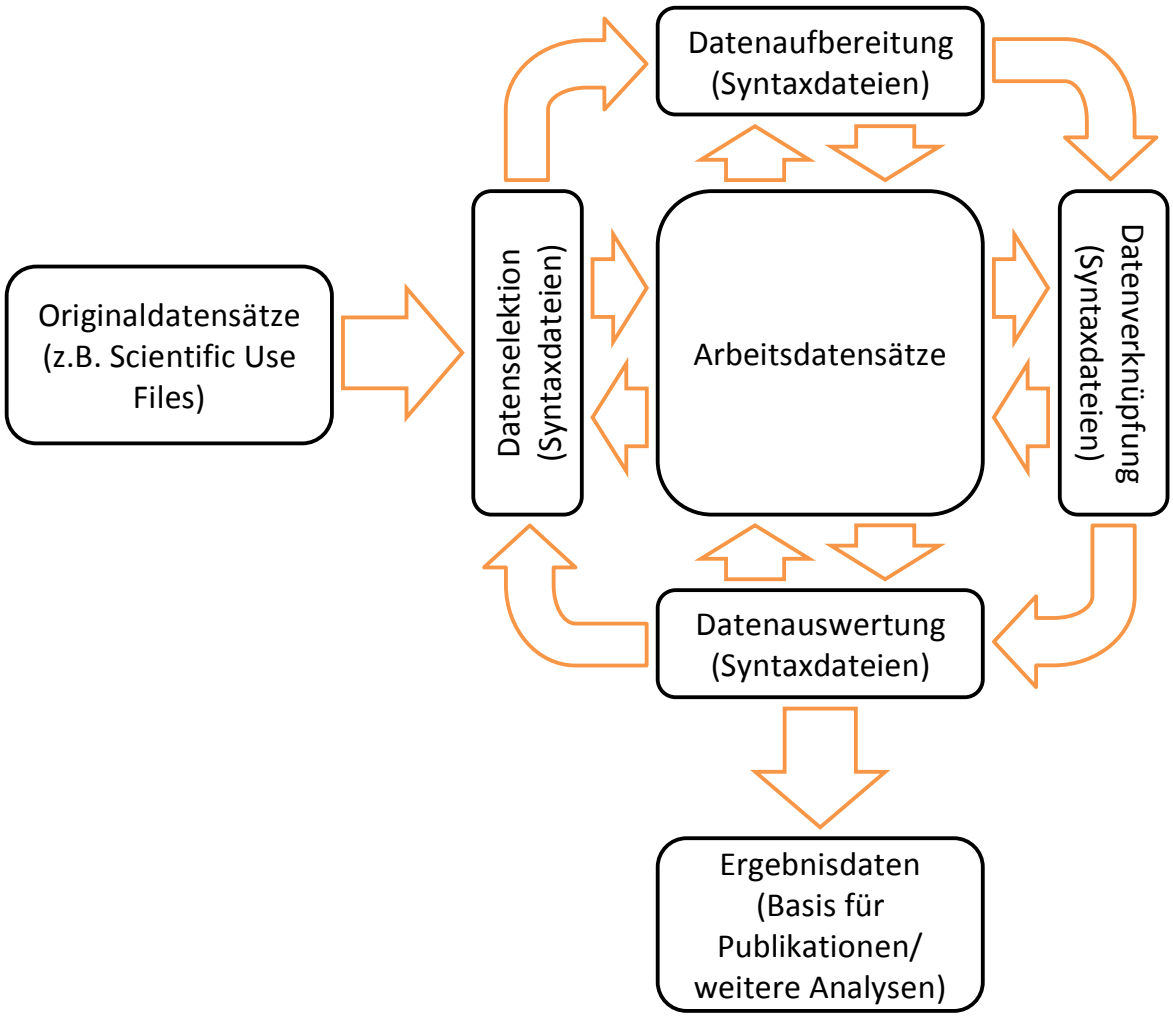

Abbildung 2 – Einordnung der verwendeten Datenarten im generischen Forschungsprozess

Trotz ihrer Bedeutung ist die Syntax, im Gegensatz zu den anderen Datenarten, eher „leichtgewichtig". Sie wird in Form einfacher Textdateien abgespeichert, deren Größe meist nicht mehr als wenige Kilobyte überschreitet (im Folgenden Syntaxdateien genannt).

Zu den Ausgangsdaten im Forschungsprozess der sozioökonomischen Berichterstattung zählen die von den FDZ bereitgestellten Originaldatensätze (Scientific-Use-Files, im Rahmen der Fernverarbeitung bereitgestellte Daten, onsite verfügbare Daten) sowie Arbeitsdatensätze oder Ergebnisdaten aus anderen Forschungsprozessen. Die Datenselektion, -aufbereitung, -verknüpfung und -auswertung erfolgt durch Algorithmen aus den Syntaxdateien. Als Ergebnis jedes dieser vier Arbeitsschritte werden jeweils Arbeitdatensätze erzeugt, die ihrerseits als Ausgangsbasis für den entsprechenden nächsten Arbeitsschritt verwendet werden können. Die Ergebnisdaten stellen das Resultat der Forschung dar und werden für Berichte, Publikationen sowie weitere Analysen genutzt. Abbildung 2 verbindet die generische Beschreibung des Forschungsprozesses aus Abbildung 1 mit den Datenarten.

Wichtig ist hierbei, dass die in einem Forschungsprozess eingesetzten Daten immer aus demselben Quellkontext stammen. Die Originaldatensätze sind aufgrund von Nutzungsbeschränkungen nicht immer frei miteinander verknüpfbar.

Weitere Informationen und Begriffsdefinition finden sich im Zwischenbericht zum Modellprojekt „Kollaborative Datenauswertung und Virtuelle Arbeitsumgebung" (VirtAUg; vgl. hierzu Fußnote 1).

2.3 Funktionale Anforderungen an eine virtuelle Arbeitsumgebung

Eine virtuelle Arbeitsumgebung dient der Unterstützung der sozialwissenschaftlichen Forschungspraxis in ihren existierenden Vorgehensweisen und Prozessen. Sie darf und kann jedoch keinen Einfluss auf diese Prozesse nehmen. Stattdessen sollte das vorrangige Ziel sein, eine effektive Unterstützung für die Prozesse anzubieten. Dabei soll die virtuelle Arbeitsumgebung bisher weniger etablierte oder technologisch anspruchsvollere Vorgehensweisen erlauben und unterstützen. Unvermeidlich ist jedoch, dass sie Einfluss auf die Entwicklung der dazugehörigen neuen Prozesse nehmen sowie Konventionen und Standardisierungen einbringen wird. Dies kann auch für die existierenden Prozesse gelten, wodurch gegebenenfalls ein Fortschritt für die beteiligten Wissenschaftler durch erhöhte Transparenz und Vergleichbarkeit im Forschungsprozess entsteht.

Insgesamt stellt die virtuelle Arbeitsumgebung einen technischen Werkzeugkasten für den Forschungsverbund dar. Dieser beinhaltet verschiedene Werkzeuge, um die Forscher/innen in ihrer täglichen Arbeit unterstützen. Die Werkzeuge sollen den bisher in der sozioökonomischen Berichterstattung eingesetzten Werkzeugen sehr ähnlich, dabei allerdings für ihren Einsatzbereich optimiert sein. Dafür ergänzen sie existierende Standards oder Prinzipien, bzw. weichen leicht von diesen ab, um zusätzliche Funktionen zu bieten, die den Anwendern einen bestmöglichen und effizienten Einsatz der Werkzeuge im Rahmen der Forschung erlauben[3]. Hinzu kommt, dass die virtuelle Arbeitsumgebung Hintergrundfunktionen übernimmt, die dem Benutzer den Umgang mit den Werkzeugen einfacher und komfortabler, dabei aber auch sicherer gestalten sollen[4].

Die im Forschungsprozess benötigten Werkzeuge sind in Abbildung 3 in Form von Basisfunktionen dargestellt (schwarze Kästen), die der Benutzer über entsprechende Schnittstellen (schwarze Ringe an den Kästen) nutzen kann. Im Hintergrund übernimmt die virtuelle Arbeitsumgebung dabei Querschnittsaufgaben bzw. stellt bestimmte allgemeingültige Anforderungen der Verbundarbeit sicher.

[3] Veranschaulichen lässt sich dies z.B. mit Haushaltsmessern. Mit einem handelsüblichen Messer kann man viele Dinge schneiden. Nur stößt ein solches Messer bei bestimmten Stoffen schnell an seine Grenzen. Leder oder ähnlich harte Materialien sind mit großem Zeitaufwand sicherlich noch schneidbar. Mit Effizienz und Effektivität hat dies jedoch nichts zu tun. Daher wird für solche Aufgaben ein Spezialmesser benötigt, welches sowohl die grundlegenden Aufgaben des Schneidens von Dingen (bisheriger Standard), als auch Spezialaufgaben, wie das Schneiden von Leder (Ergänzung des Standards) übernehmen kann. Bei einem solchen Messer ist aufgrund seiner Schärfe jedoch mehr Vorsicht geboten (leichte Abweichung vom Standard).

[4] Auf das Beispiel mit dem Messer bezogen bedeutet dies, dass das Messer außerdem mit einer dafür angefertigten Tasche geliefert wird, welche das Messer sicher verwahrt und dabei automatisch reinigt, pflegt und gegebenenfalls nachschärft.

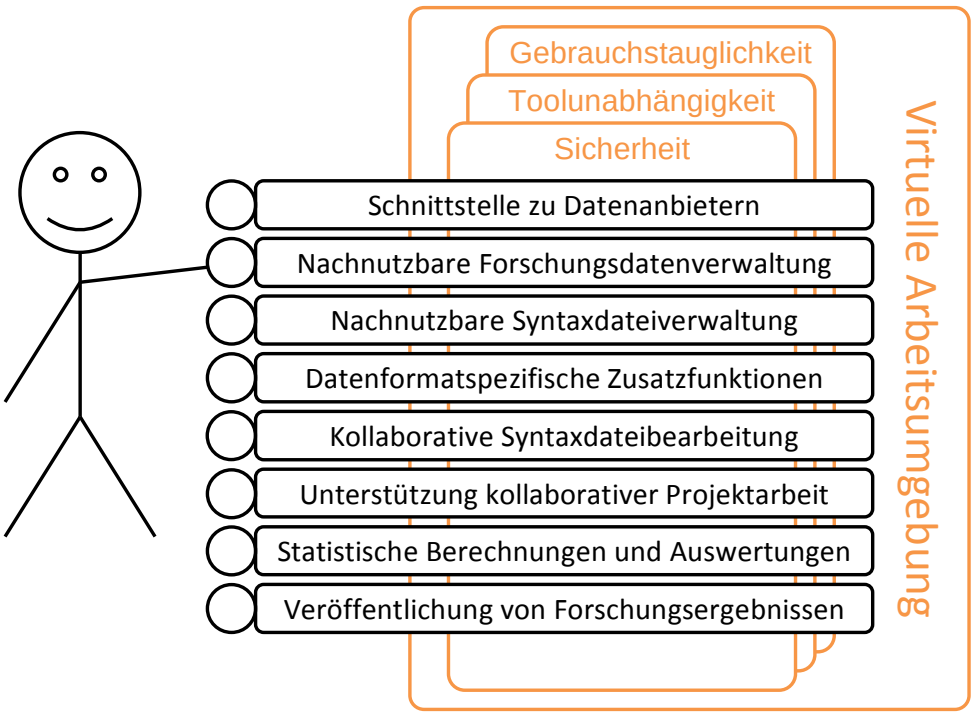

Abbildung 3 - Funktionale Anforderungen an eine virtuelle Arbeitsumgebung

Die Arbeit des Forschungsverbunds mit der virtuellen Arbeitsumgebung wird sich zunächst hauptsächlich auf die Ablage und Verwaltung von Forschungsdaten und Syntaxdateien beschränken. Hierbei sollen vor allem Möglichkeiten zum effektiven Zugriff auf Daten, gegebenenfalls direkt beim FDZ, Umsetzung finden. Allerdings würde ihre Verwendung im letzteren Fall auch die Schnittstelle zwischen Forschungsverbund (Datennutzung) und Datenprovidern verändern. Die Einrichtungen der Forschungsdateninfrastruktur beanspruchen Kontrolle über die Zugriffsweisen auf ihre jeweiligen Daten und sind daher bei der technischen Veränderung auf Seiten der DatennutzerInnen, die diese berührt, als interessierte Partei einzubeziehen. Hier könnten z.B. durch Einführung von wohlbekannten Sicherheitsmechanismen wie auf Zertifikaten basierender Authentisierung und Autorisierung neue Möglichkeiten des Zugriffs zusammen mit den Datennutzer/inne/n entwickelt werden.

In jedem Fall wird die virtuelle Arbeitsumgebung die von den FDZ vorgegebenen Nutzungsbeschränkungen beachten und deren Einhaltung durch geeignete Maßnahmen unterstützen. Dabei wird sie die Benutzer jedoch so unterstützen, dass diese sich während ihrer alltäglichen Arbeit so wenig wie möglich mit diesen Aspekten beschäftigen müssen.

Die virtuelle Arbeitsumgebung wird weiterhin spezielle Zusatzfunktionen anbieten, welche den Umgang mit den verschiedenen eingesetzten Datentypen und -formaten erleichtern. Hierzu zählt die Unterstützung zur Arbeit mit Metadaten, welche die Basis für die weitere Datenverarbeitung darstellen. Für die Arbeits- und Ergebnisdaten sowie für die Syntaxdateien wird sie die Erstellung und Erfassung der notwendigen Metadaten weitestgehend automatisieren (sofern diese innerhalb der virtuellen Arbeitsumgebung bearbeitet werden) und einfache Schnittstellen für die erforderliche Interaktion mit den Nutzern bereitstellen. Für bestimmte Daten, vor allem aber für Syntaxdateien, ist außerdem eine feingranulare Erfassung und Dokumentation von Veränderungen sinnvoll. Dies wird analog zu Versionskontrollsystemen von der virtuellen Arbeitsumgebung übernommen. Nicht zuletzt sollen, sofern möglich, Konvertierungsroutinen für unterschiedliche im Forschungsverbund eingesetzte Datenformate zur Verfügung stehen.

Außerdem soll die kollaborative Arbeit von Forschern in Verbünden durch sinnvolle Funktionalitäten vereinfacht und unterstützt werden. So soll z.B. die gemeinsame Arbeit an Daten und die Kommunikation untereinander so weit wie möglich mit Hilfe geeigneter Werkzeuge unterstützt werden.

Soweit möglich, soll die virtuelle Arbeitsumgebung auch das Ausführen statistischer Berechnungen und Auswertungen erlauben. Auch wenn diese Funktion nicht im Mittelpunkt steht, kann die virtuelle Arbeitsumgebung die für die Ablage, Bearbeitung und Verarbeitung von Daten gegebenenfalls benötigten großen Speicher- und Rechenkapazitäten transparent und einfach nutzbar für Anwender zur Verfügung stellen. So können z.B. Ressourcen des D-Grid einfach eingebunden werden, was den Aufwand für die eigene Bereitstellung solcher Infrastrukturen sowie deren Betrieb erheblich verringert.

Weiterhin soll es möglich sein, Forschungsergebnisse mit einfachen Hilfsmitteln in digitaler Form zu veröffentlichen. Dazu werden Funktionen zur Ablage und Konvertierung von Datenformaten bereitgestellt, welche für Veröffentlichungen relevant sind (z.B. verschiedene Grafikformate). Außerdem stehen Möglichkeiten zur einfachen Bearbeitung und Ergänzung einer Internetpräsenz zur Verfügung.

Die virtuelle Forschungsumgebung soll ergänzend dazu Funktionalitäten bieten, die zwar für die tägliche Arbeit weniger relevant sind, für ein Forschungsprojekt jedoch eine sehr große Bedeutung haben. So sollen z.B. die hohen Sicherheitsanforderungen an den Umgang mit Forschungsdaten, vor allem mit Originaldatensätzen, beachtet werden. Außerdem sollen Nebenaspekte Berücksichtigung finden, die erst die effektive Nutzung der virtuellen Arbeitsumgebung ermöglichen.

Damit erstreckt sich der Einsatzbereich der virtuellen Arbeitsumgebung über den gesamten Forschungsprozess. Alle Datenarten können zu jeder Zeit und unter Beachtung aller relevanten Aspekte während einer Forschungsarbeit abgelegt und verwaltet werden. Außerdem wird die Erstellung und Bearbeitung von Syntaxdateien unterstützt und durch sinnvolle Funktionen erleichtert. Ergänzt werden diese Funktionalitäten um die IT-Unterstützung für Kollaboration mit anderen Wissenschaftlern. Die einfache Ausführung statistischer Berechnungen aus der virtuellen Arbeitsumgebung heraus erleichtert gegebenenfalls weitere Schritte im Forschungsprozess. Schließlich können die Ergebnisdaten direkt veröffentlicht und anderweitig zur Verfügung gestellt werden.

Durch die beschriebenen Funktionen soll die virtuelle Arbeitsumgebung eine transparente und gut dokumentierte Basis für den Austausch der Ergebnisse zwischen den Nutzern bieten, die sich so wenig wie möglich mit den technischen Unterschieden von Dateiformaten, Features der Datenprovider usw. befassen müssen und sich mehr auf die eigentliche wissenschaftliche Arbeit mit den Daten konzentrieren können.

3 Anforderungsanalyse

Aus den beschriebenen Abläufe in der sozioökonomischen Berichterstattung sowie der sozialwissenschaftlichen Forschung im Allgemeinen und aus dem damit einhergehenden Bedarf an Funktionalitäten ergeben sich technischen Anforderungen an eine Infrastruktur, auf deren Basis eine virtuelle Arbeitsumgebung entwickelt werden kann. Diese Anforderungen sollen in diesem Kapitel benannt werden. Die folgenden Kapitel werden diese Anforderungen immer wieder aufgreifen und bei der Erarbeitung von Lösungsvorschlägen beachten.

3.1 Datenbereitstellung

Die Bereitstellung der in der sozioökonomischen Berichterstattung verwendeten Daten weist Besonderheiten auf, die mit dem Charakter der Quellen dieser Daten eng verknüpft sind. Inhaltlich werden die Originaldatensätze aus Umfragedaten der amtlichen Statistik (z.B. Mikrozensus), der Sozialwissenschaften (z.B. sozio-oekonomisches Panel) oder von Wirtschaftsverbänden und -instituten sowie aus Verwaltungsdaten (etwa gesetzlich vorgeschriebenen Meldungen an Sozialversicherungen) gewonnen. Je nach Grad der Erfassung von Details oder Grad der Aggregation werden diese Daten wegen datenschutzrechtlichen Auflagen wie auch wegen des Vertrauensschutzes nur unter besonderen Vorkehrungen (z.B. Onsite-Rechnen) oder nur mit besonderen Auflagen (wie beim Fernrechnen) und generell durch personengebundene und oft auch zeitlich limitierte Vereinbarungen für wissenschaftliche Auswertungen zur Verfügung gestellt.

Erst durch eine hinreichend weit fortgeschrittene Anonymisierung werden diese Auflagen erfüllt und die Verwendungs- und Zugangsbeschränkung kann entfallen. Die Auflagen haben nicht nur zur Konsequenz, dass Daten nur bestimmten und (dem FDZ) bekannten Personen zugänglich sind, sie wirken auch bis in die Formen der Bereitstellung hinein.

Bei der restriktivsten Form des Zugangs zu Datensätzen werden diese den Forscher/inne/n nicht zugänglich gemacht, es werden nur Syntax-Dateien für die Statistik-Programme akzeptiert, die Forscher/innen ausschließlich am Beispiel von Dummy-Datensätzen entwickeln können. Das Starten der Programme erfolgt durch Angestellte des entsprechenden FDZ. Die Ergebnisse werden erst nach Prüfung durch das FDZ an die Forscher/innen weitergegeben. Oder die Forscher/innen dürfen nur an eigens dafür vorgesehenen Arbeitsplätzen im FDZ an den Daten bzw. an den Ergebnissen arbeiten, ohne eine Möglichkeit der Speicherung auf einem elektronischen Medium. Für diese restriktive Form des Schutzes der Datensätze kann die virtuelle Arbeitsumgebung keine Erleichterung schaffen.

Bei etwas weniger restriktiven Auflagen hingegen ist eine Verbesserung der Sicherheitsstruktur durch die virtuelle Arbeitsumgebung für einzelne Forscher/innen auch mit der Möglichkeit verbunden, effizientere Formen der Bereitstellung von Ausgangsdaten zu unterstützen, sei es durch verschlüsselte Verbindungen, um z.B. Scientific-Use-Files über das Netz zu beziehen, oder durch auf Zertifikaten basierende Authentifzierungsverfahren, wenn es um die Nutzung des Fernrechnens geht. Durch eine Ablage in einem Repository ergeben sich auch vielfältige Möglichkeiten, die (erlaubte) gemeinsame Nutzung von solchen Dateisätzen zu erleichtern, oder auch die erforderlichen Löschungen vorzunehmen, ohne damit gleich den ganzen Kontext zu verlieren, ja ggf. auch die Speicherung solcher Datensätze beim FDZ zu realisieren, um die referenzierten Datensätze unter der Kontrolle des jeweiligen FDZ evtl. auch wieder erhalten zu können.

Anforderung 1. ***Die aus den Datenschutzauflagen resultierenden Nutzungsbeschränkungen der Ausgangsdaten müssen von der virtuellen Arbeitsumgebung unterstützt werden.***

Anforderung 2. ***Die Sicherheitsinfrastruktur der virtuellen Arbeitsumgebung soll eine Vereinfachung der Bereitstellung von Scientific-Use-Files gestatten.***

Anforderung 3. ***Die Sicherheitsinfrastruktur der virtuellen Arbeitsumgebung soll eine effizientere Nutzung von Fernrechnen unterstützen.***

Anforderung 4. ***Die virtuelle Arbeitsumgebung soll eine Verbesserung der gemeinschaftlichen Nutzung von Ausgangsdaten unterstützen.***

Von besonderer Bedeutung bei nur sehr eingeschränkt nutzbaren Ausgangsdaten sind mögliche Datensicherungen durch die FDZ selbst, um die Anforderungen an den Datenschutz und die Nutzungsrechte zu gewährleisten. Gleichzeitig kann so der Anspruch an gute wissenschaftliche Praxis (Aufbewahrungspflicht) erfüllt werden. Etwa bietet das Institut für Arbeitsmarkt- und Berufsforschung (IAB) bereits einen entsprechenden Dienst an, bei dem Forscher gezielt die genutzten Ausgangsdaten zur Sicherung an das IAB geben können. Bei möglicher Nachprüfung kann das IAB die Daten unter Auflagen erneut zur Verfügung stellen. Das FDZ der Statistischen Ämter der Länder erwägt, ebenfalls eine Datensicherung für Forscher bereitzustellen.

3.2 Datenverwaltung

In diesem Abschnitt wird detaillierter auf die Anforderungen hinsichtlich der Datenverwaltung in der virtuellen Arbeitsumgebung eingegangen. Dabei werden die verschiedenen Datentypen betrachtet und benötigte Funktionalitäten beschrieben.

3.2.1 Forschungsdatenverwaltung

Die Forschungsdaten der sozioökonomischen Berichterstattung liegen oft in programmspezifischen Formaten, z.B. im SPSS-Format, vor. Diese proprietären Formate enthalten meist nur wenige oder gar keine für die virtuelle Arbeitsumgebung direkt nutzbaren Metadaten. Forschungsdaten können allerdings auch zur besseren kollaborativen Arbeit in andere Formate umgewandelt werden. Bei derartigen Umwandlungen gehen die Metadaten jedoch meist verloren. Dies ist insbesondere der Fall, wenn eine Umwandlung in ein Format vorgenommen wird, welches keine Metadatenhaltung einschließt, wie z.B. CSV-Dateien (Comma Separated Value).

Ein Forschungsdatensatz besteht häufig aus nur einer einzigen Datei, welche je nach Format und Informationsgehalt zwischen wenigen MB und mehreren GB groß sein kann. In der *soeb* sind die folgenden Formate die gebräuchlichsten:

Kürzel	**Beschreibung**
*.sav, *.por	Format von SPSS, kann sich je nach Programmversion unterscheiden, (proprietär)
*.dta	Format von Stata, kann sich je nach Programmversion unterscheiden, (proprietär)

Kürzel	Beschreibung
*.sas	Format von SAS, kann sich je nach Programmversion unterscheiden, (proprietär)
*.RData	Format von R, kann sich je nach Programmversion unterscheiden
*.xls, *.xlsx	Format von Microsoft Excel, kann sich je nach Programmversion unterscheiden, enthält grundlegende Metadaten, wie Autor, Titel, usw., Metadaten durch Key-Value-Paare erweiterbar, (proprietär)
*.csv, *txt	Standardformat, enthält keine Metadaten (in kodierter Form)
*.jpg, *.jpeg	Standardformat, enthält Metadaten im EXIF-Format
*.pdf	Standardformat, enthält grundlegende Metadaten, (proprietär)
*.ps	Standardformat, enthält (fast) keine Metadaten
*.doc, *.docx	Format von Microsoft Word, kann sich je nach Programmversion unterscheiden, enthält grundlegende Metadaten, wie Autor, Titel, usw., Metadaten durch Key-Value-Paare erweiterbar, (proprietär)

Es ist zu erwarten, dass in Zukunft neue Versionen oder andere, neue Datenformate Verwendungen finden werden. Aufgrund der daraus insgesamt resultierenden Vielfalt an möglichen Datenformaten ergibt sich eine erste Anforderung an die Ablage von Forschungsdaten:

Anforderung 5. Die Ablage von Forschungsdaten muss unabhängig von Daten- und Dateiformaten möglich sein.

Die verschiedenen Formate für Forschungsdaten unterstützen das Abspeichern von Metadaten auf unterschiedlichste Weise. Einige Formate, wie z.B. CSV, erlauben gar keine Ablage von Metadaten. Metadaten für Forschungsdatensätze werden jedoch benötigt, um zumindest grundlegende Informationen über deren Herkunft ablegen zu können. Daraus ergibt sich folgende Anforderung:

Anforderung 6. Die Ablage von Forschungsdaten muss das Abspeichern von Metadaten erlauben.

In der sozialwissenschaftlichen Forschung gibt es derzeit nur wenige allgemeingültige Standards für Metadatenstrukturen von Forschungsdaten. Als Beispiel sei hier die Data Documentation Initiative[5] genannt (DDI). Einen solchen Standard zu definieren, erfordert langjährige Erfahrung in dem entsprechenden Umfeld. Und selbst dann kann es vorkommen, dass ein Standard im Nachhinein erweitert

[5] DDI will durch strukturierte Metadaten für sozialwissenschaftliche Datensätze einen Vergleich von Datensätzen aufgrund von Standardisierung ermöglichen. DDI gibt es in unterschiedlichen Versionen: DDI 2.0 dient der grundsätzlichen Beschreibung von Datensätzen, DDI 3.0 bezieht außerdem den gesamten Entstehungsprozess von Datensätzen, z.B. die Operationalisierung von Variablen, mit ein. Weitere Informationen unter: Data Documentation Initiative (DDI) (2009), letzter Zugriff: 2010.07.01, URL: http://www.ddialliance.org.

werden muss, damit neue Anforderungen mit abgedeckt werden können. Die Metadatenstrukturen für Forschungsdaten, die von der virtuellen Arbeitsumgebung benötigt werden, sollten vor allem die Ablage von Metadaten nach von der sozioökonomischen Berichterstattung vorgegebenen Schemata in einer erweiterbaren Form erlauben. Es ergibt sich demnach folgende Anforderung:

Anforderung 7. ***Die Struktur der Metadaten für Forschungsdaten soll erweiter- und veränderbar sein und der virtuellen Arbeitsumgebung somit die größtmögliche Flexibilität auch für neue Anforderungen und damit verbundene Erweiterungen sowie die Anwendung von Standards liefern.***

Forschungsdatenablage im Grid

Die Expertise evaluiert den Einsatz von Grid-Technologie und -Infrastruktur für die nachnutzbare Ablage von Forschungsdaten. Es wird daher in diesem Abschnitt analysiert, welche funktionalen Anforderungen an die Grid-Speichersysteme hinsichtlich der Ablage von Forschungsdaten gestellt werden.

Für die Ablage von Forschungsdaten und deren Metadaten sind zunächst jene grundlegenden Operationen wichtig, die in den meisten Systemen für die Ablage von Daten unterstützt werden. Zu diesen gehören

- das Ablegen eines Datensatzes,
- das Lesen eines Datensatzes,
- das Verändern einen Datensatzes und
- das Löschen eines Datensatzes,

was häufig unter der Abkürzung CRUD (Create Read Update Delete) zusammengefasst wird. Daraus lässt sich die folgende Anforderung an die Ablage von Forschungsdaten ableiten:

Anforderung 8. ***Für die Ablage von Forschungsdatensätzen und deren Metadaten werden die Funktionen Speichern, Lesen, Verändern und Löschen benötigt.***

Der Umgang mit Forschungsdaten erfordert meist eine gewisse Strukturierung der Ablage. Sicherlich können wenige Dateien noch übersichtlich in einem Verzeichnis auf der Festplatte verwaltet werden. Wenn jedoch viele Dateien zu unterschiedlichen Forschungsprozessen und -projekten abgelegt werden sollen, so sind meist komplexere Ablagestrukturen notwendig. Am gebräuchlichsten ist hier die Verwendung von tief verschachtelten Verzeichnisstrukturen. Diese folgen jedoch selten einem Schema oder das angewendete Schema ist nur dem Wissenschaftler bekannt, der die Dateien abgelegt hat. Und selbst dieser kann sich unter Umständen nach längerer Zeit nicht mehr an das Schema erinnern.

Das, was mit solchen Verzeichnisstrukturen erreicht wird, ist eine reine Gruppierung von Datensätzen. Die Bezeichnung eines Verzeichnisses sowie seine konkrete Position im Verzeichnisbaum bilden zusammen die Semantik einer Kategorie von Dateien, die alle in dem gleichen Verzeichnis liegen, also alle dieser Kategorie angehören. Es könnte z.B. folgendes Verzeichnis an irgendeiner Stelle auf der Festplatte existieren: VirtAug/SOFI/inputFiles/SPSS. Im ersten Schritt würde man vermuten, dass alle Dateien in diesem Verzeichnis zueinander gehören. Mit dem Wissen, dass VirtAug ein Projekt bezeichnet, dass das SOFI ein Partner in diesem Projekt ist, und dass SPSS ein Statistik-Programm ist, lässt sich erahnen, dass diese Dateien alle Ausgangsdaten des Projektpartners SOFI im Projekt Vir-

tAug sind und im SPSS-spezifischen Format vorliegen. Dies ist gleichzeitig die semantische Bedeutung der dadurch spezifizierten Kategorie.

Allerdings gibt es auch Kategorien, die auf globaler Ebene angesiedelt sind. Das Verzeichnis VirtAug/SOFI im obigen Beispiel ist eine Kategorie, zu der alle Dateien gehören, die dem Projektpartner SOFI im Projekt VirtAug zuzuordnen sind. Es handelt sich dabei um eine Elternkategorie der zuvor sehr konkreten Kategorie der SPSS-basierten Ausgangsdaten des Projektpartners SOFI im Projekt VirtAug. Dies zeigt, dass die Kategorien hierarchisch aufgebaut sind. Ein Datensatz einer bestimmten Kategorie gehört automatisch zur entsprechenden Elternkategorie, sofern diese existiert.

Eine derartige Kategorisierung von Forschungsdaten sollte bei deren Ablage möglich sein. Die virtuelle Forschungsumgebung kann diese Kategorien dann zur strukturierten Anzeige der Forschungsdaten, z.B. in Form einer Verzeichnisstruktur anzeigen.

Anforderung 9. ***Die Ablage von Forschungsdaten muss eine hierarchische Kategorisierung der Dateien erlauben.***

Solche Kategorisierungen werden meist mit normalen Dateisystemen realisiert. Damit unterliegen sie jedoch auch deren Einschränkungen. So sind z.B. die Metadaten, die für ein Verzeichnis angegeben werden können, relativ stark eingegrenzt. Die Möglichkeiten Metadaten Verzeichnissen zuzuordnen hängen zudem stark von dem verwendeten Dateisystem zur logischen Speicherung ab und sind wenig weit verbreitet. Die Kategorien der Forschungsdaten sollten derartige Beschränkungen auflösen und erweiterbare Metadatenstrukturen für deren Beschreibung unterstützen.

Anforderung 10. ***Kategorien von Forschungsdaten müssen, ähnlich wie Verzeichnisse in Dateisystemen, durch verschiedene grundlegende Metadaten beschrieben werden können.***

Anforderung 11. ***Die Metadaten von Kategorien von Forschungsdaten müssen mit frei definierbaren Metadaten ergänzt werden können.***

Eine weitere Begrenzung von Dateisystemen ist, dass eine hierarchische Kategorisierung von Dateien nur einmal ausgeführt werden kann. So ist es nur begrenzt möglich eine Datei in verschiedene Verzeichnisbäume gleichzeitig einzusortieren (z.B. über Links oder Verknüpfungen). Je nach Dateisystem geht dabei allerdings meist die Übersicht gänzlich verloren. Bei den Hardlinks unter Linux ist es z.B. kaum noch möglich abzuschätzen, welche Auswirkungen die Änderung einer Datei hat, die unter Umständen noch an eine anderen Stelle des Dateisystems verlinkt ist.

In der Praxis zeigt sich jedoch oft, dass eine mehrfache Kategorisierung von Dateien sinnvoll und hilfreich wäre. Im obigen Beispiel möchte man vielleicht alle Statistikdaten aus dem gesamten Projekt ohne die Zuordnung zu dem jeweiligen Projektpartner sowie ohne die Formatspezifikation und die Herkunft der Dateien angezeigt bekommen, die Mikrodaten enthalten. Daher wäre es sinnvoll, eine Kategorienhierarchie einzuführen, die nur den Projektnamen und den Dateninhalt charakterisiert. In der Praxis sollte dann zwischen den verschiedenen Kategorisierungen umgeschaltet werden können. Daraus ergibt sich eine weitere Anforderung an die Kategorisierung von Forschungsdaten.

Anforderung 12. ***Es müssen parallel mehrere hierarchische Kategorisierungen von Forschungsdaten möglich sein, wobei ein Forschungsdatensatz in jeder Kategorienhierarchie nur einmal eingeordnet werden darf.***

Die Möglichkeit mit mehreren Kategorienhierarchien zu arbeiten, kann bei falscher Anwendung den Umgang mit den Daten erschweren. Vor allem wenn verschiedene Anwender viele unterschiedliche Kategorienhierarchien einführen, die sich gegebenenfalls nur leicht unterscheiden, werden die entstehenden Sortierungen immer unübersichtlicher. Die Projekte, welche die virtuelle Arbeitsumgebung einsetzen, sollten an dieser Stelle bei Projektbeginn festlegen und spezifizieren, welche Kategorienhierarchien eingeführt werden und wie diese anzuwenden sind.

Mehrere Kategorienhierarchien bedeuten auch mehrfache Pflege. Oft ist allerdings schon die Pflege einer einzelnen Kategorienhierarchie sehr aufwendig. Es ist daher Aufgabe der graphischen Oberfläche der virtuellen Arbeitsumgebung, den Anwender bei der Nutzung mehrerer Kategorienhierarchien zu unterstützen. Ein Beispiel für eine Umsetzung wird in Kapitel 5.2.3 angeführt. Eine mögliche Lösung wäre außerdem, Kategorien mit Einsortierungsregeln zu versehen. Damit kann die Programmlogik, der virtuellen Arbeitsumgebung anhand der Forschungsdaten erkennen, wo diese in einer Kategorienhierarchie eingeordnet werden müssen. Eine weitere Unterstützung könnten Sichten auf zwei Kategorienhierarchien bilden, mit denen Dateien in der einen Hierarchie durch drag and drop in die jeweils andere Hierarchie einsortiert werden können.

Nachnutzbare Forschungsdatenablage im Grid

Der Begriff „nachnutzbar“ kann und muss bei der Ablage von Forschungsdaten sehr unterschiedlich verstanden werden. Für eine gute wissenschaftliche Praxis ist es notwendig, dass die Forschungsdaten, da sie Grundlage und Nachweis für eine Forschung sind, bis auf unbestimmte Zeit zitier- und lesbar aufbewahrt werden[6]. Die von einem FDZ festgelegten Nutzungsrechte für Daten können hingegen einer derart langen Aufbewahrung entgegenstehen. Damit wird die eigentlich implizit unbegrenzte Nachnutzung von Daten auf eine unter Umständen begrenzte Nachnutzung reduziert. Dadurch ergeben sich weitere Anforderungen an eine Ablage von Forschungsdaten. Zum einen sollten Möglichkeiten existieren, für einen Datensatz Informationen hinsichtlich bestimmter Nutzungsrechte abzulegen und diese auch aufrufen und bearbeiten zu können. Zum anderen sollte das System auf Basis dieser Informationen eine automatische Unterstützung für die Beachtung von Nutzungsrechten implementieren. Daraus ergeben sich folgende Anforderungen an die Forschungsdatenablage:

Anforderung 13. ***Die Metadaten für Forschungsdaten müssen Informationen zu Nutzungsrechten, speziell zum Ablauf von Nutzungszeiträumen und den am Ende eines Nutzungszeitraums auszuführenden Aktionen (z.B. Löschen des Datensatzes) beinhalten können.***

Anforderung 14. ***Die virtuelle Arbeitsumgebung muss automatisch feststellen können, für welchen Datensatz Nutzungsfristen abgelaufen sind, und dem Benutzer entsprechende Aktionen, z.B. das Löschen eines Arbeitsdatensatzes durch den Eigentümer, vorschlagen können bzw. Zugriffe auf diese Datensätze verhindern.***

[6] Die Empfehlungen der DFG nennen hier z.B. 10 Jahre als minimale Zeitspanne. Vgl. Empfehlung 7 in Deutsche Forschungsgemeinschaft (1998): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, WILEY-VCH, Weinheim. URL: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

Die Nutzungsrechte können jedoch auch verlangen, dass nach dem Ablauf eines Nutzungszeitraums nicht nur der Forschungsdatensatz selbst, sondern auch alle aus ihm abgeleiteten Datensätze gelöscht werden müssen. Dies stellt den Wissenschaftler vor eine besondere Herausforderung, da in einem solchen Fall unter Umständen auch Datensätze gelöscht werden müssen, die erst durch mehrere aufeinanderfolgende Verarbeitungszyklen entstanden sind. Die Arbeitsumgebung muss an dieser Stelle den Wissenschaftler bei der Ermittlung der betroffenen Datensätze unterstützen. Daraus ergeben sich zunächst weitere Anforderungen an die Forschungsdatenablage:

Anforderung 15. ***Die Metadaten für Forschungsdaten müssen Informationen zu deren Ursprung und Entstehen, der so genannten Provenienz, beinhalten können.***

Anforderung 16. ***Die Metadaten für Forschungsdaten müssen logische Verknüpfungen zwischen Datensätzen, zum Beispiel zur Abbildung von Provenienz, beinhalten können.***

Auf Basis der so abgelegten Informationen kann das System für den Benutzer ermitteln, welche Datensätze betroffen sind, in welcher Relation sie zu einander stehen und gegebenenfalls einen Vorschlag für deren Löschung unterbreiten. Auch für den Fall, dass ein Benutzer einen Datensatz manuell löschen möchte, kann auf Basis dieser Informationen festgestellt werden, welche anderen Datensätze möglicherweise auf diesem Datensatz beruhen. Der Nutzer kann darüber informiert werden und löscht somit nicht unbeabsichtigt wichtige Datensätze, die zum Nachweis bestimmter Forschungsergebnisse notwendig sind.

Wie bereits erwähnt, müssen im Rahmen der guten wissenschaftlichen Praxis unter anderem alle Forschungsdaten, die zu einem Forschungsergebnis geführt haben, aufbewahrt werden. Dies ist nicht nur speicheraufwändig, sondern, wie bereits erwähnt, aufgrund der Nutzungsrechte nicht immer möglich. Dennoch sollte die virtuelle Forschungsumgebung eine größtmögliche Dokumentation erlauben und alle in einem Forschungsprozess angewendeten Forschungsdaten, Syntaxdateien, sowie die Entstehungsreihenfolge der Forschungsdaten inklusive der Reihenfolge der Syntaxdateianwendungen anzeigen können. Dies wäre mit den bisherigen Anforderungen hinsichtlich der Dokumentation der Provenienz in den Metadaten der Forschungsdaten möglich. Wenn die Forschungsdaten aufgrund der Nutzungsrechte jedoch teilweise gelöscht werden müssen, so würden grundlegende Elemente der Provenienz verloren gehen. Dies könnte mit Hilfe von Metadaten zu „virtuellen Forschungsdaten“ gelöst werden. Diese Metadaten würden genauso aufgebaut sein, wie Metadaten zu existierenden Datensätzen, würden jedoch anstelle realer Daten erzeugbare Daten beschreiben. Dies soll an folgendem Beispiel kurz erläutert werden:

Auf Basis eines Scientific-Use-Files, welches in der virtuellen Forschungsumgebung abgelegt wurde, sind durch die Anwendung von drei Syntaxdateien drei neue Forschungsdatensätze entstanden. Die virtuelle Forschungsumgebung hat diesen Prozess in den Metadaten der Forschungsdaten dokumentiert. Aufgrund der Nutzungsrechte müssen nun das Scientific-Use-File sowie der erste erzeugte Arbeitsdatensatz gelöscht werden. In diesem Fall würden die reinen Daten beider Datensätze gelöscht werden, die Metadaten jedoch erhalten bleiben. In einem gesonderten Speicherbereich der Metadaten wird hinterlegt, dass diese Datensätze virtuell sind. Der Metadatensatz des Scientific-Use-Files sollte genügend Metadaten enthalten, damit das Scientific-Use-File gegebenenfalls erneut vom Datenanbie-

ter angefordert oder anderweitig beschafft werden kann[7]. Im Falle des notwendigen Nachweises der Forschung könnte dadurch mit Hilfe der Informationen der Metadaten aller Datensätze sowie der Syntaxdateien der gesamte Forschungsprozess nachgestellt und gegebenenfalls wiederholt werden.

Daher ergibt sich eine weitere Anforderung an die Ablage der Forschungsdaten:

Anforderung 17. ***Die Ablage von Forschungsdaten muss das Abspeichern von virtuellen Datensätzen erlauben. Solche Datensätze haben keine tatsächlichen Daten. Diese können jedoch ggf. mit Hilfe der bekannten Metadaten neu erzeugt oder beschafft werden.***

Unabhängig von Nutzungszeiträumen gilt es bei der nachnutzbaren Forschungsdatenablage jedoch auch Kriterien zu beachten, die vor Datenverlust schützen. Dies muss sowohl bei den aufgrund von Nutzungsrechten kurzzeitig abgelegten Datensätzen, als auch bei den langfristig aufzubewahrenden Datensätzen beachtet werden. Außerdem gilt es die Daten zum einen vor unerlaubter Veränderung zu schützen, bzw. unerlaubte oder unerwünschte Datenänderungen zu erkennen. Das Sicherstellen der Unversehrtheit von Daten, der so genannten Datenintegrität, kann über Prüfsummen realisiert werden. Die Nachvollziehbarkeit der Urheberschaft der Daten, der so genannten Datenauthentizität, wird durch Datensignaturen, meist auf den Prüfsummen der Daten sichergestellt. Hieraus ergibt sich eine weitere Anforderung an die Forschungsdatenablage:

Anforderung 18. ***Die Metadaten für Forschungsdaten müssen Möglichkeiten zur Ablage von datenbezogenen Prüfsummen und Signaturen bieten.***

Eine der wichtigsten Anforderungen ergibt sich aus der mehrfachen Erzeugung von Forschungsdaten. So kann es z.B. sein, dass ein Forschungsdatenzentrum ein Scientific-Use-File geliefert hat, auf dessen Basis bereits ein Forschungsprozess erfolgt ist, das Forschungsdatenzentrum dann jedoch eine neue, z.B. korrigierte, Version des Scientific-Use-Files zur Verfügung stellt. In einem solchen Fall möchte man den Forschungsprozess mit den neuen Daten wiederholen, die alten Daten jedoch nicht verlieren. Gleiches gilt, wenn eine in einem Forschungsprozess angewendete Syntaxdatei korrigiert wird, und alle Forschungsdaten, die nach der ursprünglichen Anwendung dieser Syntaxdatei erzeugt wurden, noch mal neu berechnet werden sollen. Auch hier sollen die alten Versionen der Forschungsdaten gegebenenfalls erhalten bleiben. Die erste sich daraus ergebende Anforderung ist:

Anforderung 19. ***Die Ablage von Forschungsdaten muss verschiedene Versionen eines Datensatzes speichern können.***

Damit diese Versionen jedoch effektiv genutzt werden können, werden noch weitere Funktionalitäten benötigt:

Anforderung 20. ***Die Ablage von Forschungsdaten muss unterschiedliche Versionen eines Forschungsdatensatzes benennen und auflisten können.***

[7] Möglicherweise können auf dieser Basis auch Formen der direkt referenzierbaren Speicherung solcher Daten beim FDZ vereinbart werden.

Anforderung 21. ***Zu jeder existierenden Version eines Forschungsdatensatzes sollten grundsätzliche Metadaten, wie eine Versionsnummer, ein Versionsname und ein Erstellungsdatum existieren.***

Anforderung 22. ***Zu jeder Version eines Datensatzes müssen die passenden Metadaten abgelegt sein.***

Anforderung 23. ***Die Ablage von Forschungsdaten muss eine definierte Version eines Datensatzes inklusive der dazu gehörenden Version der Metadaten zur Verfügung stellen können***

Die Nutzung dieser Funktionalität muss jedoch in der virtuellen Arbeitsumgebung implementiert sein. Es ist z.B. nicht Aufgabe der Datenhaltung, die Liste von existierenden Versionen eines Datensatzes nutzer- und kontextspezifisch aufzubereiten. Außerdem soll die virtuelle Arbeitsumgebung dem Anwender die Möglichkeit geben, die Versionsbezeichnung selbst festzulegen.

Sichere Arbeitsdatensatzablage

Das Abspeichern von Forschungsdatensätzen in normalen Dateisystemen erlaubt die Angabe grundlegender Zugriffsrechte, die der Benutzer auch für sich selbst festlegen kann. Diese werden oft so verwendet, dass der Benutzer nicht nur Fremdzugriffe beeinflusst, sondern auch seine eigenen Handlungen kontrollieren kann. Ein Beispiel ist das Setzen eines Schreibschutzes auf eine Datei, um eigene unerwünschte Änderungen zu verhindern. Allgemeiner gesagt handelt es sich dabei um so genannte Flags mit bestimmten Bedeutungen, die vom Benutzer festgelegt werden können, und welche von den Programmen, mit denen der Benutzer arbeitet, ausgewertet werden, um das Handeln des Benutzers nach seinen Vorstellungen zu sichern. Derartige Mechanismen sollten dem Benutzer auch in der virtuellen Arbeitsumgebung zu Verfügung stehen, woraus sich die folgende Anforderung ableitet:

Anforderung 24. ***In den Metadaten von Forschungsdaten müssen Flags hinterlegbar sein, die mit Hilfe der virtuellen Arbeitsumgebung vom Benutzer gesetzt werden können, und die von der virtuellen Arbeitsumgebung während der Arbeit des Benutzers interpretiert und angewendet werden.***

Die vorige Anforderung betrachtet bewusst noch nicht die Zugriffsrechte anderer Benutzer auf Forschungsdaten. Hierfür gibt es unterschiedliche Systeme, die entsprechend den Anforderungen ausgewählt werden sollten. Im Fall der sozioökonomischen Berichterstattung, die einer Nutzergruppe in einer Größenordnung im zwei- bis dreistelligem Bereich entspricht, ist das Rechtesystem von gängigen Betriebssystemen vollkommen ausreichend. Bei anderen Anforderungen stehen komplexere Systeme zur Verfügung.[8]

Die Ablage und Verwaltung derart komplexer Rechtebedingungen in den Metadaten der Forschungsdaten würde eine Rechteverwaltung erheblich erschweren und teilweise sehr unübersichtlich werden

[8] Beispiel wäre ein so genanntes RBAC-System (role based access control, rollenbasierte Zugriffskontrolle). Derartige Systeme verwalten Benutzer und Rollen. Einem Benutzer werden eine oder mehrere Rollen zugewiesen. Den Rollen wiederum werden Nutzungsrechte jeglicher Art zugeordnet. Für jede Aktion eines Benutzers wird dann mit Hilfe des Systems verifiziert, welche Rollen der Benutzer hat und ob ihm eine dieser Rollen das Ausführen der entsprechenden Aktion erlaubt.

lassen. Dennoch sollte ein Besitzer eines Datensatzes festlegen können, welchen Benutzern bzw. welchen Rollen er welchen Zugriff auf seine Daten gewährt. Um dies zu erlauben, muss die virtuelle Arbeitsumgebung das Rechteverwaltungssystem direkt ansprechen können. Zum einen muss hierbei die Bearbeitung von Rechten möglich sein, zum anderen muss die virtuelle Arbeitsumgebung die Rechtekonfigurationen auswerten und unumgänglich umsetzen. Der Zugriff auf die Rechteverwaltung sollte für den Benutzer transparent erfolgen.

Da die Rechteverwaltungsinformationen nicht in den Metadaten der Forschungsdaten abgelegt werden, ergeben sich die folgenden Anforderungen:

Anforderung 25. ***Die Ablage von Forschungsdaten sowie deren Metadaten muss die in einer Rechteverwaltung konfigurierten Zugriffsrechte beachten und umsetzen.***

Anforderung 26. ***Die Ablage von Forschungsdaten muss sicherstellen, dass unerlaubte Zugriffe jeglicher Art auf die abgelegten Daten verhindert werden.***

3.2.2 Syntaxdateiverwaltung

Bei der Ablage von Syntaxdateien müssen ähnliche Kriterien wie bei der Ablage von Forschungsdaten Beachtung finden. Aus diesem Grund können viele der Anforderungen an die Ablage der Forschungsdaten für Syntaxdateien umformuliert übernommen werden. Sie werden daher hier noch einmal aufgegriffen, jedoch nicht so ausführlich eingeleitet, wie im vorherigen Kapitel.

Auch wenn die in der sozioökonomischen Berichterstattung weithin eingesetzten Statistik-Programme die Syntax in reinen Textdateien abspeichern, so sollte die Ablage von Syntaxdateien mit dem Blick auf zukünftige Einsätze von den konkreten Dateitypen unabhängig sein:

Anforderung 27. ***Die Ablage von Syntaxdateien muss unabhängig von Daten- und Dateiformaten möglich sein.***

Auch Syntaxdateien müssen mit Metadaten beschreibbar sein. Dies sollte flexibel auf zukünftige Anforderungen erweitert werden können.

Anforderung 28. ***Die Ablage von Syntaxdateien muss das Abspeichern von Metadaten erlauben.***

Anforderung 29. ***Die Struktur der Metadaten für Syntaxdateien soll erweiter- und veränderbar sein und der virtuellen Arbeitsumgebung somit die größtmögliche Flexibilität auch für neue Anforderungen und damit verbundene Erweiterungen liefern.***

Die Ablage von Syntaxdateien muss ähnliche grundlegende Funktionalitäten bereitstellen:

Anforderung 30. ***Für die Ablage von Syntaxdateien und deren Metadaten werden die Funktionen Speichern, Lesen, Verändern und Löschen benötigt.***

Anforderung 31. ***Die Ablage von Syntaxdateien muss eine hierarchische Kategorisierung der Dateien erlauben.***

Anforderung 32. ***Kategorien von Syntaxdateien müssen, ähnlich wie Verzeichnisse in Dateisystemen, durch verschiedene grundlegende Metadaten beschrieben werden können.***

Anforderung 33. ***Die Metadaten von Kategorien von Syntaxdateien müssen mit frei definierbaren Metadaten ergänzt werden können.***

Anforderung 34. ***Es müssen parallel mehrere hierarchische Kategorisierungen von Syntaxdateien möglich sein, wobei eine Syntaxdatei in jeder Kategorienhierarchie nur einmal eingeordnet werden darf.***

Der Begriff „nachnutzbar“ gilt für Syntaxdateien in bestimmten Aspekten gleichermaßen:

Anforderung 35. ***Die Metadaten für Syntaxdateien müssen Möglichkeiten zur Ablage von datenbezogenen Prüfsummen und Signaturen bieten.***

Hinsichtlich der Versionierung gibt es bei Syntaxdateien zunächst ebenfalls ähnliche Anforderungen:

Anforderung 36. ***Zur Dokumentation von Änderungen bzw. Entwicklungsstufen an Syntaxdateien soll die Ablage eine Versionsverwaltung anbieten.***

Anforderung 37. ***Die Ablage von Syntaxdateien muss unterschiedliche Versionen einer Syntaxdatei benennen und auflisten können.***

Anforderung 38. ***Zu jeder existierenden Version einer Syntaxdatei sollten grundsätzliche Metadaten, wie eine Versionsnummer und ein Erstellungsdatum existieren.***

Anforderung 39. ***Zu jeder Version einer Syntaxdatei müssen die passenden Metadaten abgelegt sein.***

Anforderung 40. ***Die Ablage von Syntaxdateien muss eine definierte Version einer Syntaxdatei inklusive der dazu gehörenden Version der Metadaten zur Verfügung stellen können***

Auch hier bleibt die Nutzung dieser Funktionalitäten der virtuellen Arbeitsumgebung überlassen. Sinnvoll wäre auf jeden Fall mit Hilfe von Vergleichsalgorithmen Unterschiede zwischen zwei Syntaxdateiversionen zu erkennen, und für den Anwender elegant bedienbar aufzubereiten. Beispiele für solche Darstellungen liefern z.B. die in der Softwareentwicklung eingesetzten Versionsverwaltungssysteme. Auch individuelle Möglichkeiten zur Festlegung neuer Versionen sowie deren Benennung sollten vorhanden sein.

Die Ablage von Syntaxdateien unterliegt ebenso den gleichen Kriterien hinsichtlich Sicherheit, auch wenn hierbei weniger der Fokus auf dem Datenschutz, sondern eher auf Nutzerrechten liegt.

Anforderung 41. ***In den Metadaten von Syntaxdateien müssen Flags hinterlegbar sein, die mit Hilfe der virtuellen Arbeitsumgebung vom Benutzer gesetzt werden können, und die von der virtuellen Arbeitsumgebung während der Arbeit des Benutzers ausgewertet und beachtet werden können.***

Anforderung 42. ***Die Ablage von Syntaxdateien sowie deren Metadaten muss die in einer Rechteverwaltung konfigurierten Zugriffsrechte beachten und umsetzen.***

Anforderung 43. ***Die Ablage von Syntaxdateien muss sicherstellen, dass unerlaubte Zugriffe jeglicher Art auf die abgelegten Daten verhindert werden.***

3.2.3 Metadatenextraktion und Dokumentation

In den vorigen Abschnitten wurden bereits mehrfach Anforderungen hinsichtlich der Inhalte der Metadaten für Forschungsdaten und Syntaxdateien formuliert. Es wird nun detaillierter auf die Erzeugung der Metadaten und damit verbundene Anforderungen eingegangen.

Metadaten lassen sich aufteilen in technische und fachliche Metadaten. Technische Metadaten umfassen z.B. das Erzeugungsdatum einer Datei oder eines Datensatzes, das Datenformat, Daten zum Erzeugerobjekt (Kamera, Teleskop, usw.) sowie Daten zur Änderungshistorie (siehe Anforderungen zur Versionierung von Forschungsdaten und Syntaxdateien in den vorigen Abschnitten). Dabei werden Änderungen entweder durch die manuelle Bearbeitung der Daten oder durch technische Prozesse z.B. im Rahmen von Formatkonvertierungen zum Ziel der Langzeitarchivierung vorgenommen. Ebenso zählen zu den technischen Metadaten die formatspezifisch eingebundenen Metadaten, da diese i.d.R. automatisiert extrahiert werden können.

Anforderung 44. ***In der virtuellen Arbeitsumgebung soll eine integrierte Unterstützung für technische Metadaten zur Verfügung stehen.***

Fachliche Metadaten beschreiben den spezifischen Kontext der Erzeugung von Daten und orientieren sich i.d.R. an den Anforderungen der jeweiligen Wissenschaftsdisziplin. Hierzu zählt z.B. die Beschreibung der Datenbasis und der Verfahren für eine Auswertung im Rahmen der sozioökonomischen Berichterstattung, die im Zusammenhang mit den erzeugten Forschungsdaten verwendet wurden. Damit beschreiben diese Metadaten den fachlichen Kontext der verwendeten Daten. Technische Metadaten, die kontextspezifische Dokumentation enthalten bzw. enthalten können, zählen demnach ebenfalls zu den fachlichen Metadaten.

Anforderung 45. ***In der virtuellen Arbeitsumgebung soll eine integrierte Unterstützung für fachliche Metadaten zur Verfügung stehen.***

Extraktion technischer Metadaten

Die technischen Metadaten für Standard-Datenformate lassen sich überwiegend automatisiert extrahieren. Notwendig ist die jeweilige Integration von Tools oder Programmen zum Umgang mit den technischen Metadaten der Datenformate in die virtuelle Forschungsumgebung. Darüber hinaus lassen sich weitere technische Metadaten, z.B. jene, die sich aus den Anforderungen in den Kapiteln 3.2.1 und 3.2.2 ergeben, automatisch generieren und für den Benutzer transparent ablegen und verwalten. Daraus ergeben sich folgende Anforderungen:

Anforderung 46. ***In der virtuellen Arbeitsumgebung soll die Extraktion oder Erzeugung technischer Metadaten generell automatisch erfolgen.***

Die Extraktion von technischen Metadaten ist abhängig vom verwendeten Dateiformat. Zunächst müssen hier die in der sozioökonomischen Berichterstattung verwendeten Formate für Forschungsdaten

unterstützt werden. Zu diesen gehören die proprietären Formate von SPSS und Stata, das freie Format von R sowie das weitverbreitete Format CSV. Die beiden letztgenannten basieren auf ASCII-Text-Kodierung und enthalten eine spezifische Datenstruktur (in R die Syntax, in CSV die Separatoren für die Spaltenwerte).

Um die kollaborative Verwendung von Ergebnisdaten zu unterstützen, ist eine Standardisierung der dafür verwendeten Datenformate notwendig. Entsprechend der Standardisierung der Ergebnisdaten in Richtung von Standard-Datenformaten (PDF, JPEG, XLS/XLSX, ASCII-Text und CSV) wird gleichzeitig die Extraktion technischer Metadaten im Hinblick auf die entsprechenden formatspezifische Eigenschaften benötigt.

Anforderung 47. ***Die Extraktion technischer Metadaten muss für die verschiedenen Formate von Forschungsdaten (SPSS, Stata, R, CSV), Syntaxdateien (SPSS, Stata, R) und Ergebnisdaten (PDF, JPEG, XLS/XLSX, ASCII-Text, CSV) möglich sein.***

Extraktion fachlicher Metadaten

Fachliche Metadaten können i.d.R. nur teilweise automatisch erzeugt werden, da diese durch die Forschungsprozesse entstehen. Zudem ist aufgrund der explorativen Eigenschaft (siehe Abschnitt 2.1) der sozialwissenschaftlichen Forschung eine automatische Erzeugung der fachlichen Metadaten nur schwer möglich.

Die meisten fachlichen Metadaten müssen daher beim Einlagerungsprozess oder der späteren Arbeit mit den Daten in der virtuellen Forschungsumgebung erhoben werden. Hierzu ist es notwendig, Mindestangaben als Metadatenstandard für einzulagernde Daten seitens der sozioökonomischen Berichterstattung zu definieren, diese aber auch erweiterbar zu machen. Die Angaben werden durch den jeweiligen Wissenschaftler eingefügt.

Bei bestimmten Daten können fachliche Metadaten aus den Dateninhalten extrahiert werden, wodurch dem Wissenschaftler die eigene Eingabe der Informationen erspart werden kann. Syntaxdateien enthalten z.B. Befehle, mit denen die zu verarbeitenden Datensätze geladen werden sollen. Die Parameter dieser Befehle verweisen dabei auf die Datei des entsprechenden Datensatzes. Diese Informationen können extrahiert und dem Forscher als mögliches Metadatenelement vorgeschlagen werden. Gleiches gilt für einfach zu extrahierende Informationen auch aus anderen Datenformaten. Beispiele seien hier Informationen über die Struktur von CSV-Dateien, R-Arbeitsbereichen, Header-Informationen bei JPEG und PDF, usw.

Die für die Extraktion und Erstellung von fachlichen Metadaten zu betrachtenden Datenarten sind Forschungsdaten, Ergebnisdaten und Syntaxdateien. Insbesondere für die Syntaxdateien sind die fachlichen Metadaten relevant, da diese Daten für die Nachvollziehbarkeit der Forschung essenziell sind und gleichzeitig keinen Vorschriften zur Löschung durch die FDZ unterliegen. Es ist vorgesehen, dass Syntaxdateien in den Formaten SPSS, Stata und R in der virtuellen Forschungsumgebung gespeichert werden sollen.

Anforderung 48. ***Die virtuelle Arbeitsumgebung soll die Extraktion, Erstellung und Nutzung fachlicher Metadaten unterstützen.***

Anforderung 49. ***Die Extraktion fachlicher Metadaten betrifft Forschungsdaten, Ergebnisdaten und Syntaxdateien. Bei Syntaxdateien und gegebenenfalls bei***

Forschungsdaten sollen die Formate SPSS, Stata und R berücksichtigt werden.

Anforderung 50. ***Die virtuelle Arbeitsumgebung soll das Hinzufügen fachlicher Metadaten unterstützen. Das Hinzufügen ist Aufgabe des jeweiligen Wissenschaftlers, der die zugrunde liegenden Forschungsdaten, Syntaxdateien oder Ergebnisdaten erarbeitet / erzeugt hat. Die virtuelle Arbeitsumgebung soll soweit wie möglich Vorschläge zu Metadatenwerten auf Basis der Dateien, zu denen Metadaten abgelegt werden sollen, unterbreiten, indem unterschiedlichste Informationen aus den Dateien extrahiert werden.***

Dokumentation

Die Syntaxdateien müssen abhängig vom verwendeten Statistik-Programm eine bestimmte Struktur aufweisen und dürfen auch nur die vom jeweiligen Programm unterstützte Befehlskodierung beinhalten. Für das Fern- oder Onsite-Rechnen bei den FDZ müssen die in den Syntaxdateien enthaltenen Kommentare außerdem den Vorgaben der jeweiligen FDZ entsprechen. Zu solchen Vorgaben zählen neben der Kommentarstruktur auch die in die Kommentare einzufügenden Inhalte. Allerdings sind diese Vorgaben überschaubar und beschränken sich auf die meist ohnehin benötigten technischen oder fachlichen Metadaten. Die virtuelle Arbeitsumgebung sollte an dieser Stelle eine Unterstützung zur automatischen Generierung der von den FDZ geforderten Kommentarstrukturen und -elemente auf Basis der extrahierbaren oder vom Benutzer eingegebenen Metadaten bieten.

Anforderung 51. ***Die virtuelle Arbeitsumgebung soll die von den FDZ für das Fern- oder Onsite-Rechnen geforderten Metadatenstrukturen und -inhalte von Syntaxdateien automatisch auf Basis der zu den Syntaxdateien abgelegten Metadaten erzeugen können.***

3.2.4 Forschungsdatensuche

Die Ablage von Forschungsdaten soll nachnutzbar erfolgen. Der Begriff nachnutzbar impliziert auch, dass die Daten zu einem späteren Zeitpunkt wieder auffindbar sein müssen. Dies sollte grundsätzlich anhand der Metadaten, die zu den Datensätzen existieren, möglich sein. Dafür muss die Ablage von Forschungsdaten eine Suchfunktionalität unterstützen, die zumindest anhand von Metadatenwerten Datensätze finden und anzeigen kann. Dementsprechend baut dieser Bestandteil auf der Extraktion der technischen und fachlichen Metadaten auf. Eine Durchsuchbarkeit der Forschungsdaten auf Basis einer Volltextsuche in den eigentlichen Daten ist nicht notwendig, da die Daten häufig nur aufgrund der Metadaten interpretiert werden können. Somit ist die Suche auf den Metadaten ausreichend, um Daten für den jeweiligen wissenschaftlichen Kontext in der virtuellen Forschungsumgebung suchen bzw. wiederauffinden zu können.

Anforderung 52. ***Die Ablage von Forschungsdaten muss eine Suchfunktionalität auf den Metadaten der Forschungsdaten zur Verfügung stellen.***

Sinnvoll wäre außerdem das Auffinden von Datensätzen anhand ungefährer Bestimmung, so dass auch Datensätze gefunden werden, die in ihren Metadaten einen bestimmten Suchbegriff nur indirekt (z.B. als Synonym, oder Mehrzahl statt Einzahl eines Suchbegriffs) beinhalten. Demnach lässt sich folgende Anforderung ableiten.

Anforderung 53. ***Die Suchfunktionalität für Forschungsdaten sollte aktuellen Standards hinsichtlich Bearbeitungszeit und Ergebnisliste (z.B. unscharfe Suche) entsprechen.***

Wie in Abschnitt 3.2.1 bereits erwähnt unterliegen Forschungsdaten bestimmten zu beachtenden Sicherheitsaspekten. So dürfen z.B. nicht alle Benutzer einer virtuellen Arbeitsumgebung alle Forschungsdaten sehen, oder darauf zugreifen. Dies muss bei der Zusammenstellung von Suchergebnissen entsprechend berücksichtigt werden:

Anforderung 54. ***Die Suchfunktionalität für Forschungsdaten muss die in einer Rechteverwaltung konfigurierten Zugriffsrechte beachten und die Ergebnislisten entsprechend filtern.***

Für Forschungsdaten ist vor allem deren Ursprung relevant. Gerade bei Arbeitsdatensätzen sollte angezeigt werden können, durch welche Syntax sie erzeugt wurden. Diese Information ist für die Suche nach Forschungsdaten sowohl bei der Formulierung von Suchbedingungen, als auch bei der Anzeige von Ergebnissen relevant.

Anforderung 55. ***Die Suchfunktionalität für Forschungsdaten sollte sowohl bei der Auswertung von Suchbedingungen, als auch beim Anzeigen der Ergebnisse gegebenenfalls vorhandene Referenzen zu Syntaxdateien oder anderen Forschungsdaten berücksichtigen und anzeigen.***

3.2.5 Syntaxdateisuche

Zusätzlich zu der reinen Suche auf Basis der Metadaten bei den Forschungsdaten ist für die Syntaxdateien eine Volltextsuche sinnvoll, da hierbei z.B. Kommentare, die nicht in den Metadaten enthalten sind und trotzdem zur Beschreibung der verwendeten Algorithmen dienen, ebenfalls in die Suche einbezogen werden können. Auf Grundlage der Volltextsuche wird eine kollaborative Nachnutzung der Syntaxdateien unterstützt, da die Durchsuchbarkeit durch alle Wissenschaftler ermöglicht wird. Es ergibt sich darüber hinaus das Potenzial, dass Teile einer Syntax gezielt gesucht, gefunden und anschließend weiterverwendet werden können.

Mit einer Volltextsuche wird die Durchsuchbarkeit gespeicherter Daten nach jeglichen Inhalten bzw. Inhaltsmustern ermöglicht. Ohne eine Volltextsuche besteht bei großen Datenmengen das Problem, dass Informationen gegebenenfalls nicht wiederaufgefunden oder in einem anderen Kontext nachgenutzt werden können, weil die Metadaten z.B. nicht aussagekräftig genug sind. Mit einer Volltextsuche kann in einer gemeinsam genutzten virtuellen Forschungsumgebung der einzelne Wissenschaftler bei dem Auffinden bzw. Wiederauffinden von Informationen unterstützt werden. Eine Volltextsuche stellt somit eine sinnvolle Ergänzung zu einer normalen Suche auf Metadaten dar.

Eine Volltextsuche basiert auf der Indizierung aller gespeicherten Inhalte einer Datenbasis. Dementsprechend entsteht, in Abhängigkeit von dem Umfang der Datenbasis, ein hoher Rechenaufwand, um den notwendigen Index zu erstellen. In einer virtuellen Forschungsumgebung sollen dafür extern verfügbare Rechenressourcen verwendet werden. Außerdem muss der Index für eine kollaborative Verwendung zentral bereitgestellt werden. Eine virtuelle Forschungsumgebung soll dies entsprechend ermöglichen.

Anforderung 56. ***Die Suche nach Syntaxdateien soll anhand ihrer Metadaten erfolgen können.***

Anforderung 57. ***Nach Möglichkeit soll ebenso eine Volltextsuche für Syntaxdateien in die virtuelle Arbeitsumgebung integriert werden können. Diese soll, wenn möglich, auf bereits vorausgewählten Dateien, die gegebenenfalls mit Hilfe einer Suche auf den Metadaten gefunden wurden, ausgeführt werden können.***

Anforderung 58. ***Die Nutzung der Volltextsuche soll kollaborativ möglich sein.***

Auch für Syntaxdateien gibt es Sicherheitsaspekte, die zu beachten sind. Allerdings basieren diese weniger auf Datenschutzvorschriften sondern vielmehr auf der Wahrung geistigen Eigentums. Ein Forscher mag seine Arbeiten veröffentlichen und dann auch frei (z.B. Open Access) zur Verfügung stellen. Während des Forschungsprozesses soll jedoch genauso die nicht-öffentliche Arbeit erlaubt sein, damit er seiner Forschung plagiatfrei nachgehen kann. Daher dürfen in einer virtuellen Arbeitsumgebung auch nicht alle Benutzer alle Syntaxdateien sehen oder darauf zugreifen. Dies muss bei der Zusammenstellung von Suchergebnissen entsprechend berücksichtigt werden:

Anforderung 59. ***Die Suchfunktionalität für Syntaxdateien muss die in einer Rechteverwaltung konfigurierten Zugriffsrechte beachten und die Ergebnislisten entsprechend filtern.***

Ähnlich wie bei der Suche nach Forschungsdaten ist auch für Syntaxdaten die Beachtung von Referenzen wichtig. So sollten Suchbedingungen für Referenzen formuliert werden können und die Referenzen zu Forschungsdaten, die mit den Syntaxdateien in Beziehung stehen, in der Ergebnisliste angezeigt werden.

Anforderung 60. ***Die Suchfunktionalität für Syntaxdateien sollte sowohl bei der Auswertung von Suchbedingungen, als auch beim Anzeigen der Ergebnisse gegebenenfalls vorhandene Referenzen zu Syntaxdateien oder anderen Forschungsdaten berücksichtigen und anzeigen.***

3.2.6 Konvertierung und Validierung

In der sozioökonomischen Berichterstattung wird, wie bereits in den vorigen Kapiteln mehrfach erwähnt, mit unterschiedlichen Datenformaten gearbeitet. Dies ist nicht nur dem Einsatz unterschiedlicher Statistik-Programme geschuldet. Es ist oft auch notwendig, bestimmte Daten in Dateiformate zu konvertieren, die leichter austausch- oder veröffentlichbar sind. Soweit möglich, soll eine virtuelle Arbeitsumgebung entsprechende Konvertierungstools anbieten oder deren einfache Integration erlauben.

Anforderung 61. ***Die virtuelle Arbeitsumgebung soll das Einbinden von Tools zur Datenkonvertierung erlauben.***

Anforderung 62. ***Wenn möglich sollen existierende Tools zur Konvertierung verwendeter Datenformate integriert werden.***

Die Konvertierung von Forschungsdaten in die verschiedenen Formate gehört, sofern die Datenformate bekannt sind, noch zu den leichter umsetzbaren Anforderungen. Teilweise existieren hier bereits entsprechende Tools. Eine größere Herausforderung stellt allerdings die Konvertierung von Syntaxdateien dar. Hier sind komplexe Compiler notwendig, welche die Erhaltung der Semantik einer Ur-

sprungssyntaxdatei bei einer Konvertierung in ein anderes Format sicherstellen müssen. Hinzu kommt dabei, dass bestimmte Ausdrücke in der einen Syntaxsprache nicht zwingend ein semantisches Äquivalent in der jeweils anderen Syntaxsprache haben, was einen Übersetzungsprozess ebenfalls extrem erschweren würde. Dies ist auch der Grund, warum bisher kaum Tools für derartige Umwandlungen existieren. Im Rahmen der Entwicklung einer virtuellen Arbeitsumgebung für die sozioökonomische Berichterstattung sollten derartige Implementierungen ebenso nicht mit vorgesehen werden, da der Entwicklungsaufwand unter Umständen sehr groß ist. Aus diesem Grund wird eine entsprechende Anforderung an dieser Stelle nicht spezifiziert.

Ein weiterer wichtiger Punkt beim Umgang mit verschiedenen Datenformaten ist deren Validierung. Hierbei soll sichergestellt werden, dass eine Datei inhaltlich dem Datenformat entsprechend strukturiert ist. Je nach Datenformat ist derartige Funktionalität einfacher oder komplexer zu implementieren. Daher existieren nicht für alle in der sozioökonomischen Berichterstattung verwendeten Datenformate passende Validierungstools. Dennoch sollte die virtuelle Arbeitsumgebung das Einbinden derartiger Funktionalitäten erlauben und gegebenenfalls existierende Tools integrieren.

Anforderung 63. ***Die virtuelle Arbeitsumgebung soll das Einbinden von Tools zur Datenvalidierung erlauben.***

Anforderung 64. ***Wenn möglich sollen existierende Tools zur Validierung verwendeter Datenformate integriert werden.***

3.2.7 Datensicherheit und Datenschutz

Bezogen auf die automatisierte Verarbeitung und Speicherung personenbezogener Daten[9] benennt das Bundesdatenschutzgesetz die folgenden zu beachtenden Aspekte[10]:

- Zutrittskontrolle: Unbefugte dürfen keinen Zutritt zu Datenverarbeitungsanlagen bekommen.
- Zugangskontrolle: Unbefugte dürfen die Datenverarbeitungsanlagen nicht benutzen können.
- Zugriffskontrolle: Unbefugte dürfen Daten weder lesen, kopieren, verändern oder löschen können.
- Weitergabekontrolle: Unbefugte dürfen Daten während des Transports bzw. der Datenübermittlung[11] weder lesen, kopieren, verändern oder löschen können.
- Eingabekontrolle: Veränderungen an Daten müssen protokolliert werden.
- Auftragskontrolle: Die Datenverarbeitung bei Dritten muss dem Auftrag entsprechend durchgeführt werden.

[9] Das Bundesdatenschutzgesetz (BDSG) definiert personenbezogene Daten unter §3, Abs. 1.

[10] Frei formuliert nach dem Bundesdatenschutzgesetz Anlage zu §9 Satz 1 (Seite 40). Verfügbar unter: Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), das zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert worden ist (2009), Letzter Zugriff: 2010.06.11, URL: http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf.

[11] Dazu zählt z.B. die Datenübermittlung innerhalb eines geographisch verteilten Unternehmens über öffentliche Datenleitungen.

- Verfügbarkeitskontrolle: Zufällige Veränderungen oder Zerstörungen von Daten sind zu vermeiden.
- Datentrennung: für unterschiedliche Zwecke erhobene Daten müssen getrennt voneinander verarbeitet werden können.

Die sozioökonomische Berichterstattung arbeitet z.T. mit personenbezogenen Daten bzw. Daten, die auf Grundlage personenbezogener Daten erzeugt wurden. Dazu zählen z.B. die Daten des Mikrozensus oder die Daten der Rentenversicherung. Dementsprechend unterliegen die Forschungsdaten den o.g. datenschutzrechtlichen Vorgaben. Es ergibt sich daher zunächst folgende Anforderung:

Anforderung 65. ***Innerhalb der virtuellen Arbeitsumgebung sind die Vorgaben des Bundesdatenschutzgesetzes hinsichtlich der automatisierten Verarbeitung personenbezogener Daten zu beachten und entsprechend umzusetzen.***

Für prozessproduzierte Daten, wie z.B. aus der Rentenversicherung, sind zusätzlich Datenschutzvorschriften nach dem Sozialgesetzbuch zu berücksichtigen[12]. Ebenso gilt, dass die sozioökonomische Berichterstattung nicht Eigentümer der Originaldatensätze ist und die Verwertungsrechte bei den FDZ verbleiben. Die FDZ erlauben, u.a. aufgrund der anzuwendenden Datenschutzbestimmungen[13], die Nutzung der Originaldatensätze sowie der daraus abgeleiteten Arbeitsdaten nur unter Einhaltung strenger Vorschriften. Auch hieraus lässt sich eine Anforderung ableiten:

Anforderung 66. ***Der Schutz vor unberechtigtem Zugriff jedweder Art auf Forschungsdaten muss gewährleistet werden können. Da dies nicht für alle Forschungsdaten in gleicher Form gilt, ist eine angepasste Definition der Zugriffsrechte notwendig.***

Um Scientific Use Files verwenden zu können, müssen Wissenschaftler vertraglich zusichern, dass die Daten nur durch einen definierten Personenkreis und nur für einen definierten Zeitraum genutzt werden. Dabei existieren keine einheitlichen Vorgaben zu den Personenkreisen oder Zeiträumen, so dass diese jeweils individuell definiert werden müssen. Die Arbeitsdaten, die aus den SUF abgeleitet wurden, sind von den jeweiligen Vorgaben zur Nutzungsdauer und der Zugriffsberechtigungen ebenso betroffen. Eine virtuelle Forschungsumgebung soll daher die Forscher bei der Einhaltung der individuellen Vorgaben unterstützen.[14]

Anforderung 67. ***Eine Nutzung von Forschungsdaten soll gegebenenfalls ausschließlich in individuell definierbaren Zeiträumen möglich sein.***

Anforderung 68. ***Die Nutzung von Forschungsdaten soll durch ausreichend feingranulare Zugriffsberechtigungen eingeschränkt werden können.***

Anforderung 69. ***Komplexe Organisationsstrukturen, wie die der soeb, sollten bei der Definition von Zugriffsberechtigungen darstellbar sein.***

---

[12] §75 Sozialgesetzbuch X (SGB X).

[13] §3, Abs. 2 – 11 BDSG.

[14] Es gibt hier Initiativen auch der FDZ zur Standardisierung bzw. Vereinfachung der Situation. Die virtuelle Arbeitsumgebung kann hierzu beitragen.

Auch für die Metadaten von Forschungsdaten existieren in einer kollaborativen Arbeitsumgebung gewisse Anforderungen hinsichtlich Sicherheit. So soll es Mitarbeitern in unterschiedlichen Gruppen oder Projekten, in denen die virtuelle Arbeitsumgebung Anwendung findet, nicht möglich sein, Zugriff auf forschungs- oder ergebnisrelevante Information einer jeweils anderen Gruppe oder eines anderen Forschungsprojektes zu erhalten. Wesentliche Informationen können auch anhand der Metadaten von Forschungsdaten direkt abgelesen, oder zumindest abgeleitet werden. Daher, gelten die o.g. sicherheitsrelevanten Anforderungen an Forschungsdaten ebenso für die jeweiligen Metadaten.

Die technologischen Entwicklungen hinsichtlich Sicherheit haben in der Vergangenheit immer wieder gezeigt, dass es im Bereich der digitalen Datenverarbeitung nie eine allumfassende Sicherheit geben kann. Selbst die oben genannten weitreichenden Forderungen des Bundesdatenschutzgesetzes können nicht alle Aspekte und mögliche Angriffspunkte abdecken und betrachten. Vor allem dann nicht, wenn es technologische Weiterentwicklungen gibt, die bei den Forderungen noch gar nicht bedacht sind. Um dennoch den bereits genannten Notwendigkeiten an Sicherheit und Datenschutz so weit wie möglich zu genügen, wird die folgende Anforderung definiert:

Anforderung 70. ***Bei der Umsetzung von Datensicherheit und Datenschutz sollte die virtuelle Arbeitsumgebung jene Technologien einsetzen, die nach aktuellem Forschungsstand eine bestmögliche oder zumindest hinreichende Sicherheit bieten.***

Die Lieferung der Scientific Use Files erfolgt gegenwärtig auf dem Postweg. Der Einsatz einer virtuellen Forschungsumgebung böte den entsprechenden FDZ zusätzlich die Möglichkeit, die SUF den Wissenschaftlern kontrolliert digital zur Verfügung zu stellen. Hierzu sind allerdings Abstimmungen mit den FDZ notwendig, um die Möglichkeit der Integration zu erörtern. Grundsätzlich bietet sich jedoch mit einer virtuellen Arbeitsumgebung ein zuverlässiger Bereitstellungsort für die SUF und verbesserte Mechanismen für netzbasierte Wege der Bereitstellung.

3.3 Datenverarbeitung

Der Hauptaspekt der sozioökonomischen Berichterstattung in Bezug auf die Datenverarbeitung im Rahmen einer virtuellen Forschungsumgebung besteht in den eingangs beschriebenen drei Prozessen: Datenaufbereitung, Datenverknüpfung und Datenauswertung. Die Aufbereitung erzeugt dabei Arbeitsdaten, die wiederum mit weiteren Arbeitsdaten verknüpft werden können. Zur Erzeugung von Zwischenergebnissen (als Arbeitsdaten bezeichnet) oder finaler Ergebnisdaten wird die Datenauswertung betrieben. Zur Datenverarbeitung werden Syntaxdateien mit den Arbeitsdaten verknüpft und durch die jeweils verwendete Statistik-Software verarbeitet. Dabei stellen die Syntaxdateien die algorithmische Beschreibung der Berechnungsabläufe bereit, die durch die Statistik-Software auf die Arbeitsdaten angewendet wird. Für die Datenverarbeitung werden unterschiedlich hohe Rechenzeitaufwände in Abhängigkeit von der jeweiligen Aufgabenstellung und dem Umfang der zu verarbeitenden Arbeitsdaten benötigt.

Anforderung 71. ***Die virtuelle Forschungsumgebung soll die Möglichkeit bieten, Arbeitsdaten auszuwählen und diese mit Syntaxdateien für die Datenverarbeitung verknüpfen zu können.***

Integration von R oder anderen Statistik-Paketen

Für die sozioökonomische Berichterstattung ist insbesondere die Nutzung von Statistik-Software von essenzieller Bedeutung. Proprietäre Statistik-Lösungen können gegenwärtig nicht oder nur begrenzt in einer virtuellen Forschungsumgebung eingesetzt werden, da die Lizenzbedingungen der Hersteller dies nicht unterstützen bzw. die Kosten für die übergreifenden Lizenzen jeglicher Wirtschaftlichkeit entbehren. Erfahrungen in D-Grid haben gezeigt, dass es nur sehr schwer ist, kommerzielle Hersteller zur Lizenzierung ihrer Produkte in Verbundstrukturen zu bewegen. Daher wird die Nutzung einer Open-Source-Lösung als sinnvoll erachtet.

Da bereits in der sozioökonomischen Berichterstattung z.T. die Open-Source-Statistiklösung R (http://www.r-project.org/) eingesetzt wird, besteht damit eine Möglichkeit, die Lizenzierungsproblematik für bestimmte Anwendungsfälle zu umgehen. Hierbei ist allerdings zu beachten, dass voraussichtlich nur ca. 30% der *soeb*-Wissenschaftler R einsetzen werden.

Gegebenenfalls kann auch Stata für die statistische Datenverarbeitung im Grid eingesetzt werden. Das Lizenzmodell erlaubt die Installation von Rechenservern. Allerdings ist hier durchaus fraglich, welche Kosten für eine oder mehrere Installationen entstehen würden. Mit Hilfe des Grids könnten allerdings die Zugriffe auf diesen Server so kontrolliert werden, dass die Lizenzbedingungen[15] (z.B. Nutzung nur durch eine bestimmte Anzahl von Benutzern gleichzeitig – concurrent licence) Beachtung finden.

Idealerweise können Forscher der sozioökonomischen Berichterstattung auf benötigte Rechenleistung direkt aus der virtuellen Forschungsumgebung zurückgreifen, um damit die lokalen Arbeitsplatzsysteme zu entlasten und die Prozesse der Datenverarbeitung zu beschleunigen. Diese Entlastung ist essentiell für die sozioökonomische Berichterstattung, da gegenwärtig z.B. einfache Arbeitsplatzsysteme für mehr als 24 Stunden durch Datenverarbeitungsprozesse belegt werden. Ein weiterer Vorteil liegt in der möglichen parallelen Verarbeitung verschiedener Varianten zu den einzelnen Prozessen, um so das iterative Vorgehen gezielt zu unterstützen.

Anforderung 72. Die Statistik-Software R soll in die virtuelle Forschungsumgebung eingebettet sein. Gegebenenfalls sind auch weitere Software-Pakete zu unterstützen.

Ausgehend von der Situation, dass die Wissenschaftler der sozioökonomischen Berichterstattung für statistische Berechnungen vielfach lokale Kapazitäten verwenden müssen, soll bei Bedarf „On Demand“ Rechenleistung durch die virtuelle Forschungsumgebung bereitgestellt werden. Hierdurch kann, bei entsprechender Vorbereitung der Syntax-Dateien, auch eine parallelisierte Abarbeitung der Berechnungen erfolgen. Eine Unterstützung für das Starten solcher Berechnungen ist vorzusehen, die (algorithmische) Parallelisierung der Syntax bleibt jedoch Aufgabe des Forschers, die ggf. durch Beispiele unterstützt werden kann. Die einfache Form des gleichzeitigen Ausführens von mehreren voneinander unabhängigen Berechnungen ist hingegen von vornherein möglich.

Anforderung 73. Durch die virtuelle Forschungsumgebung soll externe Rechenleistung benutzerfreundlich zur Datenverarbeitung mit R oder einem Äquivalent bereitgestellt werden können.

[15] Siehe hierzu die Lizenzbedingungen von Stata unter: StataCorp LP (2010): End-User License Agreement (EULA), Letzter Zugriff: 2010.06.28, URL: http://www.stata.com/order/licterms.html.

Anforderung 74. ***Für die Nutzung der externen Rechenleistung zur Datenverarbeitung ist eine möglichst nahtlose Integration von R oder einem Äquivalent in die virtuelle Forschungsumgebung notwendig.***

3.4 Kollaborative Forschung in der sozioökonomischen Berichterstattung

Die kollaborative Forschung bekommt einen immer größeren Stellenwert, auch in der sozioökonomischen Berichterstattung. Hierdurch ergeben sich jedoch ganz neue Anforderungen an die Infrastruktur, mit deren Hilfe Forschung betrieben wird. Die schon beschriebenen Maßnahmen zur besseren, und für alle, unter Beachtung der Rechte und Vorschriften, zugänglichen gemeinsamen Datenverwaltung sind hier die tragende Struktur. Ein weiterer wesentlicher Aspekt ist dabei die Abstimmung der einzelnen *soeb*-Wissenschaftler auf dem Weg der Erarbeitung gemeinsamer Ergebnisse. Hierbei steht vor allem die Entwicklung gemeinsamer, oder aufeinander abgestimmter Syntaxdateien im Vordergrund. Bisher wurden hierzu der Datenaustausch per E-Mail und das Telefon als Kommunikationsmedium genutzt. Die heutigen technischen Möglichkeiten bieten jedoch wesentlich weiter führende Funktionalitäten an, mit denen eine kollaborative Arbeit erleichtert werden kann.

Bei der gemeinsamen Bearbeitung von Syntaxdateien liegen ähnliche Anforderungen wie bei der Softwareentwicklung vor. Daher können auch organisatorische Ansätze aus der Softwareentwicklung im Bereich der sozioökonomischen Berichterstattung Anwendung finden. Für die gemeinsame Arbeit an einer Syntaxdatei ist es z.B. sinnvoll, dass die Syntaxdatei selbst in einem zentralen System vorgehalten wird, so dass jeder Wissenschaftler sehen kann, welchen Stand die Syntaxdatei gerade hat, wer wann Änderungen vorgenommen hat und wie diese Änderungen im Detail aussehen. Um solche Herausforderungen zu adressieren, wurden bereits in den Kapiteln 3.2.1 und 3.2.2 diverse Anforderungen an eine Datenhaltung und die davor gelagerte virtuelle Arbeitsumgebung spezifiziert. Diese sollen hier nicht erneut aufgeführt werden.

Für die kollaborative Syntaxentwicklung ist neben der Datenverwaltung auch die Kommunikation ein entscheidender Faktor. Den Wissenschaftlern soll es auf einfache Weise möglich sein, während der Arbeit an einer Syntaxdatei mit den Kollegen in Kontakt zu treten, um entweder einfach Fragen stellen zu können oder über eine Lösung zu diskutieren. Gegebenenfalls macht es auch Sinn eine Diskussion mit den dazugehörenden Argumenten und Aussagen aufzubewahren, so dass bei zukünftigen und ähnlichen Diskussionen auf die erarbeitete Lösung zurückgegriffen werden kann. Daraus ergeben sich folgende Anforderungen:

Anforderung 75. ***Die virtuelle Arbeitsumgebung sollte den Wissenschaftlern verschiedene integrierte Kommunikationsmedien (online oder offline) zur Verfügung stellen.***

Anforderung 76. ***Die Kommunikationsmedien sollten wenn möglich protokollierbar sein.***

Anforderung 77. ***Die Kommunikationsmedien sollten einen direkten Datenaustausch erlauben.***

Anforderung 78. ***Wenn möglich sollte die virtuelle Arbeitsumgebung während einer Diskussion über die Kommunikationsmedien eine gemeinsame und zeitgleiche Bearbeitung von Daten durch mehrere Wissenschaftler erlauben.***

Neben diesen interaktiven Anforderungen werden in der sozioökonomischen Berichterstattung jedoch noch weitere Funktionalitäten benötigt. Z.B. sollen Informationen jeglicher Art, Syntaxbeispiele und -dateiauszüge, projektinterne Festlegungen oder sonstige Daten zentral abgelegt, verwaltet und bearbeitet werden können. Damit wird die allgemeine Projektarbeit im Team effektiv und effizient unterstützt. Weiterhin sollen Ergebnisse einer Forschung oder auch aktuelle Forschungsstände auf Projektwebseiten veröffentlicht werden können. Daraus ergeben sich folgende Anforderungen:

Anforderung 79. ***Für die teamorientierte Projektarbeit werden Funktionalitäten zur zentralen Ablage und Bearbeitung von projektinternen Informationen benötigt.***

Anforderung 80. ***Für die Veröffentlichung von Zwischen- und Endergebnissen einer Forschung werden einfache Tools zum Aufsetzen und Bearbeiten eigener Webseiten benötigt, die gegebenenfalls in die eigentlichen Projektwebseiten integriert werden können.***

4 Existierende Lösungen

4.1 Grid-Technologie als Lösungsansatz

Für die in der sozioökonomischen Berichterstattung durchgeführten statistischen Berechnungen existieren gewisse Mindestanforderungen an die Fähigkeiten der verwendeten Ressourcen. So sind die zu verarbeitenden Datensätze mit durchaus mehreren Gigabyte an Größe mittlerweile nicht mehr auf jedem System speicher- und analysierbar. Außerdem benötigen die eingesetzten Algorithmen bei diesen Datenmengen selbst auf schnellen Rechnern eine lange Verarbeitungszeit. Einzelne Analysen sind zwar noch auf modernen Desktop-PCs durchführbar, nur sind diese dann vollständig ausgelastet und für paralleles Weiterarbeiten des Wissenschaftlers während der Wartezeit nicht mehr einsetzbar. Viele Wissenschaftler arbeiten daher schon mit mehreren Desktopsystemen parallel, wobei auch diese bei mehreren parallelen Berechnungen schnell ihre Grenzen erreichen.

Die sozialwissenschaftliche Forschung wird daher in Zukunft immer häufiger die Rechen- und Speicherleistung größerer Rechnerinfrastrukturen in Anspruch nehmen. Nur dadurch können bestimmte Forschungen überhaupt noch realistisch durchgeführt werden. Außerdem erlauben derartige Systeme ganz neue Ansätze. Mit den zur Verfügung stehenden Rechenressourcen können z.B. verschiedene Versionen oder Varianten einer Syntaxdatei auf den gleichen Daten zeitgleich ausgeführt, und die Ergebnisse, die in absehbarer Zeit vorliegen, mit einander verglichen werden. Anhand der unterschiedlichen Ergebnisse lassen sich zum einen die Syntaxdateien bewerten, und zum anderen werfen die Ergebnisunterschiede gegebenenfalls ganz neue Forschungsfragen oder Erkenntnisse auf. Dadurch wird der oben beschriebene explorative Charakter der Forschung noch stärker adressiert. Vor allem aber werden längere Wartezeiten durch die Parallelisierung von unterschiedlichen Berechnungen (z.B. mit unterschiedlichen Syntaxvarianten), die sonst aufgrund von Ressourcenknappheit serialisiert werden müssten, vermieden.

Durch die Nutzung komplexerer Infrastrukturen werden die Forscher in Zukunft sicherlich auch zu ganz neuen und innovativen Herangehensweisen inspiriert, die gezielt die neuen Methoden und Möglichkeiten ausnutzen werden, deren konkreter Ansatz allerdings den Experten überlassen sei.[16]

Grid-Technologie

Der Begriff Grid wird sehr unterschiedlich interpretiert. Dies kann politische Gründe haben, ist jedoch oft auch auf mangelnde Erfahrung zurückzuführen. Die eigentliche Definition nach Foster in „What is the Grid“ von 2002 (hier frei übersetzt) betrachtet ein System als Grid, so lange es

- dezentral organisierte Ressourcen integriert und koordiniert,
- allgemeingültige, offene und standardisierte Protokolle und Schnittstellen verwendet und
- durch den kombinierten Einsatz der Ressourcen unter der Betrachtung ihrer einzelnen Qualitätsmerkmale deren Nutzwert signifikant erhöht.

[16] Es gibt durchaus Möglichkeiten, mit entsprechender Vorbereitung der Syntax, auch eine parallele Bearbeitung eines Datensatzes (hier: gleichzeitig von vielen CPUs) mit erheblichem Zeitgewinn durchzuführen. Hierfür ist allerdings das Know How des Wissenschaftlers erforderlich.

Diese Definition ermöglicht es, zu verifizieren, ob ein System ein Grid ist, sagt jedoch wenig über den aktuellen Stand der Grid-Technologie aus. Es stellt sich daher eher die Frage, wie ein Grid heutzutage aufgebaut wird, bzw. ist.

Existierende Grid-Infrastrukturen bedienen sich so genannten Grid-Middlewares, um Ressourcen mit einander zu integrieren und zu nutzen. Der Begriff „Ressourcen" muss hierbei sehr weitläufig verstanden werden, da es sich dabei nicht nur um Computersysteme, sondern auch um andere technische Geräte, beispielsweise elektronisch gesteuerte Forschungsapparaturen, handeln kann. Für den Rahmen dieser Expertise und der sich daraus ergebenden Anforderungen reicht jedoch die vereinfachte Betrachtung von Computern mit ihren Rechen- und Speicherkapazitäten als Ressourcen eines Grids aus.

Eine Grid-Middleware ist somit – vereinfacht gesehen – eine bestimmte Kombination von Softwarekomponenten, mit der die transparente Nutzung von Computern und den von ihnen bereitgestellten Ressourcen ermöglicht wird. Dafür werden die Softwarekomponenten der so genannten Middlewares[17] auf den entsprechenden Computern installiert, mit den Ressourcen des Computers konfiguriert und untereinander registriert. Diese Komponenten stellen weiterhin die von Foster geforderten einheitlichen Schnittstellen zur Nutzung der Ressourcen zur Verfügung. Als Ergebnis erscheinen die vielen heterogenen Systeme als ein großes einheitliches System mit entsprechend großen Kapazitäten.

Die Nutzung eines solchen Systems bzw. der bereitgestellten Ressourcen erfolgt nun über Schnittstellen, die von der Middleware zur Verfügung gestellt werden. Diese Schnittstellen können meist direkt von Computerprogrammen angesprochen werden. Viele Middlewares bieten aber auch graphische Oberflächen zur „manuellen" Nutzung der Ressourcen an.

Je nach Ausprägung der Schnittstellen bzw. Oberflächen und den Anwendungsanforderungen ist der Zugriff auf die Ressourcen mehr oder weniger transparent. So gibt es zum Beispiel Middlewares, die es dem Benutzer erlauben, für das Ablegen von Dateien genau zu bestimmen, auf welchen konkreten Computern eines Grids eine Datei und deren Sicherheitskopien abgelegt werden sollen. Manche Middlewares erlauben es, von den tatsächlichen Ressourcen zu abstrahieren und bieten dem Benutzer eine einfache Schnittstelle, hinter der die tatsächlich verwendete Speicherressource automatisch ausgewählt wird und auch die Erstellung von Sicherheitskopien für den Benutzer transparent passiert. In jedem Fall übernimmt die Middleware den Transport der Datei zum Zielcomputer.

Um die Rechenkapazität eines Grids zu verwenden, werden so genannte Jobs an das Grid, genauer, an die Grid-Middleware geschickt. Ein Job spezifiziert dabei, vereinfacht gesehen, einen Programmaufruf mit Aufrufparametern sowie Ein- und Ausgabedateien. Im einfachsten Fall wählt die Grid-Middleware eine freie Ressource, sprich einen Computer, aus und schickt den Job sowie die Eingabedateien dorthin. Dort wird der Job von der Grid-Middleware bearbeitet und an das lokale Betriebssystem übergeben, sprich der Programmaufruf ausgeführt. Die Ausgabedateien werden nach Abschluss des Jobs von dem Computer abgeholt und dort abgelegt, wo es in der Jobbeschreibung festgelegt wurde. Diese Schritte werden automatisch abgearbeitet und von der Middleware für den Nutzer weitestgehend transparent umgesetzt.

Auch bei der Nutzung von Rechenressourcen gibt es Unterschiede in der Transparenz. Es kann zum Beispiel für bestimmte Berechnungen notwendig sein, dass auf den Computern, auf denen die Berechnungen ausgeführt werden sollen, bestimmte Voraussetzungen vorliegen müssen, damit die Berechnung möglich ist. Solche Voraussetzungen können sich an der Ausstattung des Computers mit Re-

[17] Eine Middleware dient dem vereinfachten Zugriff auf darunterliegende komplexe Komponenten.

chenleistung, Speicherkapazität und besonderen Hardwarekomponenten, aber auch an der auf dem Computer installierten Software inklusive des Betriebssystems orientieren. Nicht jedes an ein Grid angeschlossene Teilsystem erfüllt die gleichen Kriterien. In solchen Fällen kann es Sinn machen, für einen Job genau zu bestimmen, auf welchen Computern eine Berechnung ausgeführt werden soll. Manche Middlewares bieten hierfür jedoch intelligente Auswahlalgorithmen an, die anhand der Beschreibung der Ressourcen und des Jobs automatisch das Zielsystem auswählen können.

Zusammengefasst ist die aktuelle Grid-Technologie fähig, die Ressourcen von oft sehr unterschiedlichen Systemen zu einem großen, einheitlich zu nutzenden Ressourcenpool zusammenzufassen und diese Ressourcen transparent und gezielt nutzbar zu machen.

4.2 Organisatorische Infrastruktur

4.2.1 D-Grid

Das Bundesministerium für Bildung und Forschung (BMBF) fördert seit 2005 die deutsche Initiative zum Auf- und Ausbau einer bundesweiten Grid-Infrastruktur mit dem Namen D-Grid (http://www.d-grid.de). Hierbei werden in verschiedenen Förderstufen gezielt Projekte ins Leben gerufen, deren Ziel es unter anderem ist, die technologischen und organisatorischen Grundlagen für das D-Grid zu schaffen. Des Weiteren werden durch die direkte Anwendung der Infrastruktur Anforderungen abgeleitet und damit die Grid-Technologie sowie deren Einsatz Schritt für Schritt weiterentwickelt.

In der Wissenschaft gibt es verschiedene Gebiete, die bereits seit Jahren komplexe Rechnerinfrastrukturen verwenden. Die bemerkenswertesten Beispiele kommen hier aus den Naturwissenschaften, wie z.B. aus der Hochenergiephysik mit dem Large Hadron Collider (LHC, http://lhc.web.cern.ch/lhc/), für dessen Datenverarbeitung das Worldwide LHC Computing Grid (WLCG, http://lcg.web.cern.ch/lcg/) aufgebaut wurde. Aber auch bisher eher weniger technisch orientierte Wissenschaften sehen einen immer größeren Bedarf an computergestützten Forschungsinfrastrukturen. Dies zeigt sich unter anderem an der Tatsache, dass auch im Rahmen der D-Grid-Initiative bereits Projekte gefördert werden, die z.B. den bisher eher „analog-arbeitenden“ Wissenschaften Zugang zu digitaler fachwissenschaftlicher Datenverarbeitung ermöglichen. So werden z.B. im Rahmen von TextGrid (http://www.textgrid.de) nicht nur grundlegende Funktionen für die geistes- und kulturwissenschaftliche Datenhaltung implementiert, sondern auch und vor allem eine vernetzte Forschungsumgebung für die wissenschaftliche Datenbe- und -verarbeitung geschaffen.

Auch in Zukunft setzt die D-Grid-Initiative darauf, weitere Anwendergruppen in die geschaffene Infrastruktur einzubinden. Dies wird vor allem durch das vom BMBF geförderte Projekt WissGrid (Grid für die Wissenschaft, http://www.wissgrid.de) deutlich, welches unter anderem darauf abzielt, bisher nicht im D-Grid vertretenen akademischen Communitys einen leichteren Zugang zur Grid-Technologie und speziell zum D-Grid zu schaffen.

4.2.2 WissGrid

WissGrid ist ein Projekt im Rahmen der vom BMBF (Bundesministerium für Bildung und Forschung, siehe: http://www.bmbf.de) geförderten D-Grid Initiative (siehe http://www.d-grid.de). Die Förderung von WissGrid ist Teil der aktuellen dritten Förderphase. Dabei verfolgt WissGrid das übergeordnete Ziel, die Grid-Technologie als virtuelle Forschungsumgebung für den akademischen Sektor nachhaltig zu etablieren. Dies erfolgt in den drei zentralen Arbeitspaketen:

1. AP1 – Betriebsmodell

2. AP2 – Blaupausen
3. AP3 – Langzeitarchivierung

Ein Betriebsmodell ist für ein nachhaltiges Angebot einer virtuellen Forschungsumgebung notwendig. Dabei soll ein Finanzierungsmodell mit Förderern bzw. Förderinstitutionen abgestimmt und etabliert werden. Darauf aufbauend sollen Strukturen für die Koordination von Nutzerinteressen aufgebaut werden. Der Kern der Koordination liegt in der effizienten Steuerung der Nutzeranforderungen an die virtuelle Infrastruktur. Zudem benötigen die akademischen Nutzer eine fundierte Vertretung ihrer Interessen gegenüber den Anbietern von Rechenleistung – den wissenschaftlichen Rechenzentren, die sich bereits in der GAUSS-Allianz sowie den ZKI (Zentren für Kommunikation und Informationsverarbeitung in Lehre und Forschung e.V., siehe: http://www.zki.de) organisieren.

Die so genannten Blaupausen stellen die Dokumentation von Bauplänen zum Aufbau von virtuellen Organisationen bzw. so genannten Grid Communitys im Rahmen einer virtuellen Forschungsumgebung bereit. Akademische Communitys, die noch keine Erfahrung mit virtuellen Forschungsumgebungen haben, sollen anhand der Baupläne in die Lage versetzt werden, eine auf Grid-Technologie basierende virtuelle Forschungsumgebung nutzen zu können. Ebenso wurde bereits ein Fachberaterteam aus erfahrenen Spezialisten der Grid-Projekte aus den ersten beiden Förderphasen zusammengestellt. Damit stehen Ansprechpartner für neue Communitys zur Verfügung.

Langzeitarchivierung digitaler Objekte in einer virtuellen Forschungsumgebung ist ein Aspekt, der sämtliche akademischen Communitys betrifft und die nachhaltige Nutzung ermöglicht. Die Langzeitarchivierung soll die Nachnutzung digitaler Forschungsdaten direkt in der virtuellen Forschungsumgebung ermöglichen. Unter Nachnutzung wird das Wiederauffinden und die Möglichkeit zur intellektuellen Verwendung bzw. Interpretierbarkeit langzeitarchivierter Daten verstanden. In diesem Sinne wird eine in die Grid-Technologie eingebettete Architektur zur Langzeitarchivierung erarbeitet. Des Weiteren werden generische Dienste zur Verfügung gestellt, die von den wissenschaftlichen Communitys um fachspezifische Komponenten erweitert werden können.

4.3 Datenverwaltung

Die Grid-Technologie erlaubt unter anderem die Nutzung verteilter Ressourcen zur Ablage von Daten. Die meisten Grid-Middlewares, die zum Teil auch im Rahmen der D-Grid Initiative eingesetzt werden, liefern bereits eine grundlegende Unterstützung zum Speichern von Daten, und teilweise auch zum Datenmanagement mit. Weiterhin existieren Systeme, die auf den Middlewares aufsetzen und die Datenablage und das Datenmanagement weiter optimieren und einfacher nutzbar machen. In diesem Abschnitt soll kurz auf die Middlewares und die existierenden Datenmanagementsysteme eingegangen, und deren Einsatz im Rahmen der virtuellen Arbeitsumgebung für die sozioökonomische Berichterstattung betrachtet werden.

4.3.1 Grid-Middleware

Globus Toolkit 4

Eine im D-Grid etablierte Middleware ist das Globus Toolkit 4[18]. Hier ist ein einfaches Ablegen von Daten im Grid möglich. Globus erlaubt standardmäßig die automatische Verwaltung von Dateikopien,

[18] Detaillierte Informationen unter: The Globus Alliance (2010): Globus Toolkit, Letzter Zugriff: 2010.05.27, URL: http://www.globus.org/toolkit/.

um mit Hilfe einer redundanten Datenhaltung das Verlustrisiko zu minimieren und einen effizienten parallelen Zugriff auf Daten zu ermöglichen. Der Fokus von Globus liegt jedoch eher auf der Datenverarbeitung und weniger auf der Datenhaltung. Dies zeigt sich unter anderem daran, dass keine Metadatenverwaltung integriert ist. Ein weiterer zu beachtender Aspekt ist, dass im D-Grid hauptsächlich das Globus Toolkit in der Version 4.X eingesetzt wird. Die aktuelle Version ist jedoch 5.X. Aus diesem Grund gibt es offene Fragen hinsichtlich der zukünftigen Wartung von Globus Toolkit 4.X und einer potentiell notwendigen Migration auf die Version 5.X und den damit verbundenen und zu erwartenden Inkompatibilitäten. Da Globus nicht allen Anforderungen der virtuellen Arbeitsumgebung für die sozioökonomische Berichterstattung gerecht wird und auch der zukünftige Einsatz im D-Grid Fragen aufwirft, wird die Verwendung von Globus Toolkit als alleiniges System zur Datenablage und zum Datenmanagement nicht empfohlen.

LCG/gLite

Mit LCG/gLite[19] (kurz gLite) existiert eine weitere Grid-Middleware, die ebenfalls auf diversen Ressourcen im D-Grid installiert ist. Die von gLite verwalteten Grid-Ressourcen werden als virtuelles großes Dateisystem zur Verfügung gestellt, auf welches über verschiedene Protokolle Zugriff besteht. Der interne Katalog erlaubt bei gLite auch das Ablegen von Kommentaren der Anwender, was einer rudimentären Metadatenverwaltung ähnlich ist. Doch auch hier müsste sehr viel der in der virtuellen Arbeitsumgebung benötigten Funktionalität außerhalb implementiert werden. Die Umsetzung einer Suche nach Daten anhand der Metadaten wäre nur mit großem Aufwand realisierbar. Daher wird auch gLite nicht als optimale alleinige Lösung für die Datenablage und das Datenmanagement im Rahmen der sozioökonomischen Berichterstattung angesehen.

Unicore 6

Eine der im D-Grid installierten Grid-Middlewares ist Unicore 6[20]. Es erlaubt die einfache Ablage von Dateien im Grid und beinhaltet Tools, mit denen grundlegende Operationen eines Dateisystems auf entfernten Rechnern ausgeführt werden können. Damit ist das einfache Abspeichern, Bearbeiten, Lesen und Löschen von Dateien und Verzeichnissen im Grid möglich. Unicore besitzt jedoch keine eigene Metadaten- und Dateiverwaltung. Dies erlaubt dem Anwender zwar die vollständige Kontrolle über die Lokation der Daten, zwingt ihn jedoch auch zur eigenen Verwaltung. Eine mögliche Lösung hierfür wurde im Rahmen des durch die EU geförderten Projekts Chemomentum entwickelt[21]. Zum Zeitpunkt der Erstellung der vorliegenden Expertise war die Projektwebsite jedoch nicht, oder nicht mehr

---

[19] Detaillierte Informationen finden sich unter anderem in: Burke, S; Campana, S; Lanciotti, E, et al. (2009): gLite 3.1 User Guide, Sciab`a, A. Manuals Series, Genf, Schweiz, CERN,

https://edms.cern.ch/file/722398//gLite-3-UserGuide.pdf.

[20] UNICORE - Distributed computing and data resources (2010), Letzter Zugriff: 2010.05.12, URL: http://www.unicore.eu.

[21] Siehe Rasch, K; Schöne, R; Ostropytskyy, V, et al. (2009): The Chemomentum Data Services – A Flexible Solution for Data Handling in UNICORE, Euro-Par 2008 Workshops - Parallel Processing, VHPC 2008, UNICORE 2008, HPPC 2008, SGS 2008, PROPER 2008, ROIA 2008, and DPA 2008, Las Palmas de Gran Canaria, Spain, Letzter Zugriff: 2010.05.27, URL: http://www.springerlink.com/content/ 2u70m8443180620m/ fulltext.pdf.

online. Außerdem endete der Förderzeitraum des Projekts bereits Anfang 2009[22]. Es wird daher davon ausgegangen, dass die Unterstützung für diese Erweiterung wenig zukunftsträchtig ist. Die ausschließlich von Unicore zur Verfügung gestellten Funktionalitäten decken jedoch nicht alle Anforderungen für eine Datenablage im Rahmen der virtuellen Arbeitsumgebung ab, weshalb von einem Einsatz von Unicore als alleiniges System zur Datenablage und zum Datenmanagement abgeraten wird.

4.3.2 Datenmanagement Systeme, die auf Grid-Middleware basieren

dCache/SRM

Ein auf Datenhaltung optimiertes und im D-Grid etabliertes System ist dCache/SRM[23] (kurz dCache). Es setzt auf den existierenden Middlewares auf und vereinigt, ähnlich wie gLite, die Grid-Ressourcen für den Anwender transparent zu einem großen virtuellen Dateisystem. Allerdings unterscheiden sich hierbei die Zugriffsprotokolle gegenüber gLite. Die Vorteile von dCache/SRM sind die intern je nach Konfiguration automatisch vorgenommene redundante Datenhaltung sowie die Lastoptimierung. Außerdem ist dCache fähig Datenverluste selbstständig zu erkennen und mit Hilfe der Redundanzen zu kompensieren. Nachteilig ist, dass auch dCache keine Metadatenverwaltung ermöglicht. Daher ist auch dCache allein für die Datenhaltung der virtuellen Arbeitsumgebung nicht ausreichend.

SRB

Das D-Grid wurde teilweise mit einem weiteren System zur Datenablage und zum Datenmanagement, dem Storage Resource Broker (kurz SRB[24]) ausgestattet. Auf der D-Grid Website[25] wird jedoch auf iRODS (wird im folgenden Absatz beschrieben) als Nachfolger verwiesen, weshalb auch dieses System keinen Einsatz in der sozioökonomischen Berichterstattung finden sollte.

iRODS

iRODS (Integrated Rule-Oriented Data System[26]), der Nachfolger von SRB, kann zum Aufbau von Daten-Grids, digitalen Bibliotheken, persistenten Archiven und Echtzeit-Datensystemen verwendet werden. Dafür kann es auf einfachen PC-Systemen oder auf existierenden Grid-Infrastrukturen mit entsprechenden Middlewares aufsetzen. Letzteres ist bei der Installation von iRODS im D-Grid der Fall.

---

[22] Vgl. Kurzbeschreibung des Projekts Chemomentum auf den Seiten des Forschungszentrums Jülich. Chemomentum - Grid Services based Environment to enable Innovative Research (2010), Letzter Zugriff: 2010.05.27, URL: http://www.fz-juelich.de/jsc/grid/Chemomentum.

[23] dCache (2010), Letzter Zugriff: 2010.05.21, URL: http://www.dcache.org.

[24] Dokumentation unter: SRB - The DICE Storage Resource Broker (2010), Letzter Zugriff: 2010.05.27, URL: http://www.sdsc.edu/srb/index.php/Main_Page.

[25] Zitat auf http://dgi.d-grid.de/index.php?id=24: „Ein weiteres Datenmanagementsystem, der SRB, erhielt aufgrund schwieriger Lizenzbedingungen nur in der Medizin-Community Akzeptanz. Die Community wurde vom Projekt in der Aufbauphase unterstützt. Im Rahmen der D-Grid Strategie wird aber der SRB nicht weiter vorgehalten und der weitere Focus hin zum Nachfolgesystem iRODS verschoben, das auch in Zukunft unterstützt werden soll.“ Letzter Zugriff: 27.05.2010.

[26] Beschreibung unter IRODS (2010), Letzter Zugriff: 2010.05.27, URL: https://www.irods.org.

iRODS ist sehr flexibel und kann auf unterschiedlichste Anforderungen hin angepasst werden. Kern der Software ist ein regelbasiertes System, mit welchem für eine erweiterbare Liste von Datenoperationen definierte Reihenfolgen von frei wählbaren Schritten ausgeführt werden können. Dadurch sind auch die Kernfunktionalitäten von iRODS mit begrenztem Aufwand beliebig erweiter- und veränderbar. iRODS bietet außerdem rudimentäre Möglichkeiten zur Verwaltung von Metadaten. Allerdings können mit diesen erneut nicht alle Anforderungen der virtuellen Arbeitsumgebung abgedeckt werden. Durch seine Flexibilität könnte iRODS jedoch um eine komplexere Metadatenverwaltung ergänzt werden. Auch wenn iRODS im D-Grid noch keine flächendeckende Anwendung findet, so planen bereits mehrere Projekte fest mit dessen Einsatz. Es wird unter anderem im Rahmen der im Projekt WissGrid (Grid für die Wissenschaft[27]) entwickelten Langzeitarchivierungsstrategien Anwendung finden. Ebenso existiert für iRODS eine Schnittstelle in der Repository-Software Fedora, so dass sich hiermit eine funktionale Ergänzung zu einer Gesamtlösung hin in einer verteilten IT-Infrastruktur realisieren lässt. Damit stellt iRODS, so wie es im D-Grid bereits installiert ist, eine erste mögliche Lösung für die Datenhaltung der virtuellen Arbeitsumgebung dar.

Stellaris

Im Rahmen des D-Grid Projektes Astro-Grid-D[28] wurde unter anderem das Metadatenverwaltungssystem Stellaris[29] entwickelt. Das intern verwendete RDF (Resource Description Framework[30]) erlaubt größtmögliche Flexibilität bei der Struktur und der Ablage von Metadaten. Neben den bereitgestellten Suchfunktionalitäten können außerdem gruppenbasierte Zugriffsrechte auf Metadatenkollektionen festgelegt werden. Die Ablage von reinen Forschungsdaten wird jedoch nicht unterstützt, weshalb Stellaris allein auch keine Allgemeinlösung für die Datenhaltung in der virtuellen Arbeitsumgebung darstellt.

OGSA-DAI

Das letzte im D-Grid zur Verfügung gestellte System ist OGSA-DAI[31]. Hierbei handelt es sich um ein Integrationssystem für verschiedenste Datenquellen, hauptsächlich für Datenbanken, mit dem ein homogener Zugriff auf die Datenquellen zur Verfügung gestellt werden kann. Das System selbst kann jedoch in seiner eigentlichen Form keine Daten in Dateiform ablegen oder verwalten und ist daher im Rahmen dieser Expertise nicht relevant[32].

---

[27] Die Projektwebsite ist unter WissGrid - Grid für die Wissenschaft (2010), Letzter Zugriff: 2010.05.27, URL: http://www.wissgrid.de.

[28] D-Grid-Projekt im Rahmen der astrophysikalischen Forschung. Projektwebsite unter: AstroGrid-D (2010), Letzter Zugriff: 2010.05.27, URL: http://www.gac-grid.org.

[29] Details unter: Stellaris (2010), Letzter Zugriff: 2010.05.21, URL: http://stellaris.zib.de.

[30] Beschreibung unter: Resource Description Framework (RDF) (2010), Letzter Zugriff: 2010.05.27, URL: http://www.w3.org/RDF/.

[31] Open Grid Services Architecture Data Access and Integration. Details unter: Open Grid Services Architecture Data Access and Integration (OGSA-DAI) (2010), Letzter Zugriff: 2010.05.25, URL: http://www.ogsadai.org.uk.

[32] Mittels eigner Implementierung wäre eine Dateiverwaltung möglich, jedoch ist der Aufwand ein Datenbanksystem zur Dateiverwaltung zu adaptieren zu umfangreich und wenig wirtschaftlich.

GEODE

Auf Basis der Entwicklungen in DAMES (Data Management through e-Social Science) existiert eine Lösung zur Datenablage inklusive Verwaltung technischer und weiterer Metadaten auf Basis der Grid-Technologie speziell für die sozialwissenschaftliche Forschung[33]. Das zugrunde liegende GEODE (Grid Enabled Occupational Data Environment, siehe http://www.geode.stir.ac.uk) verfolgt dabei das Ziel, Daten verschiedener Berufsgruppen zusammenzutragen und den Wissenschaftlern leicht zugänglich zur Verfügung zu stellen. Der Fokus liegt hierbei auf den entsprechenden Portalen.

GEODE setzt für den verteilt organisierten Datenzugriff auf OGSA-DAI (Open Grid Services Architecture Data Access and Integration, siehe http://www.ogsadai.org.uk). Da OGSA-DAI als eine generische Schnittstelle zu Datenbanksystemen in Grid-Umgebungen eingeordnet werden kann, ist eine Übernahme dieser Lösung weniger ratsam. Eine dateiorientierte Lösung ist aufgrund der Arbeitsweise der sozioökonomischen Berichterstattung sinnvoller.

4.3.3 Datenmanagement Systeme, die mit Grid-Middleware kooperieren

Fedora/iRODS

Eine sinnvolle Lösung steht in Verbindung mit Fedora (Flexible Extensible Digital Object Repository Architecture, siehe http://fedora-commons.org) zur Verfügung. Fedora ist ein auf nahezu beliebige Dateisysteme aufsetzendes Datenmanagement-Repository mit einem modularen Funktionsumfang. Daher kann Fedora an die spezifischen Bedürfnisse der sozioökonomischen Berichterstattung angepasst werden. Insbesondere im Bibliotheksbereich wird Fedora bereits intensiv für die Verwaltung digitaler Daten eingesetzt. Da für Fedora verschiedene internationale Projekte an der direkten Unterstützung von iRODS arbeiten (z.B. ADONIS, siehe http://www.tge-adonis.fr), besteht die Möglichkeit, iRODS als Speicherinfrastruktur für Daten einzurichten und die komplexe Verwaltung von Metadaten, so genannten Objektmodellen, und die Versionsverwaltung mittels Fedora zu realisieren. Hierzu ist anzumerken, dass unabhängig von der Untersuchung in der Expertise auch das Arbeitspaket Langzeitarchivierung in WissGrid die kombinierte Lösung aus iRODS und Fedora zur Umsetzung als Repository für Forschungsdaten verwenden wird[34].

Aus technischer Perspektive kann die Schnittstelle zwischen iRODS und Fedora auf Basis folgender Möglichkeiten implementiert werden:

- FUSE (siehe https://www.irods.org/index.php/iRODS_FUSE)
- WebDAV (siehe https://projects.arcs.org.au/trac/davis/wiki/WikiStart)
- Apache VFS (siehe http://www.omii.ac.uk/wiki/CommonsVFSExtensionsForGrids)
- JAGSA (Java implementation of the Simple API for Grid Applications, siehe http://grid.in2p3.fr/jsaga)

---

[33] Blum, JM; Warner, GC; Jones, SB, et al. (2009): Metadata Creation, Transformation and Discovery for Social Science Data Management: The DAMES Project Infrastructure, 1st Annual European DDI Users Group Meeting, Bonn, Germany, 2009.12.04, Letzter Zugriff: 2010.05.25, URL: http://www.iza.org/eddi09.

[34] Aschenbrenner, A; Enke, H; Fritzsch, B, et al. (2010): WissGrid-Spezifikation: Grid-Repository, Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, D3.5.2, 2010.04.30.

Die Lösung mittels FUSE wurde bereits im Rahmen von WissGrid prototypisch getestet und hat sich als vielversprechend erwiesen, jedoch noch Probleme mit größeren Dateien aufgezeigt. WebDAV, Apache VFS und JAGSA stellen stabile Implementierungen dar, erfordern jedoch noch Anpassungen bzw. Entwicklungen in iRODS und Fedora.

Zur Ablage und Verwaltung von Daten bietet Fedora mehrere Schnittstellen an. Zum einen können Benutzer mit Hilfe einer Web-Oberfläche direkt auf das System zugreifen, zum anderen gibt es Softwareschnittstellen, die anderen Systemen entsprechenden Zugriff erlauben. Für die Ablage der Daten in Fedora ist keine Vorbereitung notwendig, da Fedora hinsichtlich der verwalteten Daten typ- und formatunabhängig arbeitet. Eine Aktualisierung existierender Datensätze ist jederzeit sowohl über die Weboberfläche, als auch über die Softwareschnittstelle möglich. Fedora kann intern so konfiguriert werden, dass es Änderungen an Datensätzen mit protokolliert und die Datensätze selbst versioniert. Nach dem Einspielen der Daten sind diese katalogisiert und über eine Suche wiederauffindbar. Außerdem erlaubt Fedora das Einbinden effizienterer Suchmodule.

Da Fedora zur Datenverwaltung intern auf XML-Basis arbeitet, wird hier ein transparenter und zukunftssicherer Standard verwendet. Des Weiteren bietet Fedora eine Rechteverwaltung mit Unterstützung für XACML/SAML (EXtensible Access Control Markup Language / Security Assertion Markup Language) die sich gegenwärtig als verwendeter Standard in verteilten IT-Infrastrukturen etablieren. Dies gilt entsprechend auch für D-Grid, wobei hierzu Aktivitäten u.a. in dem D-Grid-Projekt Gap-SLC (siehe: http://gap-slc.awi.de) unternommen werden[35].

Schließlich implementiert Fedora das Erstellen von Prüfsummen und die Möglichkeit einer automatisierten Repository-Spiegelung. Letzteres würde die Nutzung mehrerer Rechenzentren erlauben, um einen Single-Point-Of-Failure hinsichtlich Datenhaltung zu vermeiden. Damit adressiert Fedora außerdem den Aspekt der Langzeitverfügbarkeit von Daten.

4.3.4 Metadatenextraktion und Dokumentation

Allgemein

Da die Grid-Anbieter in erster Linie Ressourcen mit Linux-basierten Open Source-Betriebssystemen verwenden, können die für die Linux-Plattform verfügbaren Möglichkeiten zur Extraktion technischer Metadaten verwendet werden.

Datenformaterkennung

Die automatische Extraktion technischer Metadaten setzt die Integration der Erkennung der Datenformate in die virtuelle Forschungsumgebung voraus, da diese Funktionalität nicht per se unterstützt wird. Zudem kann keine Verlässlichkeit bezüglich der Erkennung von Datenformaten auf der Grundlage von Dateiendungen garantiert werden.

MIME-Types

ASCII-basierte Daten können grundsätzlich als solche identifiziert werden, da dieses Datenformat auf einfachem Textinhalt beruht. Die Unterstützung zusätzlicher Formate kann mit Hilfe der Anpassung

[35] Gietz, P und Funk, SE (2010): Nutzung von SLCs und ROBOT-Zertifikaten in TextGrid, D-Grid All-Hands-Meeting 2010, Dresden, Germany, D-Grid, Letzter Zugriff: 2010.05.30, URL: http://www.d-grid.de/fileadmin/user_upload/documents/TextGrid.pdf.

der so genannten MIME-Types (Multipurpose Internet Mail Extensions, siehe http://tools.ietf.org/html/rfc2046) bereitgestellt werden. Mittels der definierten MIME-Types kann der Datentyp (z.B. Text, PDF etc.) automatisch bestimmt werden. Dabei wird der Anfang einer Datei eingelesen und anhand von definierten Merkmalen der Datentyp bestimmt. Somit wird eine von Dateiendungen unabhängige und verlässlichere Verifikation von Datenformaten ermöglicht.

Die automatische Bestimmung der Datenformate der Arbeitsdatensätze, Scientific-Use-Files und Syntaxdateien kann realisiert werden (SPSS, Stata, SAS, R). Beispielsweise unterstützen Systeme zur Langzeitarchivierung wie DSpace (siehe http://www.dspace.org, speziell zu Mime-Types: http://libraries.mit.edu/dspace-mit/build/policies/format.html) bereits die von der sozioökonomischen Berichterstattung verwendeten Statistik-Formate.

JHOVE/JHOVE2

Eine weitere Möglichkeit zur Identifikation, Charakterisierung und auch Validierung von Datenformaten ist JHOVE (JSTOR/Harvard Object Validation Environment, siehe http://hul.harvard.edu/jhove). Zur Erkennung von Datenformaten unterstützt JHOVE bereits in der Standardversion die von der sozioökonomischen Berichterstattung verwendeten Formate ASCII-Text, JPEG und PDF. Mit JHOVE2 (siehe https://confluence.ucop.edu/display/JHOVE2Info/Home) wird die Entwicklung gegenwärtig fortgesetzt, hat aber noch keinen stabilen Status erreicht. Hier liegen jedoch Ende Mai 2010 noch keine relevanten Formate für die sozioökonomische Berichterstattung zur Unterstützung vor.

Metadatenextraktionstools

SPSS: In Bezug auf die Extraktion von Metadaten kann für SPSS z.B. auf das Tool spssread (http://czep.net/data/spssread) zurückgegriffen werden, welches auf der Programmiersprache Perl basiert. Da Perl sich in eine virtuelle Forschungsumgebung integrieren lässt, kann die Unterstützung für SPSS integriert werden. Allerdings handelt es sich bei den extrahierbaren Informationen ausschließlich um technische Metadaten.

PDF: Für PDF-Daten können z.B. der Benutzername des Erzeugers, der Dokumenttitel (nicht zu verwechseln mit dem Dateinamen), Stichwörter und eine Beschreibung innerhalb des Dokuments eingebunden werden. Ebenso wird das Erzeugungsdatum sowie das Datum der letzten Änderung am Dokument hinterlegt. In einer Linux-basierten Infrastruktur steht zur Modifikation und Extraktion dieser Daten das Tool pdftk (http://www.accesspdf.com/pdftk) zur Verfügung. Das Datenformat PDF verwendet den Standard XMP (Extensible Metadata Platform, siehe (http://www.adobe.com/products/xmp/index.html).

JPEG: Bilddaten nach dem JPEG- oder TIFF-Standard können technische Metadaten nach folgenden Standards aufnehmen:

- Exif (Exchangeable image file format,

  siehe http://www.cipa.jp/exifprint/contents_e/01exif_e/ExifDCFsummary_E.pdf)
- XMP (Extensible Metadata Platform)
- Information Interchange Model (IIM) nach dem International Press Telecommunications Council (IPTC, siehe http://www.iptc.org)

Mit diesen Standards können z.B. das Erstellungsdatum, Bildauflösung, Bildgröße sowie eigene Metadatenfelder zur Definition von Suchwörtern oder Beschreibungen direkt mit den Bilddaten verknüpft

werden. Die Extraktion und auch Modifikation kann z.B. über das Tool ExifTool (http://owl.phy.queensu.ca/~phil/exiftool) in eine virtuelle Forschungsumgebung integriert werden.

XLS/XLSX: Das proprietäre Datenformat für die Tabellenkalkulation Microsoft Excel enthält ebenfalls technische Metadaten zu dem Erzeugungsdatum, dem Datum der letzten Änderung am Dokument und dem Benutzernamen der Person, die das Dokument erstellt hat. Diese Metadaten lassen sich in einer Linux-basierten Umgebung mit Hilfe der Open Source Tabellenkalkulation OpenCalc aus OpenOffice.org (http://www.openoffice.org) lesen. Eine weitere Möglichkeit besteht mit fccu-docprop (http://fccu-docprop.sourceforge.net), die jedoch seit 2006 nicht mehr aktiv betreut wird. Da OpenOffice.org i.d.R. interaktiv verwendet wird, ist eine direkte Integration in die virtuelle Forschungsumgebung nicht sinnvoll.

Da das neuere Datenformat XLSX auf XML basiert, ist eine automatisierte Extraktion technischer Metadaten grundsätzlich möglich. Hierfür steht jedoch eine explizite Entwicklung auf der Basis von Linux aus. Die Integration in eine virtuelle Forschungsumgebung über eine Eigenentwicklung ist aufgrund der Verwendung von XML grundsätzlich möglich[36].

Unabhängig von den auf Linux basierenden Tools stehen verschiedene freie Softwarebibliotheken für unterschiedlichste Programmiersprachen, unter anderem für Java, zur Verfügung, mit denen die Metadaten aus den diversen Excelformaten extrahiert werden können. Dadurch wäre es möglich Extraktionstools in die virtuelle Arbeitsumgebung zu integrieren, die auf diesen Bibliotheken aufsetzen. Allerdings ist für diese Integration noch ein gewisser Programmieraufwand gegeben.

ASCII-Text und CSV: Eine Extraktion technischer Metadaten von ASCII-Text und dementsprechend von CSV-Daten ist beschränkt auf Dateisysteminformationen, da Text-Formate aufgrund ihrer einfachen Struktur keine expliziten Metadaten aufnehmen können. Daher ist für Text-Formate die Verwendung einer dedizierten Lösung zur Metadatenverwaltung notwendig. Die Verwendung von Informationen auf der Ebene von einfachen Dateisystemen wie NTFS/ext3 etc. sind in einer verteilten IT-Infrastruktur als Basis für eine virtuelle Forschungsumgebung keine adäquate Lösung. Alternativ oder ergänzend zu einer dedizierten Metadatenverwaltung können definierte Metadaten am Anfang jeder Datei eingefügt werden.

Extraktion fachlicher Metadaten

Für Ergebnisdaten im Datenformat PDF oder JPEG können die fachlichen Metadaten z. T. in den entsprechenden Metadatenfeldern (Stichwörter, Beschreibung) gespeichert werden, um diese zu einem späteren Zeitpunkt wieder automatisiert extrahieren zu können. Hier liegt der Vorteil in der engen Verknüpfung der Metadaten mit den Inhalten selbst. Aufgrund der unterschiedlichen Datenformate und der Notwendigkeit für eine Konsistenz ist eine separate Verwaltung der technischen und fachlichen Metadaten nicht sinnvoll. Deshalb ist eine integrierte Lösung zur Verwaltung beider Metadatenarten notwendig.

---

[36] Das XLSX zugrunde liegende Format wird als Office Open XML bezeichnet. XLSX-Daten sind strukturierte XML-Dateien, die in einer gemeinsamen Datei im komprimierten Datenformat ZIP abgelegt werden. Jede XLSX-Datei basiert auf einer gleichen Grundstruktur. Hierbei befinden sich die technischen Metadaten in: XYZ.XLSX[\docProps\core.xml] und können mittels eines XML-Parsers ausgelesen werden. Der XML-Parser ist in der Lage die Struktur einer XML-Datei zu durchlaufen und es können gezielt Informationen extrahiert werden.

Data Documentation Initiative (DDI)

Die in 2003 etablierte DDI Alliance (Siehe http://www.ddialliance.org) arbeitet an Standards zur Beschreibung von Datensätzen in den Sozialwissenschaften. Daraus entstanden sind die Metadaten-Standards DDI 1, DDI 2 und aktuell DDI 3 (genauer Version 3.1). DDI 1 war weitgehend einfach gehalten und nur für simple Datensätze verwendbar. Mit Einführung von DDI 2 wurde XML als Beschreibungssprache und Unterstützung für aggregierte Daten sowie Tabellen eingeführt. Ebenso wurde in DDI 2 ein Mapping auf den Standard Dublin Core eingeführt. Die DDI Alliance ist sich dabei der Bedeutung von Dublin Core als allgemeiner Metadaten-Standard bewusst. In DDI 3 wurde die Unterstützung für komplexe Daten durch ein Lifecycle-Modell erweitert. Der Lifecycle spricht hierbei alle Ebenen von Datenerhebung bis zur Datenanalyse und Archivierung an. Zusätzlich bietet DDI 3 einen modularen Aufbau der Metadaten an, was im Zusammenhang mit dem Lifecycle-Modell die Nutzung vereinfacht.[37,38]

DDI gilt auch in Deutschland als übergreifender Standard für Metadaten in der sozialwissenschaftlichen Forschung.

Dublin Core

Zur Unterstützung der nachnutzbaren Forschungsdatenablage wird eine flexible Lösung zur Abbildung von Metadaten benötigt. Ein grundlegender Ansatz zur Gestaltung von Metadaten existiert auf Grundlage des so genannten Dublin Core Standards (Dublin Core Metadata Initiative, siehe http://dublincore.org). Dublin Core wird vornehmlich im Bibliotheksumfeld eingesetzt. Dieser Standard unterscheidet einerseits technische und fachliche Metadaten und unterstützt andererseits bereits elementare Metadateninformationen wie z.B. einen Unique Identifier. Ebenso werden Metadaten zur Provenienz und zu Berechtigungen in Dublin Core direkt unterstützt. Zur möglichen Verknüpfung von Arbeitsdaten und Syntaxdateien in der Forschungsdatenablage können außerdem die Beziehungen von Daten berücksichtigt werden. Da auch die Formate Open Document Format (ODF) und HTML auf Dublin Core als Metadatenstandard aufbauen, kann die zukünftige Entwicklung bzw. Verwendung als gegeben angenommen werden. Für die sozialwissenschaftliche Community von Bedeutung ist die Möglichkeit, DDI-Metadaten in Dublin Core zu überführen (siehe vorigen Abschnitt zu DDI). Daher ist es sinnvoll Dublin Core als Basis für das Metadatenmodell der sozioökonomischen Berichterstattung zu verwenden.

Fedora

Fedora unterstützt die oben beschriebenen MIME-Types und kann entsprechend Datenformate bestimmen. Im Rahmen des Projekts eSciDoc (Siehe https://www.escidoc.org) wird eine Integration der Extraktion technischer Metadaten auf Basis von JHOVE angestrebt. Die offene Architektur von Fedora bietet die Möglichkeit, die aufgeführten Lösungen zur Metadatenextraktion zu implementieren. Bezüglich der Integration der fachlichen Metadatenextraktion ist eine fachspezifische Anpassung der benötigten Metadaten sowie der Unterstützung im Rahmen der virtuellen Forschungsumgebung notwendig. Fedora setzt in der Grundversion für Metadaten die Verwendung des Dublin Core Standards

---

[37] Green, A (2008): Data Documentation Initiative - DDI, University of Edinburgh, UK, Letzter Zugriff: 2010.07.26, URL: http://www.disc-uk.org/docs/DDI_Green.pdf.

[38] Martinez, L (2008): The Data Documentation Initiative (DDI) and Institutional Repositories, DISC-UK, Letzter Zugriff: 2010.07.26, URL: http://www.disc-uk.org/docs/DDI_and_IRs.pdf.

voraus. In der Grundversion sind andere Standards aktuell nicht vorgesehen. Dublin Core lässt sich jedoch an individuelle Bedürfnisse anpassen. Darüber hinaus lässt sich Fedora mittels Modulen grundsätzlich auch um weitere Metadaten-Standards erweitern.

Integrationsempfehlung

Die notwendige Verwendung von Dublin Core zur Verwaltung von Metadaten ist aufgrund der weit gehenden Anpassungsmöglichkeiten von Dublin Core kein Nachteil von Fedora. Aufgrund der Bedeutung, den der Metadaten-Standard DDI für die sozialwissenschaftliche Forschung hat, ist eine Integration von DDI notwendig. Dies kann durch Überführung von Daten aus DDI in Dublin Core realisiert werden. Des Weiteren werden von Fedora bezüglich Metadatenextraktion und Dokumentation generell alle Anforderungen der sozioökonomischen Berichterstattung erfüllt[39]. Eine Erweiterung der Metadatenkategorien an die fachlichen Bedürfnisse wird ebenso durch die Erweiterbarkeit von Dublin Core unterstützt. Die empfohlene Integration technischer und fachlicher Metadaten lässt sich mit Fedora realisieren. Aufgrund der Erfüllung der gestellten Anforderungen eignet sich Fedora als mögliche Lösung zur Verarbeitung von Metadaten für die sozioökonomische Berichterstattung.

4.3.5 Versionsverwaltungssysteme

Eine Versionsverwaltung von Syntaxdateien bietet der sozioökonomischen Berichterstattung den strukturierten Umgang mit unterschiedlichen Entwicklungsstufen der Syntaxdateien im Forschungsprozess. Durch die Integration in die virtuelle Forschungsumgebung besteht für den Wissenschaftler die nahtlose Nutzung dieser Funktionalität.

SVN/CVS

Grundsätzlich werden so genannte Revision Control Systeme in der professionellen Softwareentwicklung eingesetzt, um verschiedene Entwicklungszweige einer Software und deren jeweilige Unterversionen integriert und konsistent verwalten zu können. Hierzu existiert eine Reihe von Implementierungen im Open Source-Bereich. Die bekanntesten Systeme sind CVS (Concurrent Versions System, siehe http://www.nongnu.org/cvs) und das neuere SVN (Apache Subversion, siehe http://subversion.apache.org/). Beide Systeme setzen auf einen Client-Server-Ansatz, wobei das darunter liegende Dateisystem zur eigentlichen Datenverwaltung nicht explizit festgelegt ist, sondern vielmehr von der verwendeten Betriebssystemplattform abhängt. Dabei ist festzustellen, dass sich SVN in der Zwischenzeit als Lösung zur Versionsverwaltung am Markt verstärkt durchsetzt[40]. Dementsprechend wird CVS nicht weiter in die Betrachtung einbezogen.

Kernaspekte der Versionsverwaltung mittels SVN und verwandter Systeme liegt in den Möglichkeiten, Dateikopien auf Basis eines identischen Bearbeitungsverlaufs und verschiedene Stati von Softwarequellcode verwalten zu können. Die drei möglichen Stati folgen dem so genannten Tag- und Branch-Konzept:

---

[39] The Fedora Development Team (2008): Fedora Tutorial #1 - Introduction to Fedora, Fedora Commons, 2008.07.23, Letzter Zugriff: 2010.05.26, URL: http://fedora-commons.org/confluence/download/attachments/4718930/tutorial1.pdf?version=1&modificationDate=1218459761506.

[40] Schwaber, C; Gilpin, M und Stone, J (2007): Software Change And Configuration Management, The Forrester Wave™, Q2 2007, Q2 2007, Forrester Research, I, 2007.07.29, Q2 2007, http://www.collab.net/forrester_wave_report/index.html.

1. Trunk – der zentrale Hauptentwicklungspfad
2. Branch – eine Variante, die zusätzlich zum Hauptentwicklungspfad gepflegt wird
3. Tag – eine Version, deren Status unverändert bleibt (kann entweder von dem Hauptentwicklungspfad oder einer Variante erzeugt werden)

Damit ist die Möglichkeit gegeben, auch sehr komplexe Softwaresysteme mit vielen tausend Zeilen Quellcode effektiv verwalten zu können. Mittels verschiedener Varianten können in der Softwareentwicklung sehr spezifische Kundenwünsche in Systeme integriert werden. Des Weiteren können alte Entwicklungsstände weiter gepflegt werden, in dem Fehlerkorrekturen auch für den älteren Entwicklungsstand bereitgestellt werden.

SVN wird durch das Client-Server-Prinzip primär in verteilten Arbeitsgruppen mit zentraler Datenablage verwendet. Eine explizite Verwendung von SVN zur Versionsverwaltung von Daten in verteilten Umgebungen wie der Grid-Technologie ist nicht vorgesehen. Dies liegt vor allem daran, dass der Fokus auf dem Bereich der Unterstützung von Softwareentwicklung liegt. Analog findet Versionsverwaltung der Softwareentwicklung in D-Grid via SVN separat von der eigentlichen Grid-Infrastruktur statt.

Bezüglich der Bedarfe der sozioökonomischen Berichterstattung ist festzustellen, dass eine Versionsverwaltung von Syntaxdateien einerseits notwendig ist, andererseits jedoch nicht in dem Ausmaß, der in der Softwareentwicklung von Bedeutung ist. Aufgrund des explorativen Forschungscharakters der sozioökonomischen Berichterstattung werden verschiedene Entwicklungszweige die Regel darstellen, da die einzelnen Forscher bzw. kleinere Arbeitsgruppen i.d.R. dedizierte Aufgabenstellungen bearbeiten. Die dabei durchgeführten Entwicklungen benötigen keine komplexe Versionsverwaltung, da sie in einer oder wenigen kleinen Dateien erstellt werden. Eine Zusammenführung der Ergebnisse erfolgt letztlich in einem gemeinsamen Schritt zu einem Gesamtergebnis, wobei dies davon abhängt, ob ein gemeinsames Ergebnis das Ziel der Forschungsfragestellung darstellt.

Relevant sind die gute wissenschaftliche Praxis und die allgemeine Nachvollziehbarkeit der Forschungsarbeiten in der sozioökonomischen Berichterstattung. Dazu wird der Versionsstand von Syntaxdateien benötigt, der einer Publikation oder einem Ergebnis zugrunde liegt. Somit benötigt die sozioökonomische Berichterstattung in erster Linie keine effiziente Quellcodeverwaltung wie SVN, sondern vielmehr eine Lösung, die verschiedene einfache Versionsstände von Syntaxdateien verwalten kann. Ebenso muss sich eine Versionsverwaltung für die sozioökonomische Berichterstattung möglichst harmonisch in die virtuelle Forschungsumgebung einfügen, um funktionelle Brüche zu vermeiden.

Fedora

Fedora bietet die Unterstützung einer Versionsverwaltung für jedes abgelegte digitale Objekt. Dies wäre auch entsprechend für die sozioökonomische Berichterstattung hinsichtlich Arbeitsdaten, Scientific-Use-Files und Syntaxdateien möglich. Dabei werden jedoch in der Versionsverwaltung selbst keine expliziten Versionsstände berücksichtigt, sondern vielmehr das Speicherungsdatum verwendet. Darüber hinaus wird beim Ablegen bzw. Aktualisieren von digitalen Objekten die jeweilige Nutzerzuordnung in die Versionsinformationen einbezogen. Generell wird eine Historie zu allen Änderungen (Speicherung, Modifikation, Löschung etc.) an den gespeicherten digitalen Objekten geführt.

Ein weiteres Feature von Fedora ist die Möglichkeit, Beziehungen zwischen den gespeicherten Objekten im Kontext verschiedener Versionen zu berücksichtigen. Damit können für die Arbeiten in der sozioökonomischen Berichterstattung Verknüpfungen zwischen Arbeitsdaten und Syntaxdateien hergestellt werden. Dadurch wird der Forschungsprozess nicht nur für den einzelnen Forscher, sondern

auch in Forschungsgruppen sinnvoll unterstützt. Welche Arbeitsdatenversion im Zusammenhang mit welcher Syntaxdateiversion verwendet wurde, kann somit leicht nachvollzogen werden. Für die Arbeit mit Syntaxdateien sind jedoch weitere unterstützende Funktionen wie Versionsvergleich, und Editoren mit Syntax-Highlighting sinnvoll zu integrieren.

Ein letzter wichtiger Aspekt ist, dass Scientific-Use-Files und die gegebenenfalls darauf basierenden Arbeitsdaten beim Ablauf der vorgegebenen Nutzungszeiträume vollständig gelöscht werden müssen. Dies gilt sowohl für die aktuellen Versionen, als auch für ältere Versionen der Datensätze. Fedora unterstützt hier, im Gegensatz zu vielen Versionsverwaltungssystemen, ein dezidiertes Löschen auch älterer Versionen von abgelegten Objekten.

4.3.6 Konvertierung und Validierung

Für die Konvertierung von Statistikdaten in unterschiedliche Formate bieten die meisten Statistik-Programme entsprechende Routinen. So erlaubt SPSS z.B. einen Datensatz aus dem proprietären Format nach CSV umzuwandeln. CVS kann nun wiederum genutzt werden um die Daten in andere Statistik-Programme oder Microsoft Excel zu importieren. Ein Beispiel hierzu ist im Rahmen der Dokumentation von GEODE[41] zu finden[42], einem Projekt, welches sich mit Technologien für die Distribution berufsbezogener Statistikdaten beschäftigt. Die Projektbeteiligten von GEODE stellten fest, dass entsprechende Konvertierungsroutinen benötigt werden und erwogen im Jahr 2007 eine eigene Implementierung, um von den Statistik-Programmen unabhängig zu sein. Über den aktuellen Stand dieses Vorhabens ist jedoch nichts bekannt.

Für bestimmte Umwandlungen sind im Internet diverse Tools zu finden. Ein Beispiel ist das Apache POI Project[43], eine Java API für den Zugriff auf das Microsoft Excel Datei Format. Diese könnte genutzt werden, um z.B. eine Implementierung zur Konvertierung zwischen CSV und Excel Dateien vorzunehmen[44]. Mittlerweile gibt es auch Tools für den lesenden und schreibenden Zugriff auf SPSS Dateien[45], die ebenfalls auf der Programmiersprache Java basieren. Auch hiermit wäre eine eigene Implementierung von Konvertierungstools möglich. Allerdings sind derartige Tools nicht immer kostenfrei nutzbar. Ihr Einsatz innerhalb einer virtuellen Arbeitsumgebung hängt daher vom jeweiligen Lizenzierungsmodell ab.

Auch wenn entsprechende Werkzeuge entweder implementiert oder in die virtuelle Arbeitsumgebung integriert werden können, so ist von der Anwendung dieser Tools für reguläre Arbeitsschritte abzuraten. Konvertierungen zwischen Datenformaten sind meist mit Informationsverlust verbunden. So gehen z.B. die Metadaten einer Excel-Datei bei der Umwandlung nach CSV vollständig verloren. Au-

---

[41] Dokumentation zu dem Projekt ist unter: GEODE: Grid Enabled Occupational Data Environment (2010), Letzter Zugriff: 2010.06.09, URL: http://www.geode.stir.ac.uk/index.html.

[42] Siehe hierzu http://www.geode.stir.ac.uk/file_convert_info.html. Es wird beschrieben, wie SPSS und Stata für die Umwandlung ihrer jeweils eigenen Formate in CSV und zurück eingesetzt werden können, da GEODE intern mit CSV arbeitet.

[43] Informationen unter: POI-HSSF and POI-XSSF - Java API To Access Microsoft Excel Format Files (2010), Letzter Zugriff: 2010.06.09, URL: http://poi.apache.org/spreadsheet/index.html.

[44] Ein einfaches Beispiel hierzu ist unter http://www.roseindia.net/answers/viewanswers/2707.html zu finden. Letzter Zugriff: 09.06.2010.

[45] Siehe SPSS Writer (2010), Letzter Zugriff: 2010.06.09, URL: http://spss.pmstation.com/.

ßerdem können Wertebereiche von Daten in einem neuen Format weniger genau sein, was ebenfalls einen Informationsverlust nach sich zieht. Die Projekte, in denen die virtuelle Arbeitsumgebung gegebenenfalls Anwendung findet, sollten sich daher bereits bei Projektbeginn auf wenige einheitlich verwendete Datenformate einigen um notwendige Konvertierungen zur Projektlaufzeit zu vermeiden.

Für die Validierung von Datenformaten kann unter anderem das bereits benannte Tool JHOVE[46] eingesetzt werden. Allerdings unterstützt JHOVE von Haus aus die in der sozioökonomischen Berichterstattung verwendeten Formate nicht. Es kann jedoch um entsprechende Routinen erweitert werden, sofern die Datenformatstrukturen bekannt sind. Ein Beispiel für eine einfache Validierung von SPSS Dateien mit JHOVE ist unter http://develop01.dans.knaw.nl/svn/mixed/mixed-open-source/trunk/mixed-framework/file-type-detector/src/main/java/nl/knaw/dans/mixed/framework/filetypedetector/SPSSModule.java zu finden.

4.3.7 Sicherheitslösungen im Grid

Public Key Infrastructure

Um eine eindeutige Identifikation und vertrauliche Kommunikation zu realisieren, setzen Grid-Infrastrukturen auf die so genannte Public Key Infrastructure (PKI)[47]. Die Verwendung von PKI ist ein international anerkannter Standard in wissenschaftlichen und kommerziellen verteilten IT-Infrastrukturen.

Eine PKI basiert darauf, dass jeder Nutzer ein auf seine Person nach einem vordefinierten Standard ausgestelltes Schlüsselpaar, bestehend aus einem privaten und einem öffentlichen Schlüssel, verwendet. Da diese Schlüssel nicht identisch sind, bezeichnet man dieses Verfahren auch als asymmetrisches Kryptosystem. Es existieren zwei grundlegende Anwendungsszenarien: Verschlüsselung und Signatur.

Die erste Nutzungsmöglichkeit der Verschlüsselung erlaubt anderen Personen dem Schlüsselinhaber Nachrichten zu senden, die mit seinem öffentlichen Schlüssel verschlüsselt worden sind. Der öffentliche Schlüssel kann dazu z.B. auf einer Webseite anderen Nutzern zur Verfügung gestellt werden. Nur der Schlüsselinhaber kann mit seinem privaten Schlüssel die verschlüsselte Nachricht wieder dekodieren und damit den Inhalt einsehen.

Die zweite Nutzungsmöglichkeit der Signatur dient der Authentifizierung und Autorisierung von Nutzern in einer PKI. Zur Authentifizierung sendet ein Nutzer ein so genanntes Zertifikat[48] an die Instanz, die den Nutzer anhand des Zertifikats authentifiziert und damit die angegebene Identität verifiziert. Ein Zertifikat wird von einer Zertifizierungsstelle i.d.R. gegen persönliche Authentifizierung des Nutzers vor Ort ausgestellt. Ein solches Zertifikat enthält u.a.

- einen Distinguished Name, welcher den Nutzer eindeutig identifiziert[49],
- die Gültigkeitsdauer[50] des Zertifikats

---

[46] JHOVE - JSTOR/Harvard Object Validation Environment (2010), Letzter Zugriff: 2010.06.09, URL: http://hul.harvard.edu/jhove/.

[47] Chakrabarti, A (2007): Grid Computing Security, Springer, Berlin; Heidelberg; New York.

[48] Gegenwärtig hat sich für Zertifikate X.509 als Standard etabliert.

[49] Bei Personen wird i.d.R. der Name, bei Maschinen i.d.R. der Rechnername verwendet.

[50] In D-Grid ist die Gültigkeitsdauer für Nutzer und Maschinen 1 Jahr.

- den öffentlichen Schlüssel des Nutzers
- eine Signatur des Zertifikatsausstellers (Zertifizierungsstelle, CA)

Wird das Zertifikat zur Authentifizierung von dem Nutzer verwendet, so prüft die authentifizierende Gegenstelle anhand der Signatur und des öffentlichen Schlüssels, ob das Zertifikat gültig ist. Mittels der Gültigkeitsdauer wird die Verwendung eines Zertifikats zeitlich beschränkt.

Da es i.d.R. eine Hierarchie von Zertifizierungsstellen gibt, unterscheidet man in die Certification Authority (CA), die das so genannte Stammzertifikat[51] ausstellt. Des Weiteren werden so genannte Registration Authorities (RA) aufgebaut, um die Identifikation der Nutzer bei der Ausstellung von Zertifikaten vor Ort zu gewährleisten. Dabei entsteht eine Vertrauenskette, wobei eine Instanz der jeweils übergeordneten Instanz vertraut. Bei der Verifikation eines Zertifikats kommt zusätzlich die Validation Authority (VA) zum Einsatz. Die VA erhält die Anfragen zur Verifikation und bestätigt die Echtheit eines Zertifikats. Abbildung 4 gibt einen Überblick zu der Funktionsweise einer PKI.

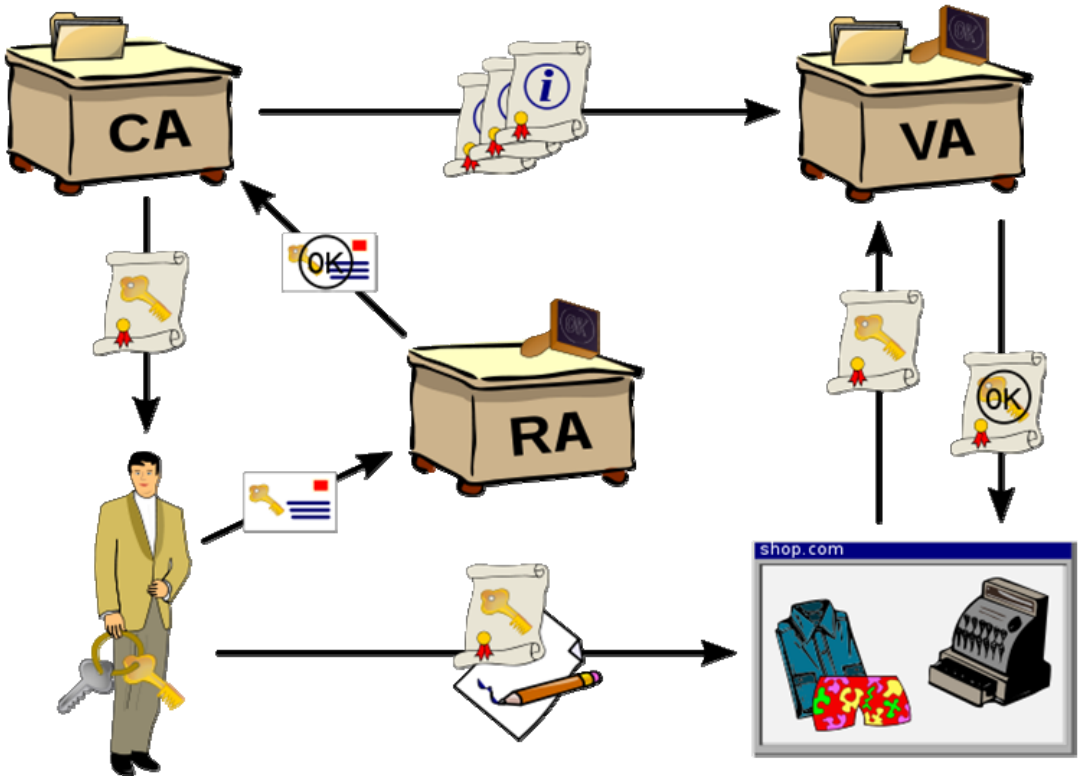

Abbildung 4 - Übersicht der Funktionsweise einer PKI[52]

[51] Die CA hält das Stammzertifikat. Alle untergeordneten Instanzen müssen der CA vertrauen. Das Vertrauen in der PKI stützt sich i.d.R. auf das in die Betreiberorganisation gesetzte Vertrauen.

[52] Quelle: Wikimedia Commons (2010): Principle of a public key infrastructure, Letzter Zugriff: 2010.06.06, URL: http://upload.wikimedia.org/wikipedia/commons/3/34/Public-Key-Infrastructure.svg.

Grid Security Infrastructure

Im Allgemeinen werden Sicherheitsimplementierungen in übergreifenden IT-Infrastrukturen als Authentication and Authorization Infrastructure (AAI) bezeichnet. Die in D-Grid und anderen internationalen Grid-Infrastrukturen eingesetzte AAI ist die Grid Security Infrastructure (GSI). GSI baut auf einer vorhandenen PKI auf. In der Grid-PKI treten an die Stelle der VA in Abbildung 4 die von den Communitys betriebenen virtuellen Organisationen, die über geeignete Dienste die Zertifikate gegenüber den Ressourcen-Providern validieren (siehe nächster Abschnitt).

Mittels GSI kann die notwendige Sicherheit für die Authentifizierung der *soeb*-Wissenschaftler in einer virtuellen Forschungsumgebung auf Basis der Grid-Technologie realisiert werden. Die in D-Grid verwendete Public Key Infrastructure wird vom Deutschen Forschungsnetz (DFN) und vom Forschungszentrum Karlsruhe (GridKA) als Certification Authorities (CA) betrieben. Beide sind EU-weit von EUGridPMA (http://www.eugridpma.org) akzeptiert. EUGridPMA[53] ist die europäische Organisation, um Vertrauensstellungen für PKI-basierte Authentifizierungsverfahren in eScience- bzw. Grid-Infrastrukturen sicherzustellen. Änderungen an der Funktionalität oder Erweiterungen an der PKI des D-Grid werden grundsätzlich zwischen DFN, GridKA und der EUGridPMA[54] abgestimmt.

Virtuelle Organisationen und Rechteverwaltung

Die Nutzer eines Grids kommen üblicherweise aus verschiedenen Organisationen, die weitgehend geographisch voneinander getrennt sind. Genau das macht den Charakter eines Grids aus. Allerdings wird in Forschungsverbünden oft eine einrichtungsübergreifende Zusammenarbeit notwendig. Es entstehen dabei so genannte virtuelle Organisationen, die einen organisationsübergreifenden Arbeitszusammenhang im Grid abbilden. Eine virtuelle Organisation (VO) kann jedoch auch identisch mit einer existierenden Organisation sein.

Eine VO umfasst eine Reihe von Personen, die das Grid nutzen, um ein gemeinsames Ziel zu erreichen. Innerhalb einer solchen Organisation nehmen die Personen unterschiedlichste Aufgaben, Positionen und damit Rollen ein. Außerdem werden die virtuellen Organisationen (VOs) nicht selten in einzelne Gruppen aufgeteilt. Der Forschungsverbund der sozioökonomischen Berichterstattung wäre zum Beispiel eine solche VO. Die (Unter-)Gruppen der VO können dabei aus den Mitarbeitern der jeweils beteiligten Forschungsinstitutionen, oder aus thematischen Arbeitsgruppen gebildet werden.[55] Die Ausgestaltung des VO Konzepts ist von den Gegebenheiten der Community abhängig.

Im D-Grid wurde bereits eine Infrastruktur etabliert, mit der es möglich ist, VOs mit deren Mitarbeitern, Gruppen (und Rollen) zu modellieren, entsprechende Rechte zu vergeben und diese Rechte bei jeder Aktion eines Benutzers im Grid zu forcieren. Neben diesen technischen Aspekten wurden auch Vorgehensweisen etabliert, die die Verwaltung solcher Organisationen strukturiert und vereinheitlicht

---

[53] European Policy Management Authority for Grid Authentication in e-Science.

[54] EUGridPMA führt Abstimmungen auf internationaler Ebene mit Asia Pacific Grid Policy Management Authority (APGrid PMA), The Americas Grid Policy Management Authority (TAGPMA) und TERENA Academic CA Repository (TACAR) durch. TERENA ist The Trans-European Research and Education Networking Association mit Sitz in den Niederlanden.

[55] Je nach Anforderungen und Gegebenheiten in der VO können die Mitglieder der VO auch verschiedene Rollen einnehmen, wie z.B. die Rolle eines Datenhalters, die Rolle eines Datenauswerters, und so weiter. Es gibt auch VO-Managementsysteme, die einen Rollen-basierten Zugriff auf Ressourcen ermöglichen.

ablaufen lassen. Die Hauptarbeiten zu diesen beiden Aspekten wurden im ersten D-Grid Integrationsprojekt (DGI) Fachgebiet 1[56] sowie im Projekt IVOM[57] geleistet.

Das Rahmenkonzept[58] für die Verwaltung von VOs und andere Dokumente des DGI[59] beschreiben, wie eine neue VO angelegt werden kann und wie sie danach verwaltet wird. Eine zentrale Rolle spielt dabei der so genannte VO-Manager, eine (oder mehrere) Person(en) zur Verwaltung einer VO. Dieser hat neben Vertragabschlüssen mit den Ressourcenanbietern außerdem die Aufgabe die Mitglieder- und Rechteverwaltung innerhalb der VO zu übernehmen. Dafür stehen ihm diverse Tools zur Verfügung, auf die hier nur oberflächlich eingegangen werden soll. Details dazu sind auf den Webseiten des D-Grid (http://www.d-grid.de) zu finden.

Zur Verwaltung virtueller Organisationen wird der so genannte Virtual Organisation Membership Registration Services (VOMRS) verwendet[60]. Hierüber hat der VO-Manager die Möglichkeit, Benutzer einzelnen Gruppen (und Rollen) zuzuordnen. Die Rechte der VO-Mitglieder und -Gruppen auf den Ressourcen sind durch Vereinbarung mit den Ressourcen-Anbietern festgelegt. Die Anbieter beziehen von den VOMRS die Informationen über die zugelassenen Mitglieder und Gruppen und bilden diese in den lokalen Nutzerberechtigungen ab. Diese Informationen werden täglich neu vom VOMRS bezogen.

Der einzelne Grid-Nutzer erzeugt mit Hilfe seines Zertifikates über die GSI ein zeitlich begrenzt gültiges Proxy-Zertifikat, welches von der Grid-Middleware als Basis zur Authentifizierung für die Nutzung der Ressourcen verwendet wird. Auf den einzelnen Grid-Ressourcen werden die durch die Proxy-Zertifikate repräsentierten Benutzer dann auf lokale Betriebssystembenutzer abgebildet. Das D-Grid hat hier die Policy, dass jedes Nutzerzertifikat auf einen lokalen Nutzer abgebildet wird.[61]

Eine vereinfachte Darstellung des Sicherheitskonzepts im D-Grid ist in Abbildung 5 dargestellt.

---

[56] Sie hierzu: DGI-1 - FG 1: D-Grid-Basis-Software - Beschreibung (2008), URL: http://dgi2.d-grid.de/index.php?id=24.

[57] Interoperabilität und Integration der VO-Management Technologien im D-Grid (2010), Letzter Zugriff: 2010.06.30, URL: http://www.d-grid.de/index.php?id=314&L=0.

[58] Milke, J-M; Schiffers, M und Ziegler, W (2008): Rahmenkonzept für das Management Virtueller Organisationen im D-Grid, Version 1.1, Letzter Zugriff: 2010.06.30, URL: http://www.d-grid.de/fileadmin/user_upload/documents/DGI-FG1-10/VO_Rahmenkonzept-final.pdf.

[59] Z.B. Grimm, C; Henne, B; Piger, S, et al. (2008): Generische VO-Strukturen für das D-Grid, D-Grid Integrationsprojekt 2 (DGI-2) Fachgebiet 3.2 AAI/VO, Letzter Zugriff: 2010.06.30, URL: http://dgi.d-grid.de/fileadmin/user_upload/documents/DGI2-FG3/FG3-2/DGI-2_FG-3.2_Generische_VO-Strukturen_D-Grid.pdf.

[60] In der Hochenergiephysik, mit der Grid Middleware gLite, wird der Virtual Organization Membership Service (VOMS) eingesetzt, der auch einen rollenbasierten Zugang zu Grid-Ressourcen implementiert. Im D-Grid bezieht der VOMS von den VOMRS-Servern die Informationen.

[61] D.h. es wird eine One-to-One Relation implementiert. Hier ist auch der wesentliche Unterschied zu einem rollenbasierten System, in dem viele Nutzer auf eine Rolle, z.B. „Daten-Bearbeiter“, lokal abgebildet werden (Many-to-One).

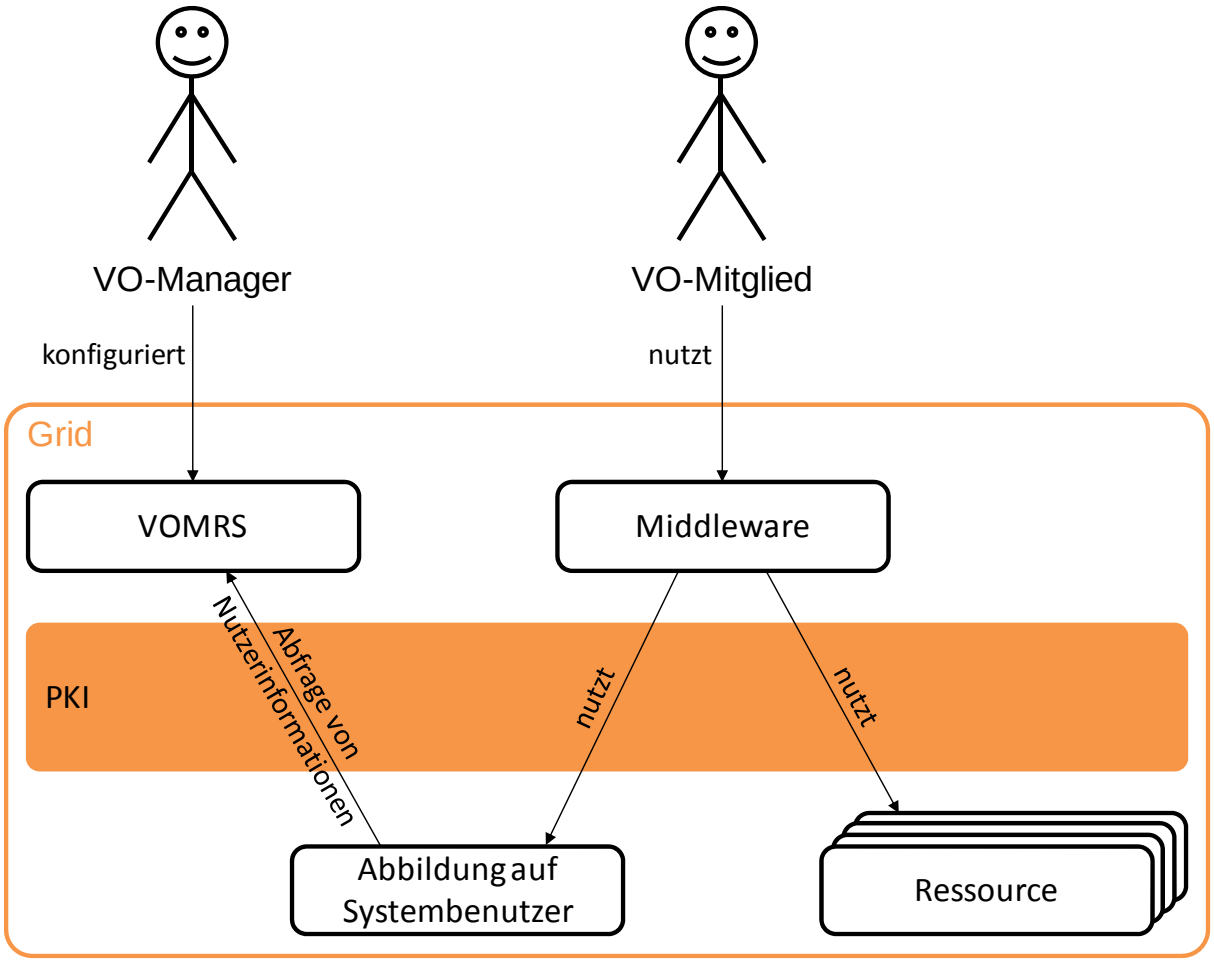

Abbildung 5 - Stark vereinfacht dargestelltes Sicherheitskonzept im D-Grid

Datensicherheit durch Verschlüsselung unter Verwendung von Parrot

Auf dem vierten D-Grid Security Workshop am 15. und 16. Oktober 2009 wurde ein Verfahren vorgestellt, mit dem eine Verschlüsselung von Dateien bei der Ablage und dem Zugriff im Grid möglich ist[62]. Herausragend war dabei, dass die auf die Dateien zugreifenden Programme nicht angepasst werden müssen. Das Verfahren beruht unter anderem auf dem Einsatz von Parrot[63], einer Software, mit der Dateizugriffe eines Programms abgefangen und ersetzt werden können. Bei jedem Dateizugriff wird, wie im Vortrag dargelegt, der Inhalt beim Lesen der Datei entschlüsselt, bzw. beim Schreiben in die Datei verschlüsselt.

Der Vorteil dieses Verfahrens ist die transparent-sichere Datenablage. Ohne den jeweiligen Schlüssel ist ein Lesen und Schreiben der Dateien nicht möglich. Das Verfahren beinhaltet jedoch noch keine endgültige Lösung dafür, wie das benötigte Schlüsselpaar sicher in das Grid übertragen werden kann. Es gab allerdings erste Ideen für eine Lösung.

---

[62] Vortrag von Steffen Limmer, Department of Computer Science, Universität Erlangen. „Transparente Dateiverschlüsselung mit Parrot und dem Secure Storage Service“. Leider keine Quellenangabe möglich.

[63] Thain, D und Livny, M (2005): Parrot: An Application Environment for Data-Intensive Computing, Scalable Computing: Practice and Experience (Band 6), Nr. 3, Seite 9-18. URL: http://www.cse.nd.edu/~dthain/ papers/parrot-scpe.pdf.

Es ist nicht bekannt, in wie weit das Verfahren weiter entwickelt wird. Der Ansatz verspricht jedoch höchste Sicherheit bei der Ablage von Dateien, nicht nur im Grid. Er kann bei der Implementierung der virtuellen Arbeitsumgebung Anwendung finden, wenn die Datensicherheitsbestimmungen diese extreme Form der Datensicherung verlangen sollten.

4.4 Datenverarbeitung

Der zentrale Arbeitsplatz zur Datenverarbeitung in der sozialwissenschaftlichen Forschung ist die Workstation der Wissenschaftler. Hier werden alle Datenverarbeitungsaufgaben, ob lokal oder entfernt ausgeführt, angestoßen. Der Einsatz einer virtuellen Arbeitsumgebung ändert nichts an dieser Arbeitsweise, so dass die Workstation weiterhin als zentrales Arbeitsinstrument bestehen bleibt. Durch die virtuelle Arbeitsumgebung werden vielmehr die Möglichkeiten der einzelnen Workstations im Hinblick auf die Datenverarbeitung wesentlich erweitert. Bereits vorhandene Funktionen soll die virtuelle Forschungsumgebung nicht ersetzen, sondern integrieren. Dies gilt insbesondere für die eingesetzten Lösungen zur Datenverarbeitung.

Die virtuelle Arbeitsumgebung erweitert die Datenverarbeitung einerseits durch den umfassenden Zugang zu Forschungsdaten und Syntaxdateien sowie um die Möglichkeit zur Nutzung integrierter Statistik-Lösungen.

Die in den Anfangskapiteln beschriebenen Arbeitsschritte der sozialwissenschaftlichen Forschung werden mit Hilfe von Statistik-Programmen, wie Stata, SPSS oder SAS durchgeführt. Deren Verwendung ist meist mit substantiellen Lizenzgebühren verbunden. Die Lizenzbedingungen der Anbieter sehen den Einsatz der Produkte in der verteilten Rechnerinfrastruktur einer virtuellen Arbeitsumgebung nur begrenzt oder gar nicht vor, und bleiben Gegenstand weiterer Verhandlungen (vgl. Abschnitt 3.3). Die virtuelle Arbeitsumgebung kann daher hauptsächlich mit dem Open Source Statistik-Programm R weitergehende Aufbau- bzw. Integrationsarbeiten beginnen, um im Bedarfsfalle („On Demand“) auch mehr Rechenkapazität zur Verfügung zu stellen. Hierauf aufbauend können auch weitere Programme, wenn entsprechende Lizensierungen möglich sind, auf entfernten Rechnern installiert und in die virtuelle Arbeitsumgebung integriert werden.

4.4.1 Zugang zu Forschungsdaten und Syntaxdateien

Einer der wesentlichen Vorteile, unmittelbar für jeden Nutzer zugänglich, wäre eine strukturierte Datenablage, sowohl von Forschungsdaten als auch Syntaxdateien in der virtuellen Arbeitsumgebung. Entsprechende Berechtigungen vorausgesetzt, erweitern Verfügbarkeit und Erreichbarkeit aller für die eigene Arbeit benötigten Daten über ein zentrales Daten-Repository ganz entschieden die Arbeitsmöglichkeiten des einzelnen Forschers (Vgl. Abbildung 8). Dies gilt unabhängig davon, ob man am eigenen Arbeitsplatz oder unterwegs mit geeignetem Netzwerkzugang arbeitet. Die über das Repository mögliche gemeinsame Nutzung der Daten strukturiert und erweitert die Möglichkeiten der Zusammenarbeit durch einfachere Wege des Datenaustausches erheblich, wie auch z.B. das Nutzen von Programmen[64], die nicht per se überall verfügbar sind.

[64] Die Einhaltung von Lizenzbedingungen wird als gegeben vorausgesetzt.

4.4.2 Integration von R

Nutzungsmöglichkeiten von R

R als Softwarepaket ist modular aufgebaut und mit so genannten Paketen erweiterbar, wobei z.T. das Problem besteht, dass bei Nutzung von zusätzlichen Paketen Funktionen mit identischem Namen eingebunden werden können. Dadurch werden Ursprungsfunktionen überschrieben und die Installation arbeitet nicht mehr planmäßig bzw. es steht nicht die erwartete Funktionalität zur Verfügung[65]. Daher ist es ratsam, für die Nutzung von sekundären oder selbst entwickelten Paketen separate R-Installationen zu betreiben. Die R-Entwicklergemeinde hat dieses Problem bereits erkannt und entsprechende Maßnahmen für die Modulentwicklung ergriffen. Dementsprechend sollten zukünftig entwickelte Pakete als „überschreibungsfrei" gelten. Zur Sicherheit sollte jedoch auf absehbare Zeit die Variante der Multi-Installation aufgegriffen werden. Dies ist relativ einfach möglich, da die jeweiligen Installationen nur in separaten Verzeichnissen abgelegt werden müssen.

Derzeitig wird R noch nicht in der ganzen Fachgemeinschaft der sozialwissenschaftlichen Forschung eingesetzt, jedoch wächst die Zahl der Anwender[66,67,68].

Integration in eine virtuelle Forschungsumgebung

In einer virtuellen Forschungsumgebung lässt sich die Statistik-Software R einfach integrieren, da sich die Syntaxdateien und Arbeitsdaten auf eine theoretisch beliebige Rechnerressource transferieren und mittels einer dortigen R-Installation abarbeiten lassen. Die Bereitstellung und Nutzung von R für Grid-basiertes Rechnen unterscheidet sich hier nicht von anderen Programmen und ist ein Standardverfahren des Grid.[69]

Für die sozioökonomische Berichterstattung wird der Einsatz von R auf Basis des Globus Toolkit empfohlen. Mit dieser Lösung wird einerseits eine hohe Kompatibilität zu bestehenden Strukturen in D-Grid erzielt und gleichzeitig eine flexible Gestaltung der virtuellen Forschungsumgebung ermöglicht. Ebenso können in der Zukunft beliebige weitere Statistik-Anwendungen integriert werden, sofern die jeweilige Anwendungslizenzierung dies erlaubt. Hinzu kommen die von Fedora/iRODS stammenden Anforderungen, die mit der Globus-Middleware zu erfüllen sind.

---

[65] Die Problematik der Funktionsüberschreibungen wurde insbesondere in der Biostatistik beobachtet.

[66] Eine Quelle, um ggf. den Umstieg von SPSS, SAS oder Stata auf R zu erleichtern bietet http://rforsasandspssusers.com.

[67] Muenchen, RA (2010): R for stata users, 1st. Auflage, Statistics and computing, Springer, New York, NY.

[68] Muenchen, RA (2009): R for SAS and SPSS users, Statistics and computing, Springer, New York, NY.

[69] Es gibt weitere, workflow-orientierte Möglichkeiten wie GridR, ein vom Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS entwickeltes R-Packet, welches Funktionen zur Ausführung von R-Befehlen auf entfernten Rechnern ermöglicht. Für den Einsatz in einer virtuellen Forschungsumgebung bietet GridR eine direkte Anbindung an das Gridge Toolkit (http://www.gridge.org/). Da das Gridge Toolkit seit 2006 keine Aktualisierung erfahren hat, ist von dem Einsatz von GridR abzuraten. Ebenso müsste diese Komponente neben bereits etablierten Diensten separat gepflegt werden, wodurch mögliche Synergieeffekte mit anderen Communitys nicht ausgeschöpft werden können.

Anforderungsprofil für Ressourcen-Provider

Ressourcen-Provider können verschiedenen Middleware-Installationen anbieten. Für die sozioökonomische Berichterstattung kommen nur solche in Betracht, die Globus Toolkit anbieten. Die Bereitstellung von Globus Toolkit wird in D-Grid durch die Referenzinstallation abgedeckt und erfordert i.d.R. keine zusätzliche Aktivität. Zusätzlich sollten die auf R basierenden fachspezifischen Dienste in einer Anwendungsdatenbank hinterlegt werden, damit ermittelt werden kann, auf welchen Systemen Berechnungen mit R ausgeführt werden können.

Aufgrund der strukturellen Anforderungen an die Datenhaltung der sozioökonomischen Berichterstattung ist eine Nutzung der D-Grid-Infrastruktur mit direkten Vereinbarungen mit dedizierten Providern zu ergänzen.

Beispiel für die Nutzung von Rechenleistung in der virtuellen Forschungsumgebung

Im Rahmen der vorliegenden Expertise wurde die Nutzung von R in Verbindung mit Globus Toolkit anhand eines einfachen Beispiels validiert. Hierbei wurde ein einfacher Arbeitsablauf mit drei Parametern zur Übergabe an R erstellt und auf einer dedizierten Rechnerressource angestoßen. Die Ergebnisausgabe erfolgte in eine temporäre Datei. Als Rechnerressource wurde eine Maschine des Grid-Clusters in der Universitätsmedizin Göttingen verwendet.

Die gewählten drei Parameter des Ablaufs definieren 1) den Batch-Aufruf von R, 2) die auszuführende R-Syntaxdatei und 3) die Datei für die Aufnahme der Ergebnisse aus R[70]. Der Aufruf des Workflows verwendet den Web-Service-Aufruf WSGRAM des Globus Toolkit, um die Grid-Middleware anzusprechen.

Abschätzung des Integrationsaufwands für R

Der Aufwand für die Integration von R auf der Grundlage von Globus Toolkit als Grid-fähige Anwendung umfasst drei zentrale Aufgabenstellungen:

1. Installation / Verteilung von R auf Ressourcen in Abstimmung mit den jeweiligen Ressourcenbetreibern. Hierzu zählt die Entwicklung eines geeigneten Linux Shell-Skripts, um die R-Aufrufe durchführen zu können.
2. Entwicklung und Integration einer grafischen Schnittstelle zur Auswahl von Arbeitsdaten und Syntaxdateien aus dem Datenmanagement.
3. Integration einer Funktionalität zum direkten Ausführen von statistischen Berechnungen in die virtuelle Arbeitsumgebung basierend auf den Schritten 1 und 2.

Zu 1.) Es fällt Koordinationsaufwand im Umfang der zu integrierenden Ressourcen an und ist abhängig von der Anzahl der zu berücksichtigenden Ressourcen-Provider. Grundsätzlich ist dies unkritisch und erfordert pro Ressourcen-Betreiber 1 Woche einmaligen Personalaufwand. Die Installation von R hängt wiederum vom Umfang der benötigten Pakete und verschiedener Installationen ab. Hier wird pro R-Installation und deren Verteilung auf die einzelnen Ressourcen ein Aufwand von 1 Woche benötigt. Dies ist ein wiederkehrender Aufwand in Abhängigkeit weiterer benötigter R-Varianten bzw. deren Pflege. Es wird dabei davon ausgegangen, dass für jede R-Installation ein eigenes Skript ge-

[70] Parameter 1: **CMD BATCH --slave,** Parameter 2: **/opt/viz/test/myscript.R,** Parameter 3: **/tmp/out2.txt**

pflegt wird. Sollten für bestimmte Berechnungen spezielle R-Pakete benötigt werden, so sollte deren Installation über den VO-Administrator (siehe Kapitel 4.3.7) in Auftrag gegeben werden.

Zu 2. und 3.) Wird eine webbasierte Benutzerschnittstelle gewählt, so können gegebenenfalls bestehende Web-Portale integriert werden – der entstehende Aufwand ist vernachlässigbar. Bei einer Eigenentwicklung müssen jedoch Teile der Funktionen der Web-Portale in die Benutzerschnittstelle integriert werden. Dieser Aufwand kann nicht fundiert geschätzt werden.

Management von bei der Verarbeitung entstehenden Datenreplikaten

Wenn eine Verarbeitung von Datensätzen im Grid mit Hilfe von R erfolgt, werden die Ausgangsdaten automatisch auf das System kopiert, auf dem die Verarbeitung stattfindet. Dabei entstehen temporäre Kopien der Daten, die von sich aus keiner weiteren Verwaltung unterliegen. Es gibt zwei Möglichkeiten zum Umgang mit dieser Situation. Zum einen können die Replikate nach der Verarbeitung direkt wieder gelöscht werden, zum anderen können diese Datenkopien gleichzeitig als Sicherungskopien in die Datenverwaltung mit aufgenommen werden. Letztere Lösung hat zwar sehr viele Vorteile, bedarf jedoch einem relativ großen Aufwand für die Umsetzung. Daher sollte in einer ersten Version der virtuellen Arbeitsumgebung auf diese Vorteile verzichtet und die erste Lösung eingesetzt werden.

Um dennoch eine sichere Datenablage zu erreichen, sind entweder eine Spiegelung der Daten bei mindestens zwei Storage-Providern, die durch entsprechende Verträge gesichert wird, oder ein tägliches Backup der Daten auf einem anderen Speicher-Medium (Tape) notwendig.

4.5 Weitere Tools zur kollaborativen Arbeit

4.5.1 Projektinterne Kommunikation

Für die projektinterne Kommunikation bieten sich unterschiedliche Werkzeuge an. Diese variieren von so genannten Wikis und Content-Management-Systemen[71] über webbasierte Foren bis hin zu Instant-Messaging-Applikationen.

Wikis

Mit einem WiKi können kollaborativ webbasierte Dokumente, bzw. Inhalte erarbeitet werden. Hier ist Wikipedia (http://www.wikipedia.de) das bekannteste Beispiel. Es bietet eine durch wenige Formatierungs-Kommandos zu erzeugende HTML-basierte Webseite. Die einzelnen Seiten sind unabhängig voneinander von verschiedenen Nutzern zu bearbeiten. Der Vorteil des Wiki sind die sehr geringen Anforderungen an den Nutzer, der eine Seite erstellt. Die Strukturierung und sinnvolle Verknüpfung der einzelnen Seiten mittels Hyperlinks ist jedoch nur für relativ unverbundenen Inhalt (Enzyklopädie etc.) ohne größeren Aufwand und ständige Wartung zu realisieren. Es ist beim Einsatz eines Wikis zu empfehlen, projektweite Festlegungen zu dessen Struktur und Nutzung zu treffen.

Webbasierte Foren

Zur dokumentierten Diskussion von Forschungsaspekten oder Fragen bieten sich webbasierte Foren an. Diese würden sich als Instrument zur schriftlichen Diskussion von Themen eignen, besser als eine Mailingliste, die eher für den Frage/Antwort-Typ von Diskurs geeignet ist. Hierfür gibt es bereits unterschiedlichste freie Software, die entweder lokal installiert, oder über das Netz bei freien Anbietern

[71] Siehe hierzu 4.5.2 weiter unten.

genutzt werden kann. Für die virtuelle Arbeitsumgebung wäre aus Sicherheitsgründen die erste Variante zu empfehlen.

Instant-Messaging

Instant-Messaging ist eine beliebte Form des Gesprächs über digitale Medien, bei dem sich mehrere Gesprächsteilnehmer über Textnachrichten mit einander unterhalten. Für diese weithin als Chatten bekannte Anwendung gibt es mittlerweile sehr viele Systeme. Die wohl bekanntesten Vertreter sind ICQ und Skype[72]. Aber auch andere Plattformen bieten Instant-Messaging immer häufiger als Teil ihrer Funktionalität zur kollaborativen Arbeit an. Ein Beispiel hierfür ist das Softwarepaket Adobe® Acrobat® Connect™ Pro[73].

Es existieren mittlerweile ganz unterschiedliche Software-Protokolle, mit denen Instant-Messaging realisiert werden kann. Weiterhin bieten diverse Tools entsprechende Programmierschnittstellen, damit die Protokolle auch durch andere Softwaresysteme genutzt werden können. Ein Beispiel hierfür ist die frei verfügbare Smack API[74], welche Instant-Messaging über das Protokoll XMPP[75] erlaubt. Aber auch andere Implementierungen sind verfügbar. Um Instant-Messaging in die virtuelle Arbeitsumgebung zu integrieren, könnten derartige Programmierschnittstellen eingesetzt werden.

Wie erwähnt ist Instant-Messaging oft nur ein Teil eines Systems zur Kommunikation oder kollaborativen Arbeit. Bei Skype ist es z.B. auch möglich Dateien untereinander auszutauschen und via Voice over IP (VoIP) mit einander zu telefonieren. Das erwähnte Adobe® Acrobat® Connect™ Pro ergänzt derartige Funktionalitäten noch dadurch, dass mehrere Benutzer gleichzeitig eine Bildschirmpräsentation verfolgen und in dieser sogar Notizen oder Markierungen vornehmen können. Die grundlegenden Funktionen sind hierbei bereits über eine Webschnittstelle und somit direkt aus einem Internetbrowser nutzbar. Der volle Funktionsumfang kann jedoch nur mit einer entsprechenden Clientsoftware genutzt werden.

Auch für derart komplexe Systeme wäre eine Integration in eine virtuelle Arbeitsumgebung sinnvoll. Dies ist relativ einfach umsetzbar, sofern die Systeme eine webbasierte Schnittstelle haben. Für die Nutzung anderer Systemschnittstellen ist unter Umständen ein großer Implementierungsaufwand notwendig. Außerdem sind proprietäre Schnittstellen meist wenig dokumentiert oder gar per Nutzungsbedingungen nicht ansprechbar. Im Fazit gibt es viele für die kollaborativ ausgerichtete sozioökonomische Berichterstattung interessanten Funktionalitäten, die mit unterschiedlichen Aufwänden in eine virtuelle Arbeitsumgebung integriert werden könnten.

Mailing-Listen

Die Mailing-Listen bieten eine einfache Form der Kommunikation innerhalb eines Projekts. Sie sind jedoch vor allem für organisatorische Mitteilungen und Abstimmungen bei vielen Teilnehmern sinnvoll. Durch Threads kann auch eine gewisse Ordnung von Diskussionsthemen erreicht werden, für

---

[72] ICQ: Details und Download unter http://www.icq.de/. Letzter Zugriff am 09.06.2010; Skype: Details und Download unter http://www.skype.de/. Letzter Zugriff am 09.06.2010.

[73] Details unter http://www.adobe.com/de/products/acrobatconnectpro/. Letzter Zugriff am 09.06.2010.

[74] Erläuterungen unter http://www.igniterealtime.org/projects/smack/index.jsp. Letzter Zugriff am 09.06.2010.

[75] Extensible Messaging and Presence Protocol, Weitere Details unter: XMPP Standards Foundation (2010), Letzter Zugriff: 2010.06.09, URL: http://xmpp.org.

eine nachvollziehbare Kommunikation und Dokumentation (z.B. Protokolle oder Materialsammlung für Konferenzen etc.) bedürfen die Mailing-Listen einer Ergänzung durch ein leicht zu bedienendes HTML-basiertes Intranet.

4.5.2 Disseminierung

Eine wichtige Anforderung der sozioökonomischen Berichterstattung ist die Möglichkeit Zwischen- und Endergebnisse einer Forschung auf Projektwebseiten veröffentlichen zu können. Wichtig ist dabei, dass sich das notwendige technische Know-How in Grenzen halten sollte. Die meisten existierenden Technologien zum Aufbau von Webseiten wurden jedoch eher für den Webtechnologie-versierten Benutzer entwickelt. Mögliche Lösungen bieten entweder viel Funktionalität unter der Voraussetzung von viel Anwenderwissen, so wie es bei einigen Content Management Systemen (CMS) der Fall ist, oder weniger Funktionalität aber dafür mit wesentlich weniger notwendiger Vorerfahrung der Benutzer. Hierunter zählen einige CMS, wie auch Wikis (siehe Abschnitt 4.5.1).

Die Anforderung der sozioökonomischen Berichterstattung hinsichtlich der möglichst einfachen Veröffentlichung von Ergebnissen deckt sich mit den Anforderungen vieler Projekte. Das System sollte ein einfaches, vorgegebenes Navigationsschema, feste Layouts mit Style-Sheets und einen WYSIWIG- Editor mit vertrauten Formatierungselementen bieten. Außerdem sollte ein minimaler redaktioneller Workflow von Erstellung/Prüfung/Veröffentlichung möglich sein. Die Lösung soll sowohl für die Public Website wie für die Intranet-Site anwendbar sein[76].

[76] Wünschenswert ist eine Lösung, die sowohl die Public Website als auch die HTML-basierte projektinterne Kommunikation bewältigen kann. Es wird an dieser Stelle empfohlen, z.B. das von WissGrid sowohl für die Projektpräsenz im Internet als für das Intranet verwendete OpenSource CMS Lenya in Erwägung zu ziehen

5 Architekturskizze

In diesem Kapitel wird auf Basis der vorhergehenden Analysen eine grobe Architektur für eine virtuelle Arbeitsumgebung erarbeitet. Zunächst wird eine Grobstrukturierung und Aufteilung der Funktionalitäten vorgenommen. Danach wird auf einzelne Funktionsblöcke genauer eingegangen und der Einsatz bestimmter Technologien vorgeschlagen.

5.1 Überblick

Bei der Architektur einer virtuellen Arbeitsumgebung muss zwischen den für die sozioökonomische Berichterstattung notwendigen Werkzeugen und den für die Werkzeuge unabdingbaren, jedoch im Hintergrund arbeitenden Funktionalitäten unterschieden werden. Die Werkzeuge und die im Hintergrund laufenden Funktionalitäten bilden zusammen die virtuelle Arbeitsumgebung.

Die Werkzeuge werden hauptsächlich in der graphischen Nutzerschnittstelle der virtuellen Arbeitsumgebung Platz finden. Sie sollten alle notwendigen Funktionalitäten zur

- Datenverwaltung – Operationen ähnlichen denen eines Dateisystems, Versionsverwaltung, logische Verknüpfungen, Festlegung von Nutzungsrechten und -zeiträumen, Suchfunktionen
- Datenbearbeitung – hauptsächlich Editieren von Syntaxdateien
- Datenvergleich – Werkzeuge zum Vergleich zweier Syntaxdateien oder Syntaxdateiversionen
- Datenverarbeitung – Durchführen von statistischen Berechnungen
- Datenanbieterzugriff – Zugang zu Originaldaten von FDZ, Fernrechnen und Ablage von Daten zur langfristigen Archivierung
- Konfiguration – system-, installations- und nutzerspezifische Einstellungen
- Verwaltung – Verwaltung von Benutzergruppen, Rollen und globalen Rechten
- Kollaboration – Werkzeuge zur Zusammenarbeit der Forscher
- Publikation – Veröffentlichung von Forschungsergebnissen im Forschungsverbund sowie auf Webseiten

bieten. Zu den dahinter liegenden Funktionalitäten, im Folgenden unter dem Begriff Systemkern zusammengefasst, gehören die

- Datenablage – Nutzung von Speicherkapazitäten im Grid zur Ablage von Dateien
- Datenorganisation – Ablage und Verwaltung der Metadaten von Dateien, Grundlage für Suchfunktionen
- Datenverarbeitung – Ausführen von statistischen Berechnungen im Grid
- Sicherheit – Gridbasierte Funktionen zur sicheren Ablage von Dateien und Ausführung von Berechnungen
- Kommunikation – Infrastruktur für die Kommunikation der Forscher, z.B. Wiki-Server
- sonstige Dienste – z.B. zur Publikation von Webseiten, zum Logging, etc.

Diese Trennung und grundlegende Struktur wird in Abbildung 6 als grobes Konzept für die virtuelle Arbeitsumgebung noch einmal dargestellt.

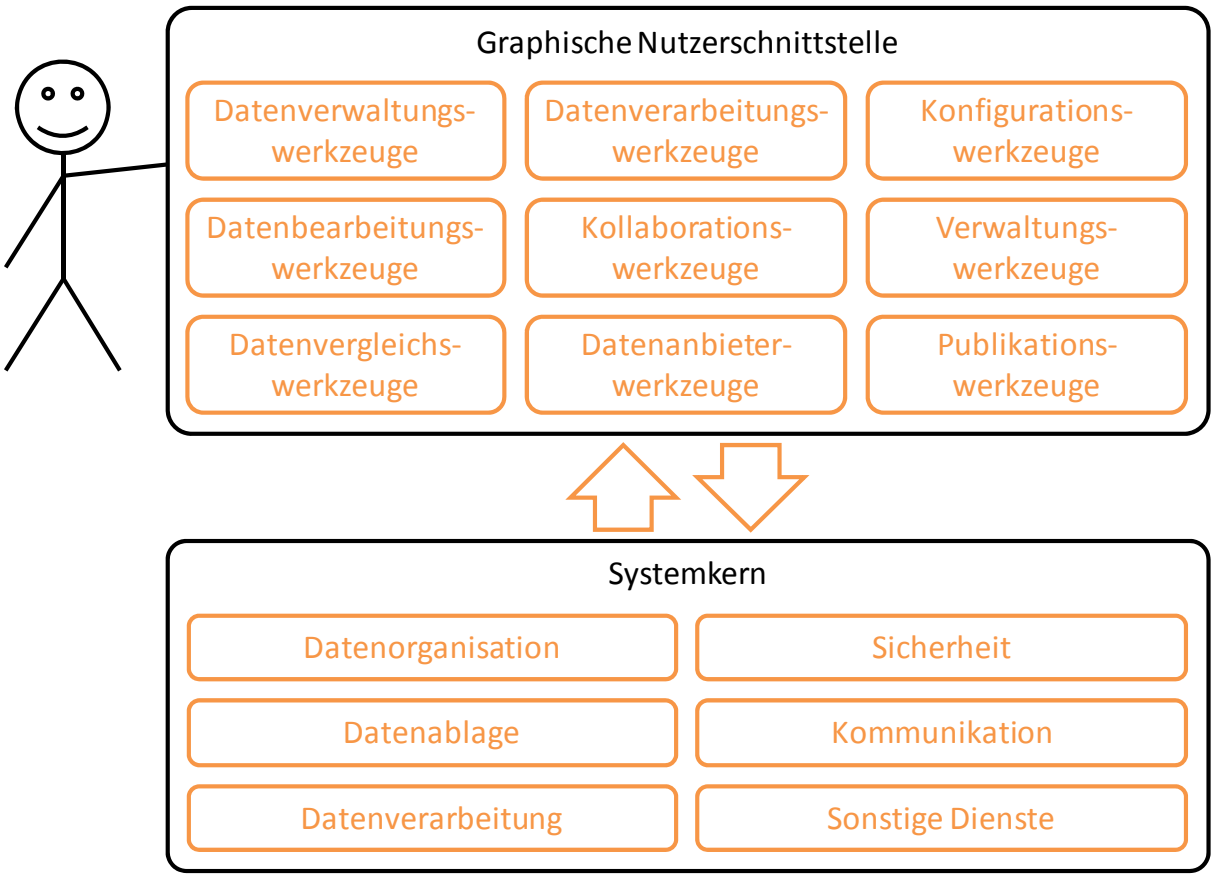

Abbildung 6 - Grobe Architektur der virtuellen Arbeitsumgebung

Die zur graphischen Nutzerschnittstelle sowie zum Systemkern gehörenden Komponenten, so wie sie in Abbildung 6 zu sehen sind, müssen als Funktionsblöcke verstanden werden, die nicht nur je eine einzelne Funktionalität umfassen, sondern mehrere zusammengehörende Funktionalitäten subsumieren. Im Folgenden soll genauer auf die jeweiligen Details eingegangen werden. Allerdings sei hierzu bemerkt, dass die Beschreibung der einzelnen Werkzeuge nicht der Fokus der Expertise ist. Stattdessen wird erläutert, an welcher Stelle die im D-Grid existierenden Systeme eingesetzt werden können und sollten und wie dieser Einsatz genau aussieht. Dennoch wird versucht die groben Elemente der Werkzeuge zu benennen und kurz zu erläutern.

5.2 Komponenten und Funktionsgruppen

In Abbildung 6 sind die einzelnen Funktionsblöcke, welche für die virtuelle Arbeitsumgebung benötigt werden, jeweils auf die graphische Nutzerschnittstelle, sowie auf den Systemkern verteilt worden. In diesem Kapitel werden sie genauer beschrieben, wobei die Reihenfolge der Beschreibungen eher inhaltlich und weniger strukturell gewählt wurde. So werden z.B. alle Funktionsblöcke, unerheblich ob zur graphischen Oberfläche oder zum Systemkern gehörend, die sich mit Datenspeicherung beschäftigen, zusammen beschrieben.

5.2.1 Sicherheit

In der sozioökonomischen Berichterstattung spielt die Sicherheit, gerade im Hinblick auf den Datenschutz eine sehr große Rolle. Die Beschreibungen zu allen Funktionsblöcken der virtuellen Arbeitsumgebung müssen, sofern notwendig, mit aufzeigen, wie sie die Sicherheitsanforderungen der sozio-

ökonomischen Berichterstattung umsetzen wollen. Dies ist umso einfacher möglich, wenn alle Funktionalitäten, die sich ausschließlich mit dem Aspekt Sicherheit beschäftigen, bereits vor der Komponentenbeschreibung eingeführt und erläutert wurden. Aus diesem Grund wird der Funktionsblock Sicherheit der virtuellen Arbeitsumgebung hier als erstes beschrieben.

Funktionsweise

Die Anforderungsanalyse aus Kapitel 3.2.7 ergibt, dass die sozioökonomische Berichterstattung teilweise sehr hohe Anforderungen hinsichtlich Sicherheit in einer verteilten Umgebung, vor allem bezogen auf die Forschungsdaten, formuliert. Allerdings sind derartige Anforderungen im Bereich anderer Forschungsinfrastrukturen nicht unbekannt. Im Rahmen der D-Grid Initiative, in denen viele Projekte zum Aufbau von Forschungsinfrastrukturen zusammenarbeiten, werden aus diesem Grund regelmäßige Workshops zum Thema Sicherheit im Grid abgehalten[77]. Besprochene Themen umfassen hier unter anderem die sichere Ablage von Daten im Grid, die sichere Datenverarbeitung, vor allem bei paralleler Nutzung von Rechenressourcen durch konkurrierende Grid-Nutzer, sowie die detaillierte Spezifikation von Zugriffsrechten für einzelne Benutzer, Benutzergruppen und Rollen. Auch wenn die in diesem Kontext spezifizierten Sicherheitsanforderungen oft einen anderen Hintergrund haben, so stellen sie dennoch ähnliche Erwartungen an die Sicherheit einer Grid-Infrastruktur wie im Rahmen dieser Expertise erarbeitet.

Mit den im Kapitel 4.3.7 vorgestellten Lösungen adressiert das D-Grid einige der von der sozioökonomischen Berichterstattung gestellten Anforderungen hinsichtlich Sicherheit. Die Systeme stellen zum einen eine sichere Authentifizierung und Autorisierung von Nutzern zur Verfügung und erlauben gleichzeitig die Spezifikation von Zugriffsrechten für Nutzergruppen und -rollen. Die existierende Grid-Infrastruktur mit den verschiedenen Middlewares ist fähig, die Sicherheitssysteme zu nutzen und die von ihr vorgegebenen Nutzungsberechtigungen zu beachten und umzusetzen.

Für die virtuelle Arbeitsumgebung wird vorgeschlagen, die im Grid existierenden Sicherheitsmechanismen zu nutzen, um die Sicherheitsanforderungen vollständig adressieren zu können. Dazu gehört der Einsatz der Grid Security Infrastructure in Kombination mit einem VOMRS. Das sich daraus ergebende Prinzip wird in Abbildung 7 noch einmal dargestellt. Darin nutzt der Anwender eine bestimmte Funktionalität, die zum Beispiel von der virtuellen Arbeitsumgebung über eines der Werkzeuge zur Verfügung gestellt wird. Die Authentifizierung des Nutzers erfolgt über dessen Zertifikat. Die Berechtigungen werden mit Hilfe des Zertifikats, der Grid Security Infrastructure und der VOMRS verifiziert. Dies findet in der virtuellen Arbeitsumgebung innerhalb des Systemkerns statt. Liegen die Nutzerrechte vor, so werden die Aktionen ausgeführt. Alle auf Basis der GSI stattfindenden Datentransfers werden verschlüsselt[78].

[77] Vgl. http://www.d-grid.de/ unter dem Menüpunkt Veranstaltungen/Abgeschlossenen Veranstaltungen.

[78] Die zugrunde liegende Technologie baut auf den derzeitigen Standards für gesicherte Netzwerk-Verbindungen auf. Die Referenzimplementationen sind hier OpenSSL (Secure Socket Layer) und Open SSH (Secure Shell).

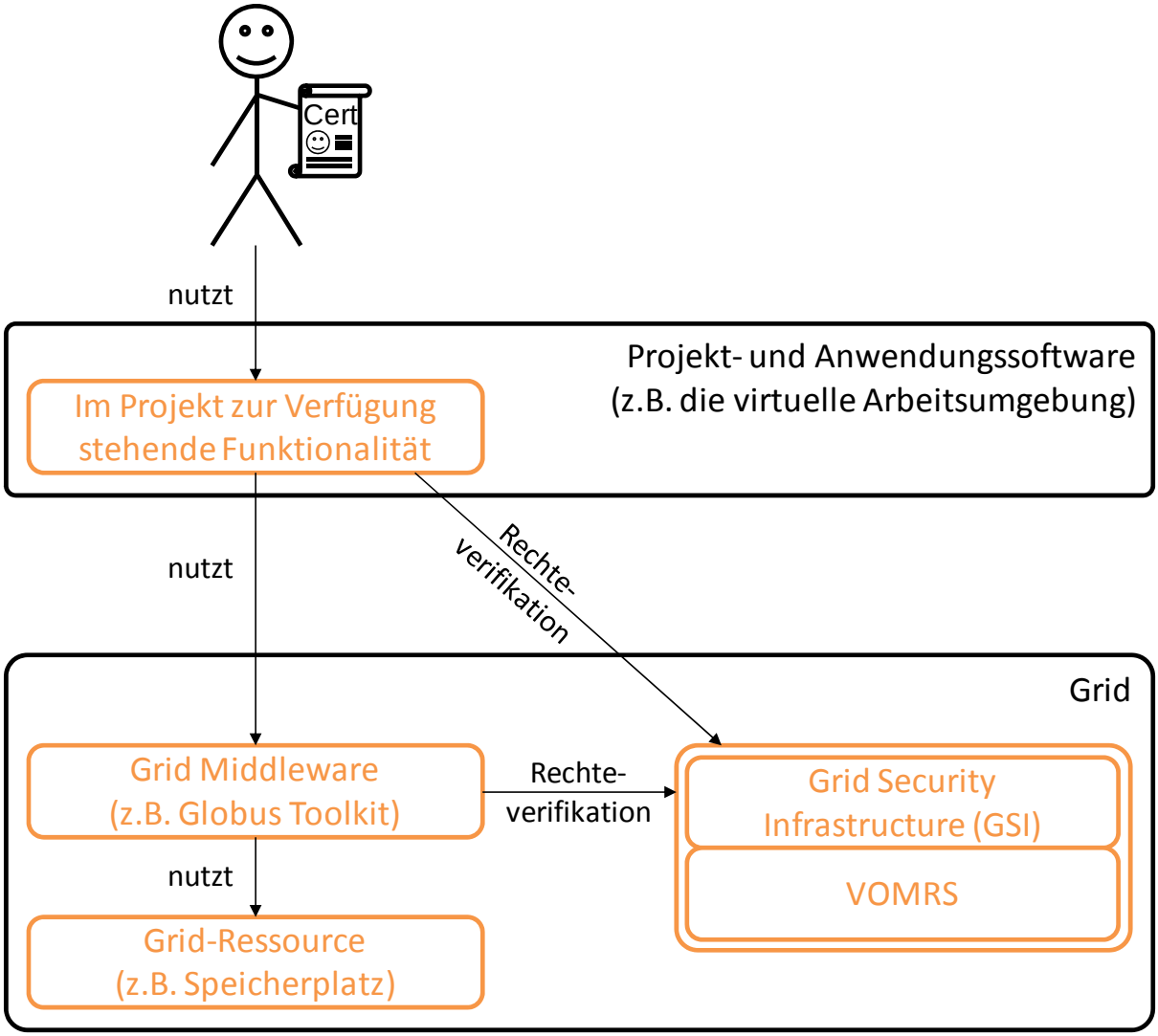

Abbildung 7 - Mehrschichtiger Aufbau einer Sicherheitsinfrastruktur

Die Middleware reicht die Überprüfung der Berechtigungen an das lokale System weiter, in dem einem Zertifikat ein lokaler Benutzeraccount zugeordnet ist. Die lokalen Berechtigungen sind für die einzelnen Nutzer (und -gruppen) durch Vereinbarung mit dem Ressourcenprovider festgelegt. So ist auch zu jedem Zeitpunkt eine Nachverfolgbarkeit der Aktionen jedes Grid-Nutzers gegeben. Außerdem können so auch die Aktionen des Benutzers gegenüber den Aktionen anderer Benutzer abgeschottet werden.

Dieses Verfahren entspricht vollständig den im Grid etablierten Standards. Es gibt auch weitere Systeme, die Anwendung finden. Sie sind je nach Anwender-Community verschieden. Die Weiterentwicklung der Sicherheits-Infrastrukturen ist ein elementarer Bestandteil des Grids.

Neben eher „schwächeren“ Anforderungen an die Sicherheit der Infrastruktur, die z.B. mit Shibboleth abgedeckt werden können, werden auch höhere Anforderungen realisiert. Soll beispielsweise die Einsicht in Datei-Inhalte auch den System-Administratoren verborgen bleiben, kann eine Verschlüsselung der Dateien vorgenommen werden, sodass beim Lesen und Schreiben von Dateien diese verschlüsselt im Dateisystem abgelegt werden. Diese Anforderung ist jedoch im Rahmen der sozioökonomischen Berichterstattung nicht erforderlich.

Anforderungsreferenz

Eine auf Basis der GSI aufbauende Datenhaltung für die virtuelle Arbeitsumgebung erfüllt die wesentlichen Anforderungen aus Kapitel 3.2.7. So wären z.B. die Daten vor unberechtigtem Zugriff geschützt und durch eine ausreichend fein definierbare Rechteverwaltung gesichert. Es könnten dadurch weiterhin Organisationsstrukturen von hinreichender Komplexität und darauf basierende Rechtedefinitionen abgebildet werden. Dieses Vorgehen nutzt aktuelle Sicherheitstechnologien. Die adressierten Sicherheitsanforderungen sind demnach Anforderung 65, Anforderung 66, Anforderung 68, Anforderung 69 und Anforderung 70.

Hinzu kommen die Anforderungen an eine Lokalität der Daten innerhalb einer Grid-Infrastruktur sowie deren Verarbeitung beim OnDemand-Rechnen. Es ist hier sinnvoll, durch Vereinbarung mit ausgewählten Anbietern im D-Grid die erforderlichen Kapazitäten durch SLA festzulegen. Hier bietet die GSI den sicheren Zugang, und die ausgewählten Anbieter die geforderte eingegrenzte Lokalität der Ressourcen[79].

Arbeitsaufgaben

- Eröffnung einer VO
    - Festlegung eines VO-Managers sowie der anderen notwendigen Strukturen
- Einrichten des VOMRS, Grundstruktur der Gruppen
- Vereinbarungen mit Grid-Ressource-Providern über Struktur und Anforderungen an die Nutzer- und Gruppenprofile, Nutzung von Software, Lizenzen-Nutzung auf dedizierten Ressourcen
- Implementierung einer Schnittstelle zur Zertifikats-Abfrage/Verbindung zur Grid-Middleware
- Weitere Anforderungen können aus Vereinbarungen mit FDZ erwachsen

5.2.2 Datenablage und Datenorganisation

Die Datenablage stellt eine der größeren Herausforderungen der virtuellen Arbeitsumgebung dar. Sie muss zum einen Daten mit ganz unterschiedlichen Formaten langzeitverfügbar und vor allem sicher speichern können, aber auch einen effizienten Zugriff auf die Daten gewährleisten. Ebenso stellen die gegebenenfalls großen Datenmengen eine gewisse Herausforderung, z.B. hinsichtlich Speicherplatz, Datentransfer und Wiederauffindbarkeit, dar.

Funktionsweise

Hinsichtlich der Datenorganisation sind die Anforderungen der sozioökonomischen Berichterstattung sehr auf Metadaten und konkrete Inhalte konzentriert. Die Metadatenstrukturen müssen verschiedene Informationen aufnehmen können, flexibel erweiterbar sein und durch Strukturen existierender Standards, wie DDI, erweitert werden können. Außerdem müssen Daten anhand ihrer Metadaten, teilweise aber auch anhand ihres Inhalts über geeignete Suchfunktionen wieder auffindbar sein.

[79] Gerade bei der Erfüllung solcher Anforderungen, die in der Wissenschaft nicht unüblich sind, besteht ein signifikanter Unterschied zwischen einer Grid- und einer Cloud-Infrastruktur.

Derartige Anforderungen sind im D-Grid Kontext nicht ganz neu. Auch andere Projekte haben teils sehr große Datenmengen, die organisiert und durchsuchbar abgelegt werden sollen. Daher gibt es bereits diverse Speichersysteme im D-Grid, welche die Anforderungen der sozioökonomischen Berichterstattung mehr oder weniger abdecken. Diese sind in Abschnitt 4.3 beschrieben. Dort wird auch deren Einsatz in der virtuellen Arbeitsumgebung bewertet. Die Kombination aus iRODS und Fedora aus Abschnitt 4.3.3 kristallisiert sich dabei als die beste Lösung heraus. Hier bildet iRODS den eigentlichen Datenspeicher und Fedora wird zur Datenorganisation eingesetzt.

Die Ablage von Daten erfolgt bei dieser Kombination über die Schnittstellen von Fedora. Im Rahmen der virtuellen Arbeitsumgebung sollten dafür Fedoras Softwareschnittstellen von den Werkzeugen der graphischen Nutzerschnittstelle aufgerufen werden. Fedora verwaltet intern die Metadaten und reicht die eigentlichen Daten weiter an iRODS. Letzteres nimmt die Daten an, startet die nach seiner Konfiguration auszuführenden Mechanismen und legt die Daten schließlich im Grid auf einem Speicherbereich ab.

Um die Daten zu lesen, zu bearbeiten oder zu löschen werden ebenfalls die Schnittstellen von Fedora aufgerufen. Auch die notwendigen Suchen können über diese Schnittstellen abgewickelt werden.

Die Metadatenverwaltung von Fedora ist sehr flexibel. Es können beliebige und frei definierbare Metadatenstrukturen abgelegt und wieder ausgelesen werden. Aufgrund der Verwendung des Dublin Core Metadaten-Standards bestehen Anpassungsmöglichkeiten an andere Standards wie z.B. DDI. Außerdem erlaubt Fedora entweder direkt, durch interne Module, oder indirekt, durch frei programmierbare Erweiterungen, die Suche auf diesen Metadaten oder auf den Dateninhalten selbst. Aus diesem Grund können alle Anforderungen zur Metadatenablage und zur Suche nach Daten über Fedora abgedeckt werden. Es gilt allerdings zu beachten, dass die Auswertung bestimmter Metadateninhalte den Werkzeugen der graphischen Nutzerschnittstelle der virtuellen Arbeitsumgebung vorbehalten bleibt.

Wie im Abschnitt 5.2.1 erwähnt, müssen gerade für die Datenablage die umfangreichen Anforderungen der sozioökonomischen Berichterstattung hinsichtlich Sicherheit und Zugriffsschutz Beachtung finden. Die Kombination aus Fedora und iRODS wird fähig sein, die Rechtevorgaben der Grid Security Infrastructure in Kombination mit VOMRS zu beachten und umzusetzen. Damit ist der im Abschnitt 5.2.1 beschriebene Ansatz zur sicherheitsorientierten Implementierung der virtuellen Arbeitsumgebung anwendbar.

Die Kombination aus iRODS und Fedora bietet sich somit insgesamt für die Datenablage und -organisation in der virtuellen Arbeitsumgebung an. Die bereitgestellten Funktionalitäten, auf die über die Schnittstellen von Fedora zugegriffen werden kann, bilden die Basis für einige Werkzeuge aus der graphischen Nutzerschnittstelle der virtuellen Arbeitsumgebung. Die groben Züge dieses Konzepts sind in Abbildung 8 dargestellt.

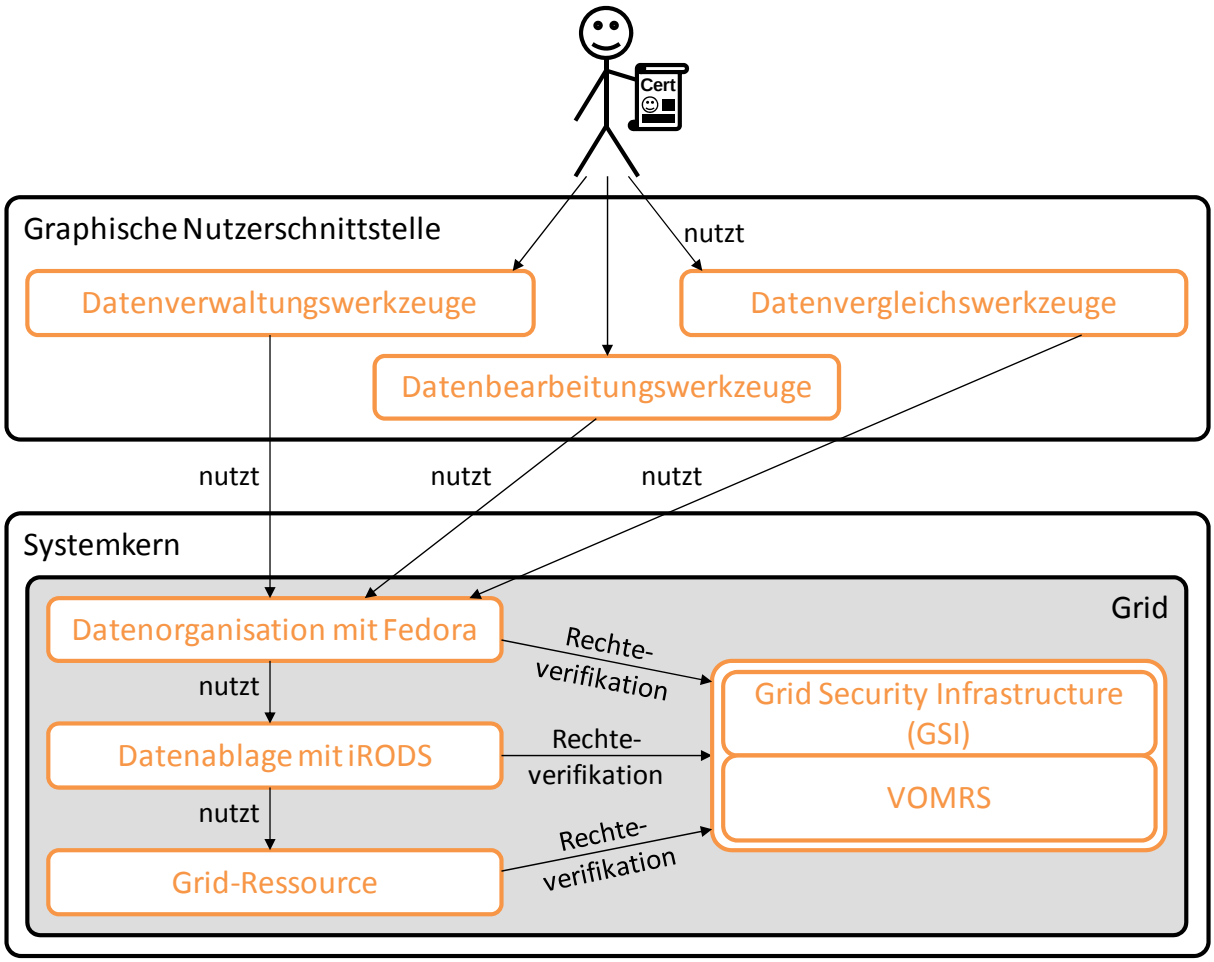

Abbildung 8 - Verwendung von Fedora und iRODS zur Datenorganisation und –ablage

Anforderungsreferenz

Mit den vielfältigen Möglichkeiten von Fedora und iRODS lassen sich diverse Anforderungen aus Kapitel 3.1 adressieren. Zum einen können die Daten formatunabhängig abgelegt und mit Metadaten beschrieben werden (Anforderung 5, Anforderung 6). Die Metadaten müssen dabei keinem vorgegebenen Schema folgen und sind damit sowohl für die technischen Belange der virtuellen Arbeitsumgebung (siehe alle Anforderungen hinsichtlich technischer Metadaten sowie Metadaten zur Beachtung von Nutzungszeiträumen, zur Abspeicherung von Provenienzinformationen oder Prüfsummen und Signaturen, etc.) als auch für die eigentliche Anwendung in der sozioökonomischen Berichterstattung (z.B. Anforderung 7) anpassbar. Die Festlegung eines passenden Schemas durch die sozioökonomische Berichterstattung ist allerdings schon frühzeitig erforderlich. Außerdem können sowohl Forschungsdaten als auch Syntaxdateien abgelegt werden. Es ist möglich, Datensätze untereinander in Beziehung zu setzen (siehe z.B. Anforderungen hinsichtlich der Provenienz von Daten), sie damit gegebenenfalls zu Kategorisieren (siehe Anforderungen hinsichtlich Kategorisierung) sowie mehrere Versionen zu speichern und zu verwalten (siehe Anforderungen hinsichtlich der Versionierung von Datensätzen). Es ist außerdem möglich, ausschließlich Metadaten ohne eigentliche Datensätze abzulegen, was direkt Anforderung 17 adressiert. Da sowohl iRODS als auch Fedora die Festlegungen einer Rechteverwaltung umsetzen können, sind auch die Anforderungen hinsichtlich Datensicherheit weitestgehend abgedeckt. Die Integration der von der GSI bereitgestellten Rechte und den entsprechenden Mechanismen in Fedora/iRODS ist u.a. auch eine wichtige Arbeitsaufgabe des AP3 von WissGrid, sodass vorhandene Synergien gut genutzt werden können. Nicht zuletzt bietet Fedora diverse Ergän-

zungen zum Aufbau einer Suchfunktionalität sowohl auf Metadaten, als auch auf den Daten selbst. Diese können außerdem durch selbstimplementierte Funktionen ergänzt oder ersetzt werden. Damit sind auch die Anforderungen an die Suchfunktionalitäten abgedeckt.

Arbeitsaufgaben

- Erarbeitung eines passenden Metadatenschemas ausgehend von Dublin Core für die unterschiedlichen Datenarten (z.B. mit Einbeziehung von DDI für Forschungsdaten) sowie für Kategorienhierarchien unter Beachtung von:
    - Technischen Metadaten
    - Fachlichen Metadaten
    - Zusätzlichen Versionsinformationen
    - Informationen über Verknüpfungen von Dateien
    - Angabe von Nutzungszeiträumen
    - Ablegen virtueller Datensätze
    - Speichern von Flags zur Implementierung diverser Funktionen der graphischen Oberfläche
- Ggf. Implementierung einer erweiterten Suchfunktionalität, unter anderem zugeschnitten auf das Metadatenschema sowie zur Unterstützung der Volltextsuche auf Syntaxdateien
- Implementierung der Interaktionen von Fedora und GSI in Bezug auf die Nutzerrechte
- Vereinbarung mit FDZ über Umgang mit zeitlich limitierten Datensätzen
- Implementierung/Konfiguration des Fedora/iRODS Speichermanagements
- Abschluss von SLAs mit ausgewählten Providern über Storage und Nutzung von Rechen-Kapazitäten

5.2.3 Datenverwaltungswerkzeuge

Fedora ist sehr flexibel und bietet nahezu unendlich viele Möglichkeiten für die Datenorganisation. Dieser generische Ansatz führt jedoch dazu, dass die Schnittstellen von Fedora ebenfalls sehr allgemein gehalten sind. Diese wären zwar im Rahmen der virtuellen Arbeitsumgebung nutzbar, nur lassen sie einen kontextbezogenen Komfort vollständig vermissen.

Es ist daher notwendig mit Hilfe von Werkzeugen von der darunterliegenden Funktionsvielfalt zu abstrahieren, um zum einen gezielt benötigte Funktionen zur Verfügung zu stellen und zum anderen bestimmte Abläufe im Hintergrund zu automatisieren. Dadurch wird dem Anwender eine auf ihn und seine Aufgaben und Vorhaben zugeschnittene Schnittstelle aufgebaut. Dies ist die Aufgabe der Datenverwaltungswerkzeuge.

Funktionsweise

Die in der virtuellen Arbeitsumgebung benötigten Datenverwaltungswerkzeuge haben als Ziel, dem Benutzer eine einfach zu verwendende, jedoch mächtige Schnittstelle zur Ablage und Verwaltung seiner Forschungsdaten und Syntaxdateien zu bieten. In der Anforderungsanalyse wurde bereits erar-

beitet, welche grundlegenden Funktionalitäten hierfür benötigt werden. Die Datenverwaltungswerkzeuge müssen genau diese Funktionalitäten zur Verfügung stellen.

Die Möglichkeiten zur Ablage von Dateien in Dateisystemen sind heutzutage auf vielen Computersystemen ähnlich umgesetzt und erfreuen sich daher einem gewissen Bekanntheitsgrad. Allerdings haben normale Dateisysteme auch Grenzen. Dennoch ist die Funktionsweise von Dateisystemen eine etablierte Form des Umgangs mit Daten und sollte daher die Grundlage für die Datenverwaltungswerkzeuge in der sozioökonomischen Berichterstattung bilden. Es gilt also in erster Linie eine Schnittstelle zur Verfügung zu stellen, die einem Dateibrowser eines Betriebssystems sehr ähnlich ist. Die grundlegenden Möglichkeiten einer solchen Schnittstelle sind in Anhang A.1 aufgelistet.

In der Anforderungsanalyse in den Kapiteln 3.2.1 und 3.2.2 wurde dargelegt, dass Verzeichnisse in Dateisystemen der Kategorisierung von Dateien dienen. Allerdings wurde dabei auch Nachteile und fehlenden Möglichkeiten beim Umgang mit einfachen Verzeichnissen benannt. Es wird daher vorgeschlagen die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung mit weiteren Funktionen anzureichern, welche die bekannten Möglichkeiten von Verzeichnissen erweitern und somit eine weiterführende Strukturierung der eigenen Arbeit erlauben. Daher sollte bei der Beschreibung der detaillierten Funktionen der virtuellen Arbeitsumgebung auch nicht von Verzeichnissen, sondern von Kategorien gesprochen werden, wobei deren Grundfunktionalität identisch mit jener von Verzeichnissen ist. Die funktionalen Ergänzungen zur Arbeit mit Kategorisierungen, welche die normalen Möglichkeiten des Umgangs mit Verzeichnissen ergänzen, sind in Anhang A.2 aufgelistet. Kernpunkte sind hierbei die erweiterten Möglichkeiten zur Arbeit mit Metadaten, die Verwendung mehrerer paralleler Hierarchien für einen Satz von Dateien sowie entsprechenden Werkzeugen zum effektiven Umgang und der Pflege dieser Hierarchien.

Ein weiterer in der sozioökonomischen Berichterstattung wichtiger Punkt ist die Versionierung von Dateien. Daher sollten die Datenverwaltungswerkzeuge außerdem Funktionalitäten für die Verwendung und Administration verschiedener Dateiversionen anbieten. Dazu zählen das Anzeigen, Vergleichen und Löschen von Versionen. Im Anhang A.3 sind mögliche Umsetzungen hierfür aufgelistet.

Wie in der Anforderungsanalyse in Kapitel 3.2.1 erarbeitet, gehört es zur guten wissenschaftlichen Praxis den Werdegang einer Forschung vollständig nachzuweisen. Aus diesem Grund ergaben sich Anforderungen hinsichtlich der logischen Verknüpfung von Daten inklusive entsprechender Beschreibungen. Damit wird außerdem die Provenienz der Daten hinreichend erfasst. Aufgrund der Nutzungsrechte für Originaldatensätze ergeben sich jedoch unter Umständen eingeschränkte Nutzungszeiträume für Forschungsdaten, was den Forscher zu deren Löschung zwingt. Daher müssen die Datenverwaltungswerkzeuge den Nutzer mit unterschiedlichsten Funktionalitäten unterstützen. Zunächst müssen Möglichkeiten zur Spezifikation von Verknüpfungen zwischen Dateien geschaffen werden. Dann sollten beim Löschen von Dateien oder Dateiversionen mögliche Verknüpfungen analysiert und ausgewertet, und der Nutzer um entsprechende Entscheidung (Löschen, Erhalten, etc.) gebeten werden. Außerdem sollten Nutzungszeiträume für Dateien definierbar sein und von der virtuellen Arbeitsumgebung ausgewertet werden. Im Falle des Ablaufs von Nutzungszeiträumen sollten dem Benutzer entsprechende Aktionen vorgeschlagen werden. Details zur möglichen Implementierung sind in Anhang A.4 zu finden.

Hinsichtlich sicherheitsrelevanter Aspekte ergaben sich in der Analyse in Kapitel 3.2.1 Anforderungen zur Spezifikation von eigenen und nutzergruppenweiten Zugriffsrechten auf Dateien. Die Datenverwaltungswerkzeuge sollten deren Spezifikation auf Datei- und Kategorienebene erlauben. Mögliche Umsetzungen dafür sind in Anhang A.5 enthalten.

Nicht zuletzt wird zur effektiven Verwaltung von Daten auch eine Suche nach Daten anhand ihrer Metadaten oder Inhalte benötigt. Diese Suche sollte dabei sowohl auf einer gerade angezeigten Kategorie, als auch global möglich sein. Eine mögliche Darstellung dieser Funktionalitäten in der graphischen Oberfläche der virtuellen Arbeitsumgebung findet sich in Anhang A.6.

Im Hintergrund arbeiten die Datenverwaltungswerkzeuge mit Zugriffen auf Fedora sowie gegebenenfalls weiteren Komponenten des Systemkerns, z.B. den Sicherheitskomponenten zur Ablage oder Verifikation von Rechten für andere Benutzer. Die Dateien, Kategorien und virtuellen Datensätze müssen als Fedora-Objekte abgelegt und mit Metadaten beschrieben werden. Für jede Datei, für jede Kategorie und für jeden virtuellen Datensatz existiert genau ein Objekt in Fedora. Die Metadatenstrukturen müssen so gewählt werden, dass alle vom Benutzer sowie von den Datenverwaltungswerkzeugen notwendigen Informationen, wie Metadaten, vom Benutzer vergebene Flags (z.B. Schreibschutz) oder Nutzungszeiträume, abgelegt werden können. Für die Verknüpfung von Kategorien mit Kategorien, von Kategorien mit Dateien sowie von Dateien untereinander werden soweit möglich die Verknüpfungsmöglichkeiten von Objekten in Fedora verwendet. Auch für die Ablage und Verwaltung verschiedener Versionen ist der Einsatz der entsprechenden Funktionalitäten von Fedora zu empfehlen. Zur Unterstützung der Suche sollten die von Fedora mitgelieferten Funktionalitäten eingesetzt und gegebenenfalls zur Optimierung (Performance) oder funktionalen Ergänzung (unscharfe Suche, Volltextsuche) erweitert werden.

Anforderungsreferenz

Durch die Datenverwaltungswerkzeuge werden eine Reihe der in Kapitel 3.1 aufgeführten Anforderungen an die virtuelle Arbeitsumgebung adressiert. Zu diesen gehören all jene, die sich mit der Kategorisierung von Dateien, den Metadaten von Dateien, teilweise den Zugriffsrechten auf Dateien sowie anderen ablageorientierten Aspekten inklusive der Suchfunktionalität beschäftigen.

Arbeitsaufgaben

- Entwicklung einer Darstellung der Repository-Inhalte unter Berücksichtigung von Kategorien
- Entwicklung einer Oberfläche zur Versions-Verwaltung und -Verwendung für Dateien
- Oberfläche zur Spezifikation, Anzeige und Auswertung von Dateiverknüpfungen
- Implementierung der Auswertung vorgegebener Nutzungszeiträume
- Implementierung des iterativen Löschens von Daten bzw. des Umwandelns in virtuelle Datensätze
- Implementierung einer Möglichkeit zur Spezifikation unterschiedlicher Flags zu Dateien (z.B. eigenes Löschen oder Bearbeiten temporär verhindern)
- Implementierung einer Anbindung zum Rechteverwaltungssystem
- Implementierung einer Oberfläche zur Suche nach Kategorien und Dateien
- Implementierung einer Oberfläche zur Volltextsuche für Syntax-Dateien

5.2.4 Datenbearbeitungswerkzeuge

Die virtuelle Arbeitsumgebung soll den Anwendern neben der reinen Ablage von Daten auch die Möglichkeit zu deren Bearbeitung geben. Hierbei liegt der Fokus vor allem auf der Bearbeitung von

Syntaxdateien und den Metadaten aller eingesetzten Dateiformate. Aber auch Konvertierungs- und Validierungsfunktionen sollten vorgesehen werden.

Funktionsweise

Die Datenverwaltungswerkzeuge erlauben bisher lediglich die Ablage und rudimentäre Verwaltung von Dateien. Ziel der Datenbearbeitungswerkzeuge ist es, diese Funktionalitäten um die Bearbeitung von Daten und deren Metadaten mit Hilfe passender Editoren zu erweitern. Außerdem sollten für bestimmte Dateiformate Konvertierungs- und Validierungsroutinen zur Verfügung stehen. Dies sollte direkt und einfach möglich und aufrufbar sein. Ein grundlegender Ansatz hierzu ist in Anhang A.7 aufgeführt.

Für die Bearbeitung von Daten und Metadaten werden unterschiedlichste Editoren notwendig. Hier gilt es soweit wie möglich existierende Systeme zu verwenden, auch um dem Forscher die Nutzung der vertrauten Werkzeuge zu gestatten. Dementsprechend sollten z.B. Datensätze in einem statistikprogrammspezifischen Format auch direkt mit diesem geöffnet und angezeigt werden können. Gleiches gilt für Syntaxdateien. Die anderen eingesetzten Formate, wie z.B. unterschiedliche Bildformate oder PDF sollten ebenfalls mit dafür passenden Programmen geöffnet werden können. Hierdurch wird sichergestellt, dass dem Forscher die jeweils notwendigen Tools zur Verfügung stehen. Diese Tools übernehmen zumeist auch eine einfache Formatvalidierung für die Dateien, da fehlerhafte Formate oft nicht geöffnet werden können.

Da die Bearbeitung von Syntaxdateien eine zentrale Rolle in der sozioökonomischen Berichterstattung übernimmt, sollte hierfür ein einfach zu bedienender aber mächtiger Editor zur Verfügung stehen. Dieser sollte Unterstützung für die korrekte Formulierung von Befehlen bieten und gleichzeitig eine Validierung von Syntaxdateien erlauben. Die verschiedenen Statistik-Programme bieten bereits Editoren an, die jedoch nicht die Anforderungen der Sozialwissenschaftler abdecken. In der sozialwissenschaftlichen Forschung hat sich vor allem TextPad (Siehe http://www.textpad.com) unter Verwendung von Erweiterungen zur Unterstützung von SPSS, Stata und R durchgesetzt. TextPad unterstützt bereits alle benötigten Syntax-Formate der sozialwissenschaftlichen Forschung in Form der Erweiterungen.

Nicht für jedes Dateiformat stehen auf jedem Rechnersystem die passenden Editoren zur Verfügung. So kann es z.B. sein, dass ein Wissenschaftler über die virtuelle Arbeitsumgebung auf eine Datei im Format eines bestimmten Statistik-Pakets zugreifen kann, aber weder eine Lizenz noch eine Installation der Software hat. Die virtuelle Arbeitsumgebung sollte hier soweit wie möglich Fall-Back-Lösungen anbieten, um den Forscher dennoch eine rudimentäre Arbeit mit diesen Dateien zu erlauben. So sollte für Syntaxdateien, sofern das passende Statistik-Paket nicht verfügbar ist, zumindest ein normaler Texteditor geöffnet werden. Für Excel-Tabellen bietet sich z.B. OpenOffice als Alternative an. Die Validierung der Dateien wird dadurch jedoch ungenauer.

Für die Bearbeitung der Metadaten aller Dateien sollte ebenso ein eigener Editor zur Verfügung stehen. Erste Details zu dessen möglichem Erscheinungsbild sind in Anhang A.8 aufgeführt. Dieser Editor sollte sowohl für Dateien, als auch für virtuelle Datensätze und verschiedene Versionen von Dateien aufgerufen werden können. Wichtig wäre außerdem, dass es für den Benutzer bei der Bearbeitung der Metadaten immer eindeutig ist, zu welcher Datei die Metadaten gehören. Dies kann am besten damit erreicht werden, dass die dazu gehörende Datei, virtuelle Datei, oder Dateiversion gleichzeitig im entsprechenden Editor mit angezeigt, oder in der Dateiverwaltung markiert wird.

Zu den mit dem Editor bearbeitbaren Metadaten sollten zumindest die von der virtuellen Arbeitsumgebung ausgewerteten Metadaten gehören. Diese umfassen unter anderem die Spezifikation von ausgewerteten Flags oder Nutzungszeiträumen. Außerdem sollten abhängig vom jeweiligen Dateiformat

weitere sinnvolle Metadatenfelder angegeben werden können. Existierende Metadatenstandards, wie z.B. DDI, sollten Anwendung finden.

Der Metadateneditor sollte einfache Möglichkeiten zur freien Ergänzung und Erweiterung von Metadatenstrukturen liefern. Dazu zählt auch, dass neue Metadatenelemente eingefügt und beschrieben, deren Semantik und Syntax festgelegt und die neuen Elemente durch die Definition von Gruppen zusammengefasst werden können. Der Metadateneditor sollte für diese neuen Strukturen genauso wie für die existierenden (siehe Vorschlag in Anhang A.8) funktionieren. Ergänzungen zu Metadatenstrukturen sollten genauso auch für die Metadaten anderer Dateien übernommen werden können. Hier bieten sich Export- und Importfunktionen für Metadatenstrukturen an.

Sollen Metadaten für eine Datei bearbeitet werden, für die noch keine Metadaten existieren, so sollte der Metadateneditor für bestimmte Metadatenwerte entsprechende Inhalte vorschlagen. Es ist z.B. möglich die technischen Metadaten teilweise auf Basis der Dateiinformationen zu generieren. Auch fachliche Metadaten können gegebenenfalls aus den Dateiinhalten extrahiert werden. Der Metadateneditor muss dafür eine Erkennung für Dateiformate und eine im Hintergrund arbeitende Informationsextraktion aufrufen können.

Die im Metadateneditor bearbeiteten Informationen sollten bei bestimmten Dateiformaten direkt mit den in den Dateien enthaltenen Metadaten abgeglichen und synchronisiert werden können. Werden z.B. die Metadaten einer Bilddatei im JPEG Format bearbeitet und ein Ersteller benannt, so sollte dieser auch auf einfache Weise in den EXIF-Header der JPEG Datei geschrieben werden können. Es muss hierbei für den Anwender transparent sein, welche Metadaten bereits synchronisiert sind, und welche nicht.

Bei der Bearbeitung der Metadaten von Syntaxdateien sollten weitere Zusatzfunktionen existieren. Syntaxdaten enthalten z.B. von sich aus keine Metadaten. Die meisten Forschungsdatenzentren erwarten jedoch für das Fern- oder Onsite-Rechnen bestimmte Kommentare in den Syntaxdateien, die eine vorgegebene Struktur haben und diverse Metadaten beinhalten. Der Metadateneditor sollte hier, ähnlich dem Schreiben von Metadaten in Headern anderer Dateien, die Möglichkeit bieten, die von den FDZ erwarteten Kommentarstrukturen auf Basis der Metadaten zu erzeugen. Gleichzeitig soll für den Benutzer ersichtlich sein, welche Metadaten in die Kommentare mit aufgenommen werden und ob die Metadaten mit den Kommentaren synchronisiert sind, oder nicht.

Zu den Metadaten von Syntaxdateien und Forschungsdaten zählen außerdem Informationen zur Provenienz. Wie in der Anforderungsanalyse erläutert, ist ein wichtiger Aspekt hierbei die Verknüpfung von Dateien. Für Syntaxdateien ist es sinnvoll die durch Befehle referenzierten Eingangs- und Ausgangsdaten mit den Syntaxdateien zu verknüpfen. Der Metadateneditor für Syntaxdateien sollte diese Befehle in den Syntaxdateien erkennen, die referenzierten Dateien auflösen und eine Möglichkeit zur Erzeugung von entsprechenden Verknüpfungen über die Datenverwaltungswerkzeuge erlauben.

Neben den Metadaten von Dateien sollen auch Metadaten von Kategorien bearbeitet werden können. Der Metadateneditor soll daher auch für Dateikategorien aufgerufen werden können. Die Bearbeitung der Metadaten von Kategorien ist jedoch ähnlich der Bearbeitung von Metadaten von Dateien und wird daher nicht weiter detailliert.

Neben den Editoren sollen außerdem Konvertierungswerkzeuge eingebunden werden können. Wie im Kapitel 4.3.6 erläutert, existieren hier jedoch nur wenige Routinen. Außerdem ist aufgrund des möglichen Informationsverlustes von dem regelmäßigen Gebrauch derartiger Funktionen abzuraten. Dennoch sollten existierende Komponenten in der virtuellen Arbeitsumgebung Einsatz finden. Wenn ein Datensatz z.B. mit dem passenden Statistik-Programm geöffnet wird, kann hierüber auch eine Um-

wandlung in ein Austauschformat (z.B. CSV) vorgenommen werden, was vom Programm i.d.R. unterstützt wird.

Obwohl die Editoren bereits eine erste Formatvalidierung vornehmen, ist der Einsatz von weiteren entsprechenden Tools sinnvoll. Daher sollten Werkzeuge, wie JHOVE, transparent in die virtuelle Arbeitsumgebung eingebunden und, wie im Anhang A.7 dargestellt, aufgerufen werden können.

Anforderungsreferenz

Die Datenbearbeitungswerkzeuge adressieren ebenfalls eine Reihe von Anforderungen an die virtuelle Arbeitsumgebung. Zu diesen zählen unter anderem Anforderung 6, Anforderung 7, Anforderung 10, und Anforderung 11, sowie Anforderung 15 bis Anforderung 17, Anforderung 39 bis Anforderung 41, Anforderung 44 bis Anforderung 51 und Anforderung 61 bis Anforderung 64.

Arbeitsaufgaben

- Ergänzung der Datenverwaltungswerkzeuge um Funktionen zum Aufruf entsprechender Editoren inklusive der Möglichkeiten zum Festlegen von Standardeditoren
- Integration eines vorhandenen Syntaxeditors mit Autovervollständigung, Syntax-Highlighting (hier der computertechnische Begriff Syntax) und weiteren Features
- Entwicklung eines zum Metadatenschema passenden Metadateneditors
    - Interaktion mit Fedora/iRODS für fachliche/technische Metadaten
    - Freie (nicht im Schema vorhandene) Metadaten hinzufügen und bearbeiten
    - Fähigkeit zur Bearbeitung von Metadaten von Kategorien
- Implementierung von Bibliotheken für die Metadatenextraktion und Ablage für einen feststehenden Satz von Dateiformaten und das Schreiben von Metadaten in die Standard-Dateien. Insbesondere sollen die Austausch-Formate unterstützt werden.
    - JHOVE zur Validierung von Datei-Formaten
    - Metadaten-Extraktion als Bibliothek
    - Metadaten-Header schreiben als Bibliothek
- Editor zum Erstellen und Anwenden von Kommentartemplates für Syntaxdateien unter Einbeziehung der Metadaten
- Editor zur automatischen Extraktion von Verknüpfungen zu Datensätzen aus Syntaxdateien inklusive Ablage in Metadaten
- Einbindung existierender Konvertierungs- und Validierungsroutinen

5.2.5 Datenvergleichswerkzeuge

Während eines Forschungsprozesses wird es gerade bei kollaborativer Arbeit notwendig, erstellte Daten mit einander zu vergleichen. Hierdurch können überhaupt erst Unterschiede erkannt und gegebenenfalls nachvollzogen werden. Dafür sind unterschiedliche Werkzeuge notwendig, die hier kurz erläutert werden sollen.

Funktionsweise

Die Datenverwaltungswerkzeuge erlauben das Ablegen und Strukturieren von Dateien. Von hier aus sollten Dateien mit einander verglichen werden können. Eine erste notwendige Funktion, die unabhängig von Dateiformaten oder -inhalten möglich sein sollte, ist der Vergleich zweier Dateien auf Identität. Zwei Dateien können als identisch angesehen werden, wenn sie auf Datenstromebene Bit für Bit identisch sind. Die virtuelle Arbeitsumgebung sollte daher ein Werkzeug zur Verfügung stellen, mit dem ein solcher Vergleich sowohl auf Dateien, als auch auf unterschiedlichen Versionen einer Datei möglich ist.

In der virtuellen Arbeitsumgebung werden Dateien und Kategorien durch Metadaten beschrieben. Auch hierfür bieten sich effektive Vergleichsmöglichkeiten an. Vor allem wenn zwei Dateien inhaltlich identisch sind, möchte man gegebenenfalls feststellen können warum sie dennoch zweimal abgelegt wurden. Dies lässt sich gegebenenfalls an Unterschieden in den Metadaten feststellen. Daher wird ein Werkzeug zum Vergleich und zum manuellen Abgleich von Metadaten zweier Dateien oder Kategorien benötigt.

Syntaxdateien nehmen im Kontext des Dateivergleichs eine besondere Rolle ein. Da es sich bei ihnen um Textdateien handelt, in denen die meisten Zeilen je eine eigene semantische Bedeutung haben, ist ein zeilenweiser Vergleich zweier Syntaxdateien sinnvoll. So können zum Beispiel unterschiedliche Versionen der gleichen Syntaxdatei mit einander verglichen und deren Entwicklung nachvollzogen werden. Aber auch parallele Entwicklungen an Syntaxdateien könnten damit effektiv wieder zusammengeführt werden. In einer erweiterten Variante sollten hier auch einzelne Wort- oder Buchstabenänderungen sichtbar gemacht werden können. Diverse Tools, u.a. aus der Softwareentwicklung (z.B. Eclipse IDE, TextPad, Emacs, etc.), liefern Teile der entsprechenden Funktionalitäten mit. Diese sollten wenn möglich wiederverwendet werden.

Anforderungsreferenz

Die Datenvergleichswerkzeuge adressieren unterschiedlichste Aspekte von Anforderungen an die virtuelle Arbeitsumgebung. Sie dienen vor allem der kollaborativen Arbeit. Dennoch sind sie durch keine der Anforderungen direkt gefordert. Sie werden jedoch, basierend auf Erfahrungen aus der kollaborativen Softwareentwicklung, bei der kollaborativen Syntaxentwicklung sowie, basierend auf Erfahrungen aus dem Datenmanagement, beim Arbeiten mit Dateien benötigt.

Arbeitsaufgaben

- Implementierung eines Vergleichs von Dateien auf Identität
- Implementierung eines Ver- und Abgleichs von Metadaten zweier Dateien oder Kategorien
- Integration und gegebenenfalls Erweiterung existierender Vergleichseditoren für Textdateien in die virtuelle Arbeitsumgebung

5.2.6 Datenverarbeitung und benötigte Werkzeuge

Die Datenverarbeitung ist einer der zentralen Punkte in der quantitativ-empirischen sozialwissenschaftlichen Forschung. Daher soll dafür, neben der Verwaltung der Daten und Syntaxdateien, zumindest eine grundlegende Unterstützung in der virtuellen Arbeitsumgebung vorgesehen werden.

Funktionsweise

Wie bereits erwähnt, werden zur Datenverarbeitung unterschiedlichste Statistik-Programme eingesetzt. Der Ablauf ist jedoch bei allen ähnlich: Es werden Syntaxdateien erstellt, welche dann von den Statistik-Programmen auf den zu untersuchenden Datensätzen ausgeführt werden. Je nach verwendeter Software kann eine solche Verarbeitung auch auf Kommandozeilenebene angestoßen werden.

Hauptsächlich werden solche Berechnungen derzeit lokal, unter Verwendung von vorhandenen Statistik-Programmen auf Workstations ausgeführt. Hier müssen Programm, Syntaxdatei und Datensatz lokal vorhanden sein. Eine Bearbeitung des gleichen Datensatzes oder der Syntaxdatei auf anderen Workstations führt zu vielfältig vorhandenen Duplikaten ohne einfache Abgleichmöglichkeiten. Hier bietet die virtuelle Arbeitsumgebung mit dem Daten-Management und den Vergleichswerkzeugen wesentliche Arbeitserleichterungen. Hinzu kommt, dass durch ein von vielen Arbeitsplätzen zugängliches Daten-Repository auch das lokale Ausführen von Berechnungen durch einfachen Zugriff auf den gewünschten Datensatz erleichtert wird. In Folge der verbesserten Forschungsdatenversorgung durch die virtuelle Arbeitsumgebung erweitert dies die Möglichkeiten der lokalen Datenverarbeitungsaufgaben für die Wissenschaftler. Unter Einsatz der virtuellen Arbeitsumgebung auf der lokalen Workstation erhalten die einzelnen Wissenschaftler eine optimale Versorgung mit Forschungsdaten als Grundlage der Datenverarbeitung mit den verwendeten Statistik-Anwendungen SPSS, Stata und SAS.

Durch die Sicherheitsinfrastruktur der virtuellen Arbeitsumgebung kann ggf. auch der Prozess des Fernrechnens in die vereinheitlichte Oberfläche einbezogen werden. Die auf der Verwendung von Zertifikaten beruhende Authentifizierung und Verschlüsselung des Datenverkehrs wird von einigen FDZ als eine ausreichende Voraussetzung für eine Implementierung eines solchen Zugangs angesehen. Eine Implementierung dieser Komponente kann jedoch nur in sehr enger Kooperation mit den FDZ realisiert werden und ist daher als ein gemeinsames zusätzliches Projekt anzustreben.

Nicht zuletzt kann die virtuelle Arbeitsumgebung durch die Nutzung der vom D-Grid zur Verfügung gestellten Infrastruktur unterschiedlichste Berechnungen auf dafür vorgesehenen und eingerichteten Rechnerressourcen ausführen, wenn die lokal vorhandenen Kapazitäten nicht mehr ausreichend sind.

In einem ersten Schritt müssen die dafür benötigten Statistik-Pakete auf den Ressourcen des D-Grid installiert werden. Hierbei gilt es, die Lizenzbestimmungen der Statistik-Pakete sowie deren Fähigkeiten zum computergesteuerten Starten von Berechnungen zu beachten. Als einfachste Variante bietet sich hier das Statistik-Paket R sowohl hinsichtlich der Lizenzen, als auch der kommandozeilenbasierten Ausführung von Berechnungen an. Die Installation weiterer Statistik-Programme sollte in einem nächsten Schritt erwogen werden.

Um statistische Berechnungen im Grid aus einer virtuellen Arbeitsumgebung heraus anstoßen zu können, müssen entsprechende Jobs an die Grid-Middleware geschickt werden. Dies sollte für den Anwender transparent passieren, da die Spezifikation solcher Jobs bestimmtes Fachwissen erfordert und in einem abgeschlossenen Kontext, wie dem der virtuellen Arbeitsumgebung, sich je Job nur geringfügig unterscheiden wird. Die virtuelle Arbeitsumgebung muss demnach in einem Teil ihrer Oberfläche eine einfach zu nutzende Möglichkeit bieten, um eine solche Verarbeitung anzustoßen. Im Hintergrund wird dabei eine Jobbeschreibung erstellt und an die Grid-Middleware gesendet. Die graphische Oberfläche muss danach eine Überwachung der Verarbeitung ermöglichen und den Benutzer über die Fertigstellung der Berechnung, z.B. per Anzeige auf dem Bildschirm oder per E-Mail, informieren. Außerdem sollen die Ergebnisse direkt zugreifbar zur Verfügung gestellt werden.

Die Erstellung einer Jobbeschreibung muss gegebenenfalls auf Inhalte der Syntaxdatei zurückgreifen. Der Grund ist, dass die Syntaxdateien Referenzen auf die Ein- und Ausgangsdaten einer Berechnung

beinhalten. Diese werden ebenso innerhalb einer Jobbeschreibung benötigt, um dem Grid die Möglichkeit zum so genannten Staging, also dem zur Verfügung stellen der benötigten Dateien auf dem Zielsystem, zu geben. Dabei müssen die Pfade zu den Dateien gegebenenfalls auf Grid-interne Pfade angepasst werden. All dies soll wenn möglich transparent für den Benutzer passieren.

Im Idealfall ist der Unterschied zwischen lokaler und Grid-basierter Berechnung nur in wenigen Punkten, z.B. in nur einem Flag beim Starten der Berechnung, sichtbar.

Eine mögliche Umsetzung für die Dialoge zum Starten und Überwachen von Berechnungen ist in Anhang A.9 beschrieben.

Anforderungsreferenz

Die von der Datenverarbeitung und den dazugehörigen Werkzeugen adressierten Anforderungen sind Anforderung 71, Anforderung 72, Anforderung 73 und Anforderung 74.

Arbeitsaufgaben

- Implementierung einer graphischen Oberfläche zum lokalen oder Grid-basierten Starten einer Berechnung
    - Implementierung des lokalen Startens von Berechnungen
    - Einbindung einer Dateiabholung aus dem Repository für lokale Berechnungen
    - Implementierung einer Job-Abwicklung über das Grid
    - Implementierung der Nutzung von Fernrechnen
- Installation von R auf den dedizierten Grid Ressourcen

5.2.7 Kollaborationswerkzeuge und Kommunikation

Ein wichtiger Aspekt der virtuellen Arbeitsumgebung ist die Unterstützung der Teamarbeit in sozialwissenschaftlichen Projekten. Hierbei ist die räumliche Trennung der Projektmitglieder zu beachten. Dies wird umso wichtiger, je weitreichender ein Projektkontext ist. Während in einem deutschlandweiten Projekt kurze und schnelle Dienstreisen für Besprechungen und Workshops finanziell und zeitlich oft noch sinnvoll sind, trifft dies für europaweite Projekte nicht mehr unbedingt zu. Ziel ist es daher, die Wissenschaftler mit geeigneten Mitteln trotz der dazwischen liegenden Distanz so nah wie möglich zusammen zu bringen und eine effektive gemeinsame Arbeit an dem Projekt zu ermöglichen.

Funktionsweise

Kommunikation und gemeinsame Dokumentation sind zwei tragende Säulen guter Teamarbeit. Hierdurch wird sichergestellt, dass alle Projektmitglieder zu jeder Zeit über die vergangenen, aktuellen und zukünftigen Arbeiten der jeweils anderen Partner informiert sind. Weiterhin werden Ziele noch effektiver und effizienter erreicht, wenn in kleinen Untergruppen gemeinsam an der gleiche Aufgabe gearbeitet wird. Genau für diese Tätigkeiten sollten passende und leicht einsetzbare Werkzeuge mit in die virtuelle Arbeitsumgebung aufgenommen werden.

Die Kommunikation in Teams, die am gleichen Ort arbeiten, erfolgt üblicherweise über Gespräche. Allerdings werden hier aus organisatorischen Gründen Termine vereinbart oder auch e-Mails verschickt, da nicht jeder Kollege immer Zeit für ein Gespräch hat. Bei der räumlichen Trennung von Teams ist dies genauso. Lediglich das Medium, mit dem Gespräche geführt werden können, muss ein

anderes sein. Es ist sicherlich nicht Aufgabe der virtuellen Arbeitsumgebung Telefongespräche oder Videokonferenzen zu ermöglichen, zumal es hierfür bereits professionelle Lösungen gibt. Jedoch sollten Technologien, wie Instant-Messaging, Voice over IP oder Videotelefonie eingesetzt werden können. Hierfür bieten sich existierende Systeme, wie etwa Skype, für die Integration in eine virtuelle Forschungsumgebung an.

Eine Besonderheit von Gesprächen im selben Raum ist die Möglichkeit des Materialaustausches. Es können Präsentationen vorgeführt oder Ausdrucke und Grafiken verteilt werden. Die Gestiken des Redners helfen den Zuhörern beim Verstehen. Auch hierfür gibt es bereits Systeme, wie z.B. Adobe® Acrobat® Connect™ Pro, die Teil einer virtuellen Arbeitsumgebung sein können. Der gemeinsame Zugriff auf zentral gelagerte Materialien kann außerdem über die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung stattfinden.

Zur Dokumentation der Projektarbeit werden üblicherweise zentrale Systeme benötigt, in denen von jedem Teammitglied Informationen abgelegt und abgerufen werden können. Was mit einem Dokumentenschrank mit vielen Ordnern in der realen Welt möglich ist, kann in der digitalen Welt durch Wikis oder zentrale Dateiverwaltungssysteme ersetzt werden. Letzteres wird durch die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung bereits abgedeckt. Wikis sollten zusätzlich integriert werden. Wichtig ist jedoch, sowohl bei einem Dokumentenschrank, als auch bei Wikis oder anderen Dateiverwaltungssystemen, dass es teamweite Vorgaben für deren Ordnung geben muss, und dass diese Vorgaben eingehalten und deren Einhaltung überprüft werden.

Ein letzter Punkt der lokalen Zusammenarbeit ist das gemeinsame Bearbeiten vorliegender Materialien, Dokumente oder Dateien. Es ist möglich, sich den gleichen Ausdruck oder das gleiche digitale Dokument auf einem Computer gemeinsam anzusehen und auch zu bearbeiten. Bei einer räumlichen Trennung ist dies wesentlich schwerer. Üblicherweise wird hier quasiparallel gearbeitet, in dem eine Person eine Änderung vornimmt, diese an die anderen Personen weiterleitet, dann eine Diskussion zu der Änderung stattfindet, weitere Änderungen entstehen und diese wieder an alle weitergeleitet werden. Zum einen ist dieses Vorgehen aufwändig und zum anderen verlangt es sehr viel Disziplin aller Teilnehmer.

In einer virtuellen Arbeitsumgebung sollte das gemeinsame Bearbeiten von zumindest Syntaxdateien trotz räumlicher Trennung möglich sein. Hierzu sollten Syntaxdateien in einem eigens dafür vorgesehenen Editor geöffnet werden können. Jeder Teilnehmer, der diesen Editor benutzt, sieht die gleiche Syntaxdatei und kann eigene Änderungen vornehmen, oder Änderungen der anderen Teilnehmer sehen und verfolgen. Unter paralleler Nutzung eines Kommunikationssystems, wie z.B. Skype entsteht somit eine Zusammenarbeit, die einer lokalen, gemeinschaftlichen Arbeit sehr ähnlich ist. Daher sollte ein solcher Editor mit in die virtuelle Arbeitsumgebung aufgenommen werden. Eine mögliche Implementierung der entsprechenden Oberfläche ist rudimentär in Anhang A.10 beschrieben.

Eine quasiparallele Bearbeitung sollte jedoch weiterhin über den Dateiaustausch mit Hilfe der Datenverwaltungswerkzeuge möglich sein.

Für einige dieser Funktionalitäten werden entsprechende Komponenten im Systemkern benötigt. So sollten z.B. ein für ein Wiki notwendiger Webserver oder ein im Hintergrund arbeitendes System zur kollaborativen Syntaxbearbeitung installiert und, falls nicht bereits vorhanden, zunächst implementiert werden.

Anforderungsreferenz

Die in diesem Abschnitt beschriebenen Werkzeuge zur entfernten Teamarbeit sind die Umsetzungen von Anforderung 75 bis Anforderung 79.

Arbeitsaufgaben

- Integration einer Lösung für Instant-Messaging und Videotelefonie
- Installation und Konfiguration des Intranets
- Festlegung von Nutzungsvorgaben für das Intranet
- Klärung der Verwaltung von Servern und Installation für die im Systemkern notwendigen Funktionalitäten, je nachdem, welche Komponenten ausgewählt werden
- Konzeption, Implementierung und Installation eines kollaborativen Syntaxeditors

5.2.8 Schnittstelle zu Datenanbietern

Damit ein Forschungsprozess beginnen kann, werden in der virtuellen Arbeitsumgebung Werkzeuge für den Zugriff auf Originaldatensätze benötigt. Die Originaldatensätze werden in verschiedener Form von den Datenanbietern zur Verfügung gestellt. Des Weiteren bieten die FDZ teilweise an, über Fernrechnen direkt Auswertungsprogramme auf Datensätzen zu auszuführen. Drittens ermöglichen einige FDZ für eigene Datensätze auch die Archivierung von Resultaten solcher Auswertungen.

Die virtuelle Arbeitsumgebung sollte – unabhängig von den unterschiedlichen Verfahren für den Erwerb der Ausgangsdaten - mit Werkzeugen ausgestattet sein, die Ausgangsdaten in geeigneter Form einbringen zu können. Im Folgenden werden daher mögliche Schnittstellen vorgestellt, den Prozess des Erwerbs der Ausgangsdaten zu erleichtern.

Funktionsweise

Die virtuelle Arbeitsumgebung kann einen direkten und einheitlichen Zugriff auf online verfügbare Datensätze bereitstellen. Hierbei sollte zunächst anhand von Kriterien nach Datensätzen gesucht werden können. Diese Suche sollte über verschiedene Datenanbieter gleichzeitig ausgeführt werden können. Ein Suchergebnis sollte dann über einfache Methoden direkt mit Hilfe der Datenverwaltungswerkzeuge in die virtuelle Arbeitsumgebung übertragen werden können. Dabei sollte die virtuelle Arbeitsumgebung automatisch vorhandene Metadaten in die eigene Metadatenverwaltung übernehmen und diese gegebenenfalls mit automatisch erzeugbaren Informationen anreichern[80].

Am Ende einer Forschung muss das gesamte verwendete Datenmaterial im Rahmen der guten wissenschaftlichen Praxis aufbewahrt werden. Wie bereits mehrfach erwähnt, ist dies aufgrund von Nutzungsrechten für Original- und Arbeitsdatensätze nur begrenzt möglich. Dies wurde von den Datenanbietern erkannt. Daher können Wissenschaftler mittlerweile von einigen Angeboten Gebrauch machen, und ihre Daten direkt beim Datenanbieter archivieren. Damit werden keine Nutzungsrechte verletzt und die Forschungsdokumentation bleibt vollständig erhalten.

Die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung sollten, wie oben beschrieben, den Wissenschaftler bei der Einhaltung von Nutzungsfristen unterstützen. Dabei soll unter anderem das

[80] Eine solche Suche kann auf Basis von gemeinsam genutzten Metadaten sinnvoll implementiert werden.

Löschen der Daten am Ende des Nutzungszeitraums vorgeschlagen werden. Dabei ist es durch Dienstleistungen mancher Datenanbieter möglich, die Daten vor der Löschung direkt an das FDZ zur Archivierung zu übertragen. In der virtuellen Arbeitsumgebung werden die Daten weiterhin als virtuelle Datensätze, mit einem Verweis auf den Archivierungsstandort, erhalten bleiben.

Für solche Verfahren wird es notwendig, gemeinsam mit den FDZ geeignete Schnittstellen zu entwickeln. Die schon vorhandenen Überlegungen und Angebote können langfristig auch auf Basis der Sicherheitsinfrastruktur der virtuellen Arbeitsumgebung besser integriert werden.

Anforderungsreferenz

Die von den hier beschriebenen Werkzeugen umgesetzten Anforderungen sind Anforderung 2, Anforderung 3 und Anforderung 4.

Arbeitsaufgaben

- Erarbeitung eines gemeinsamen Metadatenformats (in enger Abstimmung mit den FDZ) für die Suche
- Implementierung einer Suchfunktion, die auch verschiedene Datenanbieter einbezieht
- Ergänzung der Datenverwaltungswerkzeuge um die Funktionalität zur Archivierung von Daten bei den Datenanbietern

5.2.9 Konfigurations- und Verwaltungswerkzeuge

Je komplexer Softwaresysteme werden, umso größer werden auch die Anforderungen hinsichtlich Konfigurierbarkeit. Daher sollte die virtuelle Arbeitsumgebung Werkzeuge zur Verfügung stellen, mit denen unterschiedlichste benutzerdefinierte Einstellungen vorgenommen und abgelegt, sowie nutzerübergreifende Festlegungen vorgenommen werden können.

Funktionsweise

Über die Konfigurationswerkzeuge sollte der Benutzer seine eigenen Einstellungen der virtuellen Arbeitsumgebung vornehmen und abspeichern können. Solche Einstellungen umfassen unter anderem

- das Erscheinungsbild der graphischen Oberfläche,
- die Konfiguration der Systeminstallation, z.B. mit Pfaden zu Verzeichnissen oder Hyperlinks für spezielle Webverbindungen und
- die beeinflussbaren Verhaltensweisen und -muster des Systems.

Denkbar wären demnach u.a. Farbeinstellungen, die Angabe eines temporären Verzeichnisses, die Verbindung zu einem Server oder das Festlegen von so genannten Shortcuts (Tastenkürzel).

Aufgrund des frühen Entwicklungsstadiums kann an dieser Stelle nicht genau aufgelistet werden, welche Parameter mit welchen Werten einstellbar sein sollten. Dafür sind noch zu wenige Systemaspekte und -details bekannt. Wichtig ist jedoch, dass die Entwicklung entsprechender Werkzeuge vorgesehen werden muss. Außerdem muss jede Komponente der virtuellen Arbeitsumgebung Zugriff auf die Konfiguration haben, um sich entsprechend den Einstellungen des Benutzers zu verhalten. Hier ist der gegebenenfalls höhere Implementierungsaufwand für die Komponenten zu berücksichtigen, damit diese die Konfigurationseinstellungen umsetzen.

Neben der anwenderspezifischen Konfiguration gibt es notwendige Verwaltungsmaßnahmen und -einstellungen, die für mehrere Benutzer der virtuellen Arbeitsumgebung, meist für eine bestimmte Benutzergruppe, vorgenommen werden. Beispiele hierfür sind

- die Nutzerrechte, die für verschiedene Funktionen und Zugriffe vergeben werden,
- strukturelle Vorgaben, die von den Komponenten der virtuellen Arbeitsumgebung umzusetzen sind, wie z.B. Vorgaben zum Aufbau von Kategorienhierarchien für Dateien und
- grundlegende Metadatenstrukturen für unterschiedliche Dateitypen.

Für derartige Funktionalitäten sollten Verwaltungswerkzeuge in der virtuellen Arbeitsumgebung vorgesehen werden. Diese erlauben entsprechende Einstellungen vorzunehmen, zu managen und gegebenenfalls mit anderen Projektteilnehmern abzustimmen.

Auch bei den Verwaltungseinstellungen können so früh im Entwicklungsstadium nur wenige Aussagen über deren konkrete Ausprägung gemacht werden. Dennoch müssen auch sie bereits vorgesehen werden. Außerdem sollte bei der Entwicklung der anderen Komponenten bereits berücksichtigt werden, dass gegebenenfalls Verwaltungseinstellungen für eine Komponente beachtet werden müssen.

Anforderungsreferenz

Da die Konfigurations- und Verwaltungswerkzeuge Auswirkungen auf die gesamte Funktionalität der virtuellen Arbeitsumgebung haben werden, gibt es keine spezielle Anforderung, die von ihnen direkt adressiert wird. Eher sind sie Teil der Umsetzung einer jeden spezifizierten Anforderung an die virtuelle Arbeitsumgebung. Auf eine genaue Angabe der Einzelanforderungen wird daher verzichtet.

Arbeitsaufgaben

- Entwicklung eines lokalen Konfigurationssystems
- Entwicklung der Werkzeuge zur Konfiguration
- Beachtung der Konfiguration in den einzelnen Komponenten der virtuellen Arbeitsumgebung
- Entwicklung eines globalen Verwaltungssystems
- Entwicklung der Werkzeuge zur Verwaltung der virtuellen Arbeitsumgebung
- Beachtung der Verwaltungseinstellungen in den einzelnen Komponenten der virtuellen Arbeitsumgebung

5.2.10 Publikationswerkzeuge und sonstige Dienste

Am Ende des Forschungsprozesses werden die Ergebnisse veröffentlicht. Auch hierfür sind unterschiedlichste Werkzeuge und Komponenten notwendig.

Funktionsweise

Mit Hilfe der Publikationswerkzeuge soll der Wissenschaftler bei der Veröffentlichung seiner Forschungsergebnisse unterstützt werden. Einige der bisher beschriebenen Tools können hierfür Verwendung finden. Soll z.B. eine Grafik veröffentlicht werden und muss diese vorher in ein anderes Format umgewandelt werden, so bieten sich die Konvertierungstools der Datenbearbeitungswerkzeuge an.

Die virtuelle Arbeitsumgebung sollte jedoch darüber hinaus über Funktionen verfügen, mit denen Publikationen zusammen- und zur Verfügung gestellt werden können. Allerdings gibt es hierbei ein sehr

breites Spektrum an möglichen Verfahren. Während der eine Wissenschaftler Dokumente lieber mit einem Textverarbeitungsprogramm verfasst und das resultierende Dokument über einen Verlag verbreiten lässt, möchte ein anderer seine Ergebnisse direkt auf einer Webseite präsentieren. Die virtuelle Arbeitsumgebung kann an dieser Stelle nicht für jeden erdenklichen Weg Unterstützung bieten. Allerdings sollte sie für ausgewählte Varianten die passenden Werkzeuge bereithalten, insbesondere für die Veröffentlichung von Informationen über Projektwebseiten.

Nahezu jedes Forschungsprojekt pflegt eine Internetpräsenz, um über dessen Vorhaben, Fortschritte oder entsprechende Ergebnisse zu informieren (s.a. http://www.soeb.de/). Die Publikationswerkzeuge der virtuellen Arbeitsumgebung sollten Zugriffe auf jene Systeme bieten, mit denen die Projektwebsite aufgebaut und gepflegt wird. Handelt es sich hierbei um Content-Management-Systeme, so muss Zugriff auf deren Konfigurationsoberfläche bestehen. Sind es aber eher Wikis, so reicht ein Zugang zur entsprechenden Einstiegsseite. Die Zugriffe müssen dabei einen einfachen Transfer von bereits erstellten Teilen der Veröffentlichung in die Systeme zur Websiteverwaltung erlauben. Demnach sollte es einfach möglich sein, z.B. eine Grafik oder einen Text über die Datenverwaltungswerkzeuge aus der virtuellen Arbeitsumgebung herauszuholen und über die Verwaltung der Projektwebsite zu veröffentlichen. Außerdem sollten Reviewfunktionalitäten der eingesetzten Websitesysteme, sofern sie von diesen angeboten werden, aus der virtuellen Arbeitsumgebung heraus nutzbar sein. Hierdurch kann sichergestellt werden, dass mehrere Mitglieder eines Forschungsverbunds eine geplante Publikation vor deren tatsächlicher Veröffentlichung verifizieren.

Im Rahmen guter wissenschaftlicher Praxis sollten Datensätze und Syntaxdateien, die als Basis eines Forschungsprozesses gedient haben, langzeitverfügbar abgelegt und in den entsprechenden Publikationen eindeutig referenziert werden. Vor diesem Hintergrund sind die Datenablage, die Datenorganisation und die Datenverwaltungswerkzeuge der virtuellen Arbeitsumgebung ebenso als eine Art der Publikationswerkzeuge zu verstehen. Je nach Nutzung und detaillierter Ausimplementierung stellen sie ein zitierbares Langzeitdatenrepositorium dar.

Für Projektwebseiten, wie auch für andere Projekt-Dienste, werden entsprechende Server benötigt. Diese und andere Systeme, werden unter dem Begriff „sonstige Dienste“ zusammengefasst, welche Teil des Systemkerns sind. Hierunter fallen auch weitere Komponenten. Sollen z.B. Verwaltungseinstellungen abgelegt werden, die über die Nutzerverwaltung hinausgehen, so müssen Systeme bereitstehen, die diese Konfigurationsinformationen aufnehmen und den anderen Komponenten der virtuellen Arbeitsumgebung zur Verfügung stellen. Weiterhin werden gegebenenfalls zusätzlich Dienste, z.B. zum Überwachen von Verarbeitungsstatus oder für in dieser Expertise noch nicht betrachtete, aber in Zukunft implementierte Funktionen benötigt. Auch hier ergibt sich wieder ein sehr breites Spektrum, dessen vollständige Aufzählung und Beschreibung den Rahmen der Expertise sprengen würde.

Anforderungsreferenz

Mit den Publikationswerkzeugen wird direkt Anforderung 80 umgesetzt. Die sonstigen Dienste haben, ähnlich den Konfigurations- und Verwaltungswerkzeugen, eine Querschnittsaufgabe und adressieren daher nahezu alle Anforderungen aus Abschnitt 3.

Arbeitsaufgaben

- Gemeinsame Erstellung und Nutzung der Projekt-Website
    - Auswahl eines geeigneten CMS, Klärung von Look & Feel, Design usw.
    - Festlegung / Implementierung von Workflows für die gemeinsame Redaktion

- Bereitstellung und Maintenance von Servern, die Dienste der virtuelle Arbeitsumgebung bereitstellen
- Entwicklung oder Installation sonstiger Dienste

5.2.11 Aufwandsabschätzung

Die Aufwandabschätzung gehört zu den schwierigsten Aufgaben einer Projektvorbereitung. Abhängig von der Güte der Schätzung variiert deren Verlässlichkeit und Genauigkeit teils sehr stark. Dennoch werden die Zahlen oft als tatsächliche Werte angesehen und fest in die Projektplanung übernommen. Dies kann bei Fehlschätzungen von Aufwänden allerdings auch zum Scheitern eines Projektes führen. Es ist daher wichtig, die Werte von Schätzungen auch als solche anzusehen und entweder mit zeitlichen Puffern im Projekt zu arbeiten oder gegebenenfalls eine Nachfinanzierung vorzusehen.

Es gibt unterschiedliche Herangehensweisen für Aufwandsabschätzungen. Allen gemein ist jedoch, dass die Schätzungen umso genauer werden, je besser die Details des zu entwickelnden Systems sowie des Einsatzgebietes bekannt sind. Außerdem spielt die Erfahrung des Schätzers eine große Rolle. Unbekannte Komponenten der Schätzung sind jedoch unerwartete Probleme und die konkreten Fähigkeiten der Entwickler. So ist der Aufwand zum Entwickeln einer Funktionalität für jemanden ohne Erfahrung sowohl im Anwendungskontext, als auch im Bereich der eingesetzten Technologien wesentlich größer, als für jemanden, der bereits seit Jahren in dem Kontext tätig ist[81].

Die in dieser Expertise gemachten Aufwandsabschätzungen versuchen die Erfahrungen der Expertisensersteller, die erarbeiteten Einzelkomponenten und deren Funktionalitäten sowie die unbekannten Faktoren bestmöglich mit einander in Beziehung zu setzen, so dass die geschätzten Aufwände eine bestmögliche Validität aufweisen. Es sei dennoch darauf hingewiesen, dass die gelisteten Werte Schätzungen mit einer gegebenenfalls großen Abweichung zum tatsächlichen Aufwand darstellen.

Für die Schätzungen werden zu erledigende Einzelaufgaben ermittelt und deren Aufwand aus verschiedenen Perspektiven betrachtet. Dazu zählen die Erstellung detaillierter Konzepte und Designs der Funktionalitäten, deren Umsetzung, die gegebenenfalls notwendige Kommunikation, das Testen, die Dokumentation sowie die Überführung in den Betrieb. Die Einzelaufwände werden dann zu Gesamtaufwänden addiert und gegebenenfalls mit einem Risikofaktor multipliziert. Der Risikofaktor soll dabei bewerten, wie groß die Wahrscheinlichkeit für unerwartete Probleme bei der Umsetzung einer Funktionalität ist. Die Darstellung erfolgt in Form einer Tabelle, welche die einzelnen Funktionen, die kategorisierten Aufwände und deren Summe beinhaltet.

Die zu erwartenden, geschätzten Aufwände hinsichtlich aller Funktionalitäten in der virtuellen Arbeitsumgebung sind in Anlage B (gesondertes Dokument[82]) aufgelistet. Darin sind außerdem Aufwände für das Projektmanagement sowie die Erstellung eines gemeinsamen technologischen und optischen Rahmens für die einzelnen Komponenten des Systems enthalten.

---

[81] Vgl. hierzu http://www.stefan-baur.de/cs.se.aufwand.html. Letzter Zugriff 21.07.2010.

[82] Siehe AnlageB_GridExpertiseVirtAug_v10.xls. Dieses Dokument wird gesondert und mit anderen Zugriffsrechten publiziert, da nicht alle enthaltenen Informationen öffentlich zugänglich sein sollten und dürfen.

6 Exemplarische Arbeitsabläufe anhand der Architekturskizze

Im Kontext der in Abschnitt 2.1 definierten Anwendungsszenarien und der in Abschnitt 5 spezifizierten Architekturskizze werden im Folgenden exemplarische Arbeitsabläufe vorgestellt. Das Ziel ist es dabei, die Arbeitsweise mit der virtuellen Arbeitsumgebung aus den Perspektiven eines einzelnen Forschers und eines themenbezogenen, standortübergreifenden Forschungsverbundes sowie der FDZ darzustellen. Der Onsite-Zugriff auf die Ausgangsdaten und die Ablage der Ausgangsdaten werden, da aus allen drei Perspektiven nahezu identisch, übergreifend in Abschnitt 6.1 dargestellt.

Die beschriebenen Vorgehen stellen nur einen Ausschnitt der möglichen Prozesse exemplarisch dar. Eine detaillierte Beschreibung aller möglichen Nutzungsmöglichkeiten ist aufgrund der Komplexität an dieser Stelle nicht möglich.

6.1 Onsite-Zugriff auf Ausgangsdaten

Im Kontext der Onsite-Nutzung kann der Zugriff auf die Ausgangsdaten, wie in Abschnitt 2.1 beschrieben, nur vor Ort bei dem jeweiligen FDZ erfolgen. In Folge dessen werden der einzelne Forscher, die Forschergruppe oder der Forschungsverbund entsprechend die Datenselektion und eine erste Datenaufbereitung für die benötigten Zwischenergebnisse direkt beim FDZ (und unter dessen Aufsicht) durchführen. Die Vorbereitung zur Datenselektion, wenn möglich auf Grundlage von Dummy-Datensätzen (bereitgestellt durch das entsprechende FDZ), erfolgt innerhalb der virtuellen Arbeitsumgebung. Hier kann von dem jeweiligen Wissenschaftler die Datenselektion getestet werden. Ebenso kann der Wissenschaftler idealerweise im Intranet der virtuellen Arbeitsumgebung bereits verfügbare Informationen zu der Datenstruktur des jeweiligen FDZ entnehmen, um den Vorbereitungsprozess zu optimieren. Abschließend werden die zusammengestellten Zwischenergebnisse physisch zum Arbeitsort des/der Wissenschaftler(s) transportiert.

Werden seitens des FDZ keine Dummy-Datensätze bereitgestellt, ist der Wissenschaftler gehalten, die Vorbereitung in enger Abstimmung mit dem FDZ und erfahrenen Kollegen durchzuführen. Die Abstimmung mit Kollegen wird durch die virtuelle Forschungsumgebung durch die Kollaborationswerkzeuge unterstützt. Der Wissenschaftler kann einen Termin mit einem erfahrenen Kollegen vereinbaren und die Abstimmung findet dann mittels Videokonferenz in der virtuellen Arbeitsumgebung statt.

Die virtuelle Arbeitsumgebung nimmt diese OnSite erzeugten Ergebnisdaten auf und referenziert die beim Onsite-Rechnen genutzten Daten geeignet.

6.2 Perspektive des Wissenschaftlers

6.2.1 Forschungsdatenablage

Ausgangspunkt ist ein Einzelwissenschaftler, welcher eine dedizierte Fragestellung bearbeitet. Dies kann im Rahmen einer Forschergruppe oder eines Forschungsverbundes stattfinden. In einem ersten Schritt werden die Ausgangsdaten (SUF oder Zwischenergebnisse der Onsite-Nutzung[83]) in der virtuellen Arbeitsumgebung abgelegt. Dazu meldet sich der Einzelwissenschaftler an der virtuellen Ar-

[83] Bezüglich Ablage und Verknüpfung mit den für die Onsite-Nutzung notwendigen Syntaxdaten sei auf Abschnitt 6.2.2 verwiesen.

beitsumgebung an. Aufgrund der Verwendung der PKI für die virtuelle Arbeitsumgebung benutzt der Wissenschaftler sein persönliches Zertifikat für die Anmeldung. Die virtuelle Arbeitsumgebung stattet den Wissenschaftler, in Form der Autorisierung, mit allen ihm zur Verfügung stehenden Berechtigungen zur Nutzung der virtuellen Arbeitsumgebung aus. Dies geschieht für den Wissenschaftler völlig transparent.

Anschließend navigiert der Wissenschaftler zu dem Bereich für die Forschungsdatenverwaltung (Datenverwaltungswerkzeuge) und wählt die Funktion zur Ablage neuer Forschungsdaten aus. Beim Ablegen werden automatisch das Dateiformat sowie relevante technische Informationen zu den Daten bestimmt und zur Verifikation angezeigt. Gegebenenfalls können diese Informationen durch den Wissenschaftler korrigiert werden. Im gleichen Schritt wird der Forscher gebeten fachliche Metadaten in einer dafür vorgesehenen Eingabemaske einzugeben. Der Forscher beschreibt dabei die Herkunft der Daten, ihren fachlichen Inhalt sowie den Kontext, in dem die Daten verwendet werden. Im Zusammenhang mit der Herkunft der Daten wird ein möglicher automatischer Löschzeitpunkt definiert, um die Anforderungen des bereitstellenden FDZ bezüglich der maximalen Aufbewahrungsdauer zu erfüllen.

Des Weiteren können die Zugriffsberechtigungen festgelegt werden. Als Standard wird dem Wissenschaftler der auf seine Person eingeschränkte Einzelzugriff vorgeschlagen. Hier kann der Wissenschaftler weitere Berechtigungen auf der Basis weiterer einzelner Personen oder ganzer Arbeitsgruppen bzw. Rollen auswählen. Für jede zusätzliche Berechtigung kann der Zugriffsmodus spezifiziert werden: Nur-Lesen, Lesen und Schreiben, Vollzugriff inkl. Löschen. Zusätzlich werden mögliche dedizierte Speicherorte/Speicheranbieter der virtuellen Arbeitsumgebung angeboten, die der Wissenschaftler bei Bedarf auswählen kann.

Abschließend bestätigt der Wissenschaftler die eingegebenen Daten und leitet damit den Speicherprozess der Forschungsdaten in die virtuelle Arbeitsumgebung ein.

6.2.2 Forschungsprozess

Im Anschluss an die Ablage der Forschungsdaten widmet sich der Forscher der Erstellung von Syntaxdaten zur Datenbearbeitung und -analyse. Dementsprechend verwendet er seinen lokalen Editor zum Erstellen der Syntax. Im Rahmen der Erstellung der Syntaxdaten werden Kommentierungsvorgaben als integrierte fachliche Metadaten in der Syntax hinterlegt.

Nach Fertigstellung der Syntaxdaten verwendet der Wissenschaftler erneut die Datenverarbeitungswerkzeuge der virtuellen Arbeitsumgebung. Der Speicherungsprozess verläuft analog zu den Forschungsdaten. Zusätzlich wird der Wissenschaftler um Bestätigung bzw. Korrektur der Informationen zur Verknüpfung der Syntaxdateien mit bereits abgelegten Forschungsdaten gebeten. Letzteres ist insbesondere dann notwendig, wenn Forschungsdaten und Syntaxdateien getrennt gespeichert werden.

Zur Datenaufbereitung wählt der Wissenschaftler in den Datenverarbeitungswerkzeugen die zuvor abgelegten Forschungs- und Syntaxdaten aus. Für die Bearbeitung und -analyse mittels SPSS, SAS oder Stata werden die Daten auf die lokale Workstation heruntergeladen. Sollen die Daten mit R verarbeitet werden, wählt der Wissenschaftler in den Datenverarbeitungswerkzeugen die entsprechende Funktion und startet die Ausführung direkt in der virtuellen Arbeitsumgebung. Dabei kann eine Emailadresse hinterlegt werden, an die die virtuelle Arbeitsumgebung Informationen zum Status des Datenverarbeitungsauftrags senden kann. Als sofortige Rückmeldung erhält der Wissenschaftler Informationen zu dem Datenverarbeitungsauftrag inklusive des Status, ob die Ausführung erfolgreich

angestoßen werden konnte. Anschließend kann sich der Wissenschaftler anderen Aufgaben widmen, bis er eine Email über die abgeschlossene Ausführung seines Datenverarbeitungsauftrags erhält.

Um zur weiteren Verfeinerung der Ergebnisse zusätzliche relevante Forschungsdaten mit den Zwischenergebnissen aus der Datenaufbereitung zu erhalten, setzt der Wissenschaftler die Suche aus den Datenverwaltungswerkzeugen der virtuellen Arbeitsumgebung ein. Anhand einer unscharfen Suche verschafft er sich einen übergreifenden Überblick über die zur Verfügung stehenden potenziellen Forschungsdaten. Hierbei werden die Metadaten zu den Forschungsdaten durchsucht. Diese Möglichkeit kann vor allem auch dann genutzt werden, wenn zusätzliche Funktionen in die Syntax eingebaut werden sollen und diese bereits durch Kollegen implementiert wurden. Die virtuelle Arbeitsumgebung bietet dementsprechend die Suchbarkeit relevanter Syntaxdateien an.

Nach erfolgreicher Durchführung des Datenverarbeitungsauftrags erhält der Wissenschaftler über die virtuelle Arbeitsumgebung Zugriff auf die Ergebnisdaten. Diese kann er zur Prüfung direkt aus der virtuellen Arbeitsumgebung einsehen und herunterladen. Darauf aufbauend können die Zwischenergebnisse mit weiteren Daten aus der virtuellen Arbeitsumgebung verknüpft werden und die Datenauswertung durchgeführt werden. Seitens der virtuellen Arbeitsumgebung können die Ergebnisse der vorangegangenen Suchvorgänge zu Forschungsdaten verwendet werden und der Wissenschaftler selektiert einen der gefundenen Forschungsdatensätze für die Auswertung. Der Vorgang zur Datenauswertung verläuft in Bezug auf die Ausführung analog zu dem der Datenaufbereitung.

Abschließend erhält der Wissenschaftler Ergebnisdaten, die er in seinen Forschungsbericht einbindet. Der Forschungsbericht wird analog zu dem Ablegen von Forschungsdaten in der virtuellen Arbeitsumgebung gespeichert. Als Format wird der für den Arbeitskontext vereinbarte Standard (z.B. das Open Document Format) verwendet. Hierbei werden die technischen Metadaten des Dokuments verarbeitet und eine Formatvalidierung durchgeführt. Der Wissenschaftler ergänzt die Metadaten um Informationen zum Forschungshintergrund und den Autoren.

6.3 Perspektive des themenbezogenen, standortübergreifenden Forschungsverbundes

Die zu einem bestimmten Thema organisierte Gruppe von Wissenschaftlern nutzt die virtuelle Arbeitsumgebung einerseits aus der jeweils individuellen Perspektive des Einzelwissenschaftlers (siehe Abschnitt 6.2), und verwendet darüber hinaus weitere Elemente, um die gemeinsame Forschungsleistung und den Abstimmungsprozess zu optimieren. Im Rahmen eines standortübergreifenden Forschungsverbundes werden nicht nur die individuellen oder gruppenspezifischen Leistungsmerkmale der virtuellen Arbeitsumgebung benötigt. Es wird insbesondere den Kommunikationswerkzeugen hohe Bedeutung beigemessen. Durch die geografische Trennung ist die Möglichkeit ad hoc eine Kommunikation mit Partnern eines anderen Standortes durchzuführen eine wesentliche Erleichterung für den Forschungsprozess.

Nach Projektstart legen die einzelnen Wissenschaftler, soweit vorhanden und möglich, ihre Forschungsdaten, die dazugehörigen Metadaten sowie die Syntaxdaten in der virtuellen Arbeitsumgebung ab. Dazu werden durch die Projektkoordination eine Struktur in der virtuellen Forschungsumgebung bzw. Vorgaben zur Standardisierung definiert, um die Ablage der Daten möglichst homogen zu gestalten. Durch Setzen von Zugriffsberechtigungen können SUF allen berechtigten Wissenschaftlern zentral bereitgestellt werden.

Der Forschungsverbund betreibt mittels der Datenbearbeitungswerkzeuge die Entwicklung von Syntaxdateien gemeinsam und erzielt bessere Ergebnisse als bei rein individuellen Arbeiten. Darüber hin-

aus besteht mit den Datenverwaltungswerkzeugen die Möglichkeit vorhandene freigegebene Syntaxdateien zu durchsuchen und auf relevantes Material zurückzugreifen. Hierdurch wird der Forschungsprozess für den Verbund optimiert.

Auf Basis der mittels der Datenverwaltungswerkzeuge erfassten Metadaten durchsuchen die Arbeitsgruppen des Verbundes vorhandene Forschungsdaten. Damit können Sie auf eine breitere relevante Datenbasis zurückgreifen, die ohne einen strukturierten Überblick in vergleichbarer Form nicht erzielt werden kann.

Der Forschungsverbund entwickelt im Laufe des Projekts komplexe Zusammenhänge, die im Intranet mittels der vorhandenen Kollaborationswerkzeuge dokumentiert werden. Die damit verbundene Historie zu allen Änderungen unterstützt die Forscher im Hinblick auf die Transparenz des gemeinsamen Forschungsprozesses und der gemeinsamen Entwicklung wissenschaftlich wertvoller Erkenntnisse.

Der Forschungsverbund stimmt sich gemeinsam im Rahmen eines wöchentlichen Jour Fix ab, welche mittels der Kommunikationswerkzeuge der virtuellen Arbeitsumgebung realisiert werden. Das Protokoll der Abstimmungen wird im Intranet allen Partnern zur Verfügung gestellt. Die geografische Trennung spielt somit keine direkte Rolle mehr für den Fortschritt der wissenschaftlichen Arbeit.

Zum Abschluss von Projektphasen bzw. Projekten werden die erarbeiteten Ergebnisdaten sowie die verwendete Syntax mit Hilfe der Datenverwaltungswerkzeuge in der virtuellen Forschungsumgebung abgelegt. Dabei werden Verknüpfungen der Daten untereinander in den Metadaten gespeichert.

Die jeweils finalen Forschungsberichte werden ebenso in der virtuellen Arbeitsumgebung gespeichert. Das Intranet bietet dabei die Funktion an, einen strukturierten Review-Prozess für jedes der zu publizierenden Dokumente zu durchlaufen. Die an einem Bericht oder Publikation beteiligten Wissenschaftler erhalten im Rahmen des Review-Prozesses Zugriff auf das jeweilige Dokument und können Änderungen vorschlagen bzw. mit Hilfe der Kommunikationswerkzeuge über inhaltliche Fragen diskutieren. Abschließend kann das Dokument weiteren Wissenschaftlern zur Verfügung gestellt, publiziert und in der virtuellen Arbeitsumgebung abgelegt werden. Als Format für die Ablage wird der für den Arbeitskontext vereinbarte Standard (z.B. das Open Document Format) verwendet. Bei allen Speicherprozessen werden technischen Metadaten der Dokumente verarbeitet und Formatvalidierungen durchgeführt. Der jeweils verantwortliche Wissenschaftler ergänzt die Metadaten um Informationen zum Forschungshintergrund und den Autoren. Zusätzlich werden editierbare Versionen der Berichte abgelegt, um diese zu einem späteren Zeitpunkt ggf. für die Publikation einer aktuellen Version anpassen zu können.

6.4 Einsatz der virtuellen Arbeitsumgebung in Verbindung mit FDZ

Auf Grundlage der vollständigen Protokollierung von Änderungen an Daten in der virtuellen Arbeitsumgebung kann sichergestellt und verifiziert werden, dass die maximalen Aufbewahrungszeiten für die Ausgangsdaten (hier SUF) nicht überschritten werden. Die Wissenschaftler der sozioökonomischen Berichterstattung können somit gegenüber den FDZ den ordnungsgemäßen Umgang mit den Ausgangsdaten belegen.

Sofern eine Abstimmung mit den entsprechenden FDZ erreicht werden kann, lassen sich SUF über die virtuelle Arbeitsumgebung bereitstellen. Der Zugang zu bestimmten SUF via Postweg könnte somit der Vergangenheit angehören. Damit kann eine Entlastung der FDZ und der betroffenen Forscher durch den Einsatz der virtuellen Arbeitsumgebung realisiert werden.

7 Empfehlungen und Ausblick

Die obige Darlegung und Analyse der Hauptakteure in der Community (der Wissenschaftler und Institute im Umfeld der sozioökonomischen Berichterstattung), deren Arbeitsprozesse, der Rahmenbedingungen und der Anforderungen an IT basierte Infrastruktur und Dienste ermöglichte eine differenzierte Darstellung der Elemente einer virtuellen Arbeitsumgebung für diese Community.

Als Hauptpunkte der Analyse sind zu nennen: die Verbesserung und Erweiterung der kollaborativen Arbeitsmöglichkeiten der Community liegt wesentlich in einer Datenmanagement-Infrastruktur. Diese kann auf den von der Grid-Technologie des D-Grid bereitgestellten Sicherheits-, Authentifizierungs- und Autorisierungsmechanismen aufbauen und so auch wesentliche Anforderungen einer Reihe von Datenprovidern (FDZ) realisieren. Die Hereinnahme von einigen aus der Software-Entwicklung bekannten Tools bietet hervorragende Möglichkeiten, die fachliche Diskussion über Methoden und Resultate zu verbessern (Syntaxdateien). Ein konsistentes Metadaten-System, kombiniert mit dem schon weit entwickelten Archivierungs-System Fedora/iRODS, deckt die wesentlichen Anforderungen für eine Community- oder projektweite Verfügbarkeit von gemeinsam genutzten Daten unter Beachtung der rechtlichen und Urheber-rechtlichen Aspekte der Datenhaltung ab. Die Datenhaltung und -bereitstellung für die virtuelle Arbeitsumgebung kann über die vorhandene D-Grid Infrastruktur effizienter gestaltet werden. In der Ausarbeitung der einzelnen Aspekte der virtuellen Arbeitsumgebung geht die Expertise weit über den gesteckten Rahmen hinaus, die Verwendung des D-Grid als ein Kern-Element der virtuellen Arbeitsumgebung darzustellen.

Bei der Darstellung der Elemente der virtuellen Arbeitsumgebung sind alle Werkzeuge und Funktionen einer virtuellen Arbeitsumgebung genannt und IT-Lösungen dafür skizziert worden. Naturgemäß sind bei einer derartig detaillierten Beschreibung der einzelnen Komponenten keine Gewichtungen oder Priorisierungen vorgenommen worden. Die Bewertung und Einordnung der dargestellten Elemente ist jedoch notwendig, um einen Weg zur Realisierung im Rahmen der angestrebten 3. Phase der sozioökonomischen Berichterstattung aufzuzeigen. Es ist zu empfehlen, dass keine Trennung der technischen Implementierung der virtuellen Arbeitsumgebung von der Durchführung der 3. Phase vorgenommen wird.

Aus den Erfahrungen der Projekte, die an der Entwicklung des D-Grids teilgenommen haben, wie auch solcher Infrastrukturprojekte wie dem internationalen „Virtual Observatory“ in der Astronomie, ist recht klar zu ersehen, dass eine enge Koppelung von der wissenschaftlichen Arbeit und der Entwicklung des IT basierten Supports die erfolgversprechendste Umgebung bildet. Natürlich ist diese Kombination dann erfolgreich, wenn zu Beginn eines solchen Vorhabens schon eine klare Vorstellung der zu implementierenden IT-Komponenten entwickelt worden ist. Für die 3. Phase der sozioökonomischen Berichterstattung sind diese Voraussetzungen vorhanden. Auch die im Umfeld des Projektes stattfindenden Entwicklungen können so mit aufgenommen werden, beispielsweise die zu erwartenden Bemühungen der Datenprovider, die Hindernisse einer effizienten wissenschaftlichen Nutzung von vorhandenen Daten weit möglichst zu reduzieren.

Die in den unter „Komponenten und Funktionsgruppen“ (Abschnitt 5) aufgelisteten Teilaufgaben können in 3 Kategorien gegliedert werden, die als Leitlinie bei der Implementierung dienen:

> **Kernelemente**: Die Implementation dieser Elemente ist unabdingbar für eine effiziente und nutzbare virtuelle Arbeitsumgebung. Eine unvollständige oder nur partielle Realisierung dieser Elemente würde insbesondere eine Akzeptanz durch die gesamte Community unwahrscheinlich machen und der Zweck der Aufbauarbeit wäre gefährdet. Zeitlich gesehen muss die Realisierung dieser Teilaufgaben im Rahmen von soeb III geschehen.

Hierzu zählen das Datenspeicherungs- und Datenmanagement-System inklusive eines Metadaten-Systems und Komponenten der virtuellen Arbeitsumgebung, die unabdingbar für die Nutzung dieses Systems von der Workstation des Wissenschaftlers aus sind. Die elementaren Anforderungen aus Datenschutz und Urheberschutz sind durch das vorgeschlagene Sicherheits-System unter Beibehaltung eines durch das Projekt geführten Systems von Berechtigungen beachtet. Und – nicht zuletzt – sind hier die Instrumente der Online-Publikation und der projektinternen Kommunikation zu nennen.

Langfristig notwendige Elemente: Die Möglichkeit der Implementation dieser Teilaufgaben ist nicht vom *soeb*-Projekt allein zu bestimmenden oder realisierbaren Entwicklungen abhängig. Hierzu gehören insbesondere alle Tools, die von den zu verhandelnden und gemeinsam zu schaffenden Schnittstellen zu verschiedenen FDZ abhängig sind. Im Verlauf der Projektdurchführung sind Entwicklungen zu erwarten, die in der virtuellen Arbeitsumgebung antizipiert sind, deren konkrete Ausgestaltung zwischen dem *soeb*-Projekt und anderen Akteuren der Community weiter spezifiziert und vereinbart werden müssen.

Als Beispiele seien hier die möglichen Änderungen bei der Bereitstellung von Daten durch die Forschungsdatenzentren, oder auch eine Lösung für das Problem der zeitlich limitierten Nutzung von Forschungsdaten genannt, welche derzeit nur in Ansätzen vorhanden ist, da hier das Bedürfnis der Forschung nach permanenter Verfügung von untersuchten Daten im Konflikt mit den durch Gesetze und den vom Vertrauensschutz gezogenen Grenzen der Nutzung steht (Archivierung von SUF). Ein weiterer Punkt ist die Nutzung von Diensten, die in den FDZ selbst stattfinden, entweder durch Remote-Zugriff (Fernrechnen) oder durch physische Anwesenheit des Forschers im FDZ (Onsite-Rechnen).

Elemente zur Erweiterung des Funktionsumfangs: Die Implementation dieser Aufgaben würde die Nutzerfreundlichkeit der virtuellen Arbeitsumgebung erweitern und zusätzliche Funktionen einbringen. Das kann insbesondere bei der über das Projekt hinausgehenden Verwendung helfen, die Anpassung an andere Vorhaben zu erleichtern.

Unter diese Kategorie fällt als ein wichtiges Element die Möglichkeit zur Erstellung, Darstellung und Nutzung verschiedener Kategorisierungen der Dateien des gemeinsamen Repositories. Diese ist nicht weniger wichtig als die Charakterisierung der Daten durch Metadaten, kann jedoch bei korrekter Implementierung in einer späteren Phase nachgezogen werden, und ist andererseits keine unverzichtbare Komponente für eine funktionsfähige Basisimplementation der virtuellen Arbeitsumgebung.

Auch die Integrierung von Skype oder ähnlichen Web 2.0 Komponenten ist nicht essentiell im Sinne der Funktionsfähigkeit. Bei den Kommunikationskomponenten ist vor allem auch die Abwägung zwischen dem Zusatz-Nutzen und dem Zusatz-Aufwand erforderlich. Beispielsweise ist ein Forum anstelle einer Mailing-Liste sicher besser imstande, Diskussions-Stränge nachvollziehbar zu ordnen, erfordert aber von allen Teilnehmern auch das Erlernen der neuen Technik. Gleiches gilt, wenn es um die gemeinsame Gestaltung und Nutzung der Projekt Website und des Intranet geht: möglicherweise ist ein System für Intra- und Extranet sinnvoller als zwei getrennte Systeme.

Eine weitere Aufgabe ist hier einzuordnen: die Restriktionen, welche durch die Lizenzbedingungen der genutzten Software entstehen, sind insbesondere in der virtuellen Arbeitsumgebung of-

fensichtlich kontraproduktiv. Es gilt hier, zusammen mit den Lizenz-Gebern, an die stattfindende Entwicklung angepasste Modelle zu finden. Diese Diskussion kann das Projekt mit gestalten, sollte es aber nicht zu einer der ureigenen Aufgaben machen. Hier sind gegebenenfalls, durch die verstärkte (und vereinfachte) Nutzung von Open Source-Komponenten wie R, alternative Wege zu beschreiten.

Langfristig ist auch eine Entwicklung der direkten Verwendung von HPC-Computing Methoden (Paralleles Ausführen von Berechnungen auf vielen CPUs) eine unbestrittene Notwendigkeit. Beim derzeitigen Stand der Community-Entwicklung ist jedoch noch kein großer Handlungsdruck vorhanden, und andererseits ist, wenn die Entwicklung der Syntax und der Methode zur Nutzung dieser Parallel-Berechnung so weit gediehen ist, durch die Remote-Komponenten der virtuellen Arbeitsumgebung der Rahmen schon vorhanden, um diese neue Arbeitsweise problemlos zu integrieren.

Aus dieser Kategorisierung ergeben sich sowohl sachliche als zeitliche Strukturierungen der Projektplanung. Es ist, auch durch die Erstellung der Expertise, für das geplante Projekt ein Bezugsrahmen vorhanden, der die organisatorischen und technischen Aspekte der Projektplanung vereinfacht und dadurch die inhaltliche Planung entlastet. Es werden Ressourcen zugänglich, die durch systematische Auswertung von anderen Projekt-Erfahrungen gewonnen sind (über WissGrid). Diese Ressourcen sollten auch während der Projektlaufzeit genutzt werden, z.B. bei der Anleitung und Überprüfung der Implementierung der virtuellen Arbeitsumgebung, oder bei den notwendigen Verhandlungen mit Ressourcenprovidern über geeignete Ressourcen für das Vorhaben.

Mit der Vorbereitung und der Durchführung des Projekts wird ein wichtiges zukünftiges Arbeitsinstrument in der sozialwissenschaftlichen Forschung geschaffen, das in seiner Anlage auch weit über das Projekt hinaus genutzt werden kann. Es wird die Grundlage verbessert, mit den Datenprovidern effizientere Wege zur wissenschaftlichen Nutzung der wertvollen Datenbestände zur gesellschaftlichen Entwicklung zu finden. Und durch Standardisierungen, die in der virtuellen Arbeitsumgebung gezielt berücksichtig werden, wird auch die Zusammenarbeit über die Ländergrenzen hinweg mit europäischen und internationalen Projekten und Entwicklungen substantiell verbessert.

8 Literaturverzeichnis

Aschenbrenner, A; Enke, H; Fritzsch, B, et al. (2010): WissGrid-Spezifikation: Grid-Repository, Göttingen, Niedersächsische Staats- und Universitätsbibliothek Göttingen, D3.5.2, 2010.04.30.

AstroGrid-D (2010), Letzter Zugriff: 2010.05.27, URL: http://www.gac-grid.org.

Baur, SK (2010): Aufwandsschätzung - Dauer und Kosten eines Software-Projekts werden geschätzt, Letzter Zugriff: 2010.07.21, URL: http://www.stefan-baur.de/cs.se.aufwand.html.

Blum, JM; Warner, GC; Jones, SB, et al. (2009): Metadata Creation, Transformation and Discovery for Social Science Data Management: The DAMES Project Infrastructure, 1st Annual European DDI Users Group Meeting, Bonn, Germany, 2009.12.04, Letzter Zugriff: 2010.05.25, URL: http://www.iza.org/eddi09.

Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), das zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert worden ist (2009), Letzter Zugriff: 2010.06.11, URL: http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf.

Burke, S; Campana, S; Lanciotti, E, et al. (2009): gLite 3.1 User Guide, Sciab`a, A. Manuals Series, Genf, Schweiz, CERN, https://edms.cern.ch/file/722398//gLite-3-UserGuide.pdf.

Chakrabarti, A (2007): Grid Computing Security, Springer, Berlin; Heidelberg; New York.

Chemomentum - Grid Services based Environment to enable Innovative Research (2010), Letzter Zugriff: 2010.05.27, URL: http://www.fz-juelich.de/jsc/grid/Chemomentum.

Data Documentation Initiative (DDI) (2009), Letzter Zugriff: 2010.07.01, URL: http://www.ddialliance.org.

dCache (2010), Letzter Zugriff: 2010.05.21, URL: http://www.dcache.org.

Deutsche Forschungsgemeinschaft (1998): Vorschläge zur Sicherung guter wissenschaftlicher Praxis: Empfehlungen der Kommission „Selbstkontrolle in der Wissenschaft“, WILEY-VCH, Weinheim. URL: http://www.dfg.de/aktuelles_presse/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf.

DGI-1 - FG 1: D-Grid-Basis-Software - Beschreibung (2008), URL: http://dgi2.d-grid.de/index.php?id=24.

Foster, I (2002): What is the Grid? A Three Point Checklist, Letzter Zugriff: 2009.12.09, URL: http://www-fp.mcs.anl.gov/~foster/Articles/WhatIsTheGrid.pdf.

GEODE: Grid Enabled Occupational Data Environment (2010), Letzter Zugriff: 2010.06.09, URL: http://www.geode.stir.ac.uk/index.html.

Gietz, P und Funk, SE (2010): Nutzung von SLCs und ROBOT-Zertifikaten in TextGrid, D-Grid All-Hands-Meeting 2010, Dresden, Germany, D-Grid, Letzter Zugriff: 2010.05.30, URL: http://www.d-grid.de/fileadmin/user_upload/documents/TextGrid.pdf.

Green, A (2008): Data Documentation Initiative - DDI, University of Edinburgh, UK, Letzter Zugriff: 2010.07.26, URL: http://www.disc-uk.org/docs/DDI_Green.pdf.

Grimm, C; Henne, B; Piger, S, et al. (2008): Generische VO-Strukturen für das D-Grid, D-Grid Integrationsprojekt 2 (DGI-2) Fachgebiet 3.2 AAI/VO, Letzter Zugriff: 2010.06.30, URL: http://dgi.d-

grid.de/fileadmin/user_upload/documents/DGI2-FG3/FG3-2/DGI-2_FG-3.2_Generische_VO-Strukturen_D-Grid.pdf.

Hartung, M; Loebe, F; Herre, H, et al. (2008): A Platform for Collaborative Management of Semantic Grid Metadata, Badica, C; GiuseppeMangioni; Carchiolo, V, et al., Proceedings of the 2nd International Symposium on Intelligent Distributed Computing – IDC 2008, Catania, Italy, 2008 (Band 162) Seite 115-125, Springer, Berlin ; Heidelberg.

Interoperabilität und Integration der VO-Management Technologien im D-Grid (2010), Letzter Zugriff: 2010.06.30, URL: http://www.d-grid.de/index.php?id=314&L=0.

IRODS (2010), Letzter Zugriff: 2010.05.27, URL: https://www.irods.org.

Jensen, S; Plale, B; Pallickara, SL, et al. (2006): A Hybrid XML-Relational Grid Metadata Catalog, Proceedings of the 2006 International Conference Workshops on Parallel Processing, Columbus, Ohio, USA, URL: http://doi.ieeecomputersociety.org/10.1109/ICPPW.2006.10.

JHOVE - JSTOR/Harvard Object Validation Environment (2010), Letzter Zugriff: 2010.06.09, URL: http://hul.harvard.edu/jhove/.

Martinez, L (2008): The Data Documentation Initiative (DDI) and Institutional Repositories, DISC-UK, Letzter Zugriff: 2010.07.26, URL: http://www.disc-uk.org/docs/DDI_and_IRs.pdf.

Milke, J-M; Schiffers, M und Ziegler, W (2008): Rahmenkonzept für das Management Virtueller Organisationen im D-Grid, Version 1.1, Letzter Zugriff: 2010.06.30, URL: http://www.d-grid.de/fileadmin/user_upload/documents/DGI-FG1-10/VO_Rahmenkonzept-final.pdf.

Muenchen, RA (2009): R for SAS and SPSS users, Statistics and computing, Springer, New York, NY.

Muenchen, RA (2010): R for stata users, 1st. Auflage, Statistics and computing, Springer, New York, NY.

Open Grid Services Architecture Data Access and Integration (OGSA-DAI) (2010), Letzter Zugriff: 2010.05.25, URL: http://www.ogsadai.org.uk.

POI-HSSF and POI-XSSF - Java API To Access Microsoft Excel Format Files (2010), Letzter Zugriff: 2010.06.09, URL: http://poi.apache.org/spreadsheet/index.html.

Pukacki, J; Kosiedowski, M; Mikołajczak, R, et al. (2006): Programming Grid Applications with Gridge, Computational Methods in Science and Technology (Band 12), Nr. 1, Seite 47–68. URL: http://www.gridge.org/files/gridge/publications/programming-grid-applications-with-gridge.pdf.

Rasch, K; Schöne, R; Ostropytskyy, V, et al. (2009): The Chemomentum Data Services – A Flexible Solution for Data Handling in UNICORE, Euro-Par 2008 Workshops - Parallel Processing, VHPC 2008, UNICORE 2008, HPPC 2008, SGS 2008, PROPER 2008, ROIA 2008, and DPA 2008, Las Palmas de Gran Canaria, Spain, Letzter Zugriff: 2010.05.27, URL: http://www.springerlink.com/content/2u70m8443180620m/fulltext.pdf.

Razum, M; Schwichtenberg, F und Fridman, R (2007): Versioning of Digital Objects in a Fedora-based Repository, German eScience Conference 2007, Baden-Baden, Germany, URL: http://edoc.mpg.de/315520.

Resource Description Framework (RDF) (2010), Letzter Zugriff: 2010.05.27, URL: http://www.w3.org/RDF/.

SRB - The DICE Storage Resource Broker (2010), Letzter Zugriff: 2010.05.27, URL: http://www.sdsc.edu/srb/index.php/Main_Page.

Schwaber, C; Gilpin, M und Stone, J (2007): Software Change And Configuration Management, The Forrester Wave™, Q2 2007, Q2 2007, Forrester Research, I, 2007.07.29, Q2 2007, http://www.collab.net/forrester_wave_report/index.html.

SPSS Writer (2010), Letzter Zugriff: 2010.06.09, URL: http://spss.pmstation.com/.

StataCorp LP (2010): End-User License Agreement (EULA), Letzter Zugriff: 2010.06.28, URL: http://www.stata.com/order/licterms.html.

Stellaris (2010), Letzter Zugriff: 2010.05.21, URL: http://stellaris.zib.de.

Thain, D und Livny, M (2005): Parrot: An Application Environment for Data-Intensive Computing, Scalable Computing: Practice and Experience (Band 6), Nr. 3, Seite 9-18. URL: http://www.cse.nd.edu/~dthain/papers/parrot-scpe.pdf.

The Fedora Development Team (2008): Fedora Tutorial #1 - Introduction to Fedora, Fedora Commons, 2008.07.23, Letzter Zugriff: 2010.05.26, URL: http://fedora-commons.org/confluence/download/attachments/4718930/tutorial1.pdf?version=1&modificationDate=1218459761506

The Globus Alliance (2010): Globus Toolkit, Letzter Zugriff: 2010.05.27, URL: http://www.globus.org/toolkit/.

UNICORE - Distributed computing and data resources (2010), Letzter Zugriff: 2010.05.12, URL: http://www.unicore.eu.

Wegener, D; Sengstag, T; Sfakianakis, S, et al. (2007): GridR: An R-Based Grid-Enabled Tool for Data Analysis in ACGT Clinico-Genomics Trials, Proceedings of the Third IEEE International Conference on e-Science and Grid Computing, Bangalore, India.

Wikimedia Commons (2010): Principle of a public key infrastructure, Letzter Zugriff: 2010.06.06, URL: http://upload.wikimedia.org/wikipedia/commons/3/34/Public-Key-Infrastructure.svg.

WissGrid - Grid für die Wissenschaft (2010), Letzter Zugriff: 2010.05.27, URL: http://www.wissgrid.de.

XMPP Standards Foundation (2010), Letzter Zugriff: 2010.06.09, URL: http://xmpp.org.

Anhang A. Erste Design-Details

Anhang A.1. Grundlegende Funktionalitäten von Dateibrowsern

Die verschiedenen Dateibrowser der unterschiedlichen Betriebs- und Dateisysteme verfügen über die folgenden Funktionalitäten, die innerhalb der Datenverwaltungswerkzeuge (siehe Kapitel 5.2.3) der virtuellen Arbeitsumgebung ebenfalls zur Verfügung stehen sollten:

- Darstellung von Verzeichnissen und Dateien, ggf. in Form eines aufklappbaren Verzeichnisbaums
- Erstellen von neuen Verzeichnissen oder Dateien mit Hilfe von Kontextmenüs
- Hinzufügen von Verzeichnissen und Dateien mittels Drag & Drop
- Löschen von Dateien und Verzeichnissen mit Kontextmenüs oder entsprechenden Shortcuts inklusive einer Papierkorbfunktion
- Verschieben von Dateien und Verzeichnissen mit Drag & Drop
- Ausschneiden und (Wieder-)Einfügen von Dateien und Verzeichnissen mit Kontextmenüs oder bekannten Shortcuts
- Umbenennen von Dateien oder Verzeichnissen
- Öffnen von Dateien oder Verzeichnissen, direkt durch anklicken oder indirekt über das Kontextmenü, ggf. mit Auswahl des entsprechenden Programms
- Anzeige von grundlegenden Metadaten von Dateien und Verzeichnissen mit Hilfe von Detailansichten, Kontextmenüpunkten, über eine Statusleiste sowie mittels Hover-Effekten
- Vorschau auf Dateien mit dafür adäquatem Inhalt (z.B. Bilddateien) mittels Kontextmenü, besonderem Vorschaubereich oder Hover-Effekten
- Ggf. einfache Anzeige des Inhalts von Archivdateien mit und ohne Komprimierung
- Navigationsmöglichkeiten ähnlich derer eines Browsers (Vor, Zurück, Adressleiste, etc.)

Anhang A.2. Funktionale Ergänzungen zur Kategorisierung von Dateien

Wie in Kapitel 5.2.3 dargestellt müssen die im Anhang A.1 aufgelisteten Funktionalitäten von Dateisystembrowsern in den Datenverwaltungswerkzeugen für eine effektivere Kategorisierung von Dateien ergänzt werden. In der folgenden Liste sind einige Beispiele hierfür enthalten:

- Darstellung von Kategorienhierarchien mit allen im Anhang A.1 genannten Funktionen von Dateibrowsern, wobei ein Verzeichnis einer Kategorie entspricht
- Bearbeitung und freie Ergänzung von Metadaten von Kategorien mit Hilfe entsprechender Eingabemasken, die direkt sichtbar sind oder über Kontextmenüs und Shortcuts aufgerufen werden können
- Darstellung von, und freier Wechsel zwischen unterschiedlichen Kategorienhierarchien, z.B. mit Hilfe einer Tabbed View

- Funktionen zur Erstellung und Definition neuer Kategorienhierarchien, welche über ein Kontextmenü verfügbar sind
- Nebeneinanderdarstellung von Kategorienhierarchien zum effizienten Einordnen von Dateien in mehrere Kategorien gleichzeitig, aufrufbar über ein Kontextmenü
- Darstellung einer Liste von mehreren Wurzel-Kategorien, welche den Ausgangspunkt für Dateisammlungen mit unterschiedlichen Kategorienhierarchien bilden. Per Klick auf ein Element der Liste wird die Darstellung für die unterschiedlichen darunterliegenden Kategorienhierarchien angezeigt.

Anhang A.3. Funktionen zur Arbeit mit Dateiversionen

Um in den Datenverwaltungswerkzeugen effektiv mit verschiedenen Versionen von Dateien arbeiten zu können, werden unterschiedlichste Funktionalitäten benötigt. Diese sind in Kapitel 5.2.3 benannt. An dieser Stelle werden ein paar entsprechende Implementierungsvorschläge aufgelistet.

- Anzeige der Historie von Dateien mit dazugehörigen Versionsnummern, Versionszeitpunkten und gegebenenfalls Versionskommentaren (Menüpunkt im Kontextmenü beim Rechtsklick auf eine Datei)
- Möglichkeit zum Anzeigen einer bestimmten Version z.B. durch Mausklick oder Kontextmenü, gegebenenfalls mit Auswahl des zu verwendenden Programms
- Möglichkeit zum textuellen Vergleich zweier Dateiversionen durch Markierung und einen entsprechenden Eintrag im Kontextmenü (ein dafür notwendiger Editor wird bei den Datenvergleichswerkzeugen in Kapitel 5.2.5 beschrieben)
- Möglichkeit zum Löschen nicht mehr benötigter Versionen, z.B. durch Kontextmenueintrag oder Shortcuts in der Anzeige der Historie einer Datei

Anhang A.4. Verknüpfungen und Nutzungszeiträume

Kapitel 5.2.3 nennt grob die zu implementierende Unterstützung des Anwenders hinsichtlich Dateiverknüpfungen, z.B. zur Darstellung von Provenienz, sowie zur Spezifikation und Auswertung von Nutzungszeiträumen. Hier sind nun ein paar Ideen zur möglichen Implementierung dieser Funktionalitäten in den Datenverwaltungswerkzeugen aufgelistet:

- Über das Kontextmenü aufrufbarer Editor zur Definition von logischen Verknüpfungen zwischen Dateien. Diese Verknüpfungen müssen beschreibbar, und deren Art mit fest vorgegebenen Typen klassifizierbar sein (z.B. „wurde erstellt mit Hilfe“, „wurde erstellt auf Basis von“, etc.)
- Dialog zum Löschen von Dateien, sofern diese mit anderen verknüpft sind oder von anderen referenziert werden. Der Dialog muss dabei die Verknüpfungen analysieren und je nach Verknüpfungsart den Nutzer auf die Konsequenzen des Löschens hinweisen. Der Dialog wird automatisch beim Löschen von Dateien oder Dateiversionen gestartet. Er erlaubt außerdem, dass nur die Dateien oder Dateiversionen, jedoch nicht deren Metadaten gelöscht werden (siehe virtuelle Datensätze)
- Über das Kontextmenü aufrufbarer Editor zum Festlegen von Nutzungszeiträumen für Dateien (kann gegebenenfalls in den Editor für Metadaten integriert werden, welcher im Kapitel 5.2.4 beschrieben wird)

- Über das Kontextmenü aufrufbare sowie im Hintergrund automatisch laufende Überprüfung, ob Nutzungszeiträume für Dateien demnächst ablaufen oder bereits abgelaufen sind, inklusive entsprechender Hinweise und Vorschläge für Benutzer (z.B. Möglichkeit zum Aufrufen des Löschen Dialogs). Diese Überprüfung muss auch immer dann für eine Datei stattfinden, wenn für diese eingeschränkte Nutzungszeiträume definiert sind und diese geöffnet oder anderweitig gelesen (z.B. kopieren, verschieben, etc.) wird.
- Spezielle Markierung virtueller Datensätze in der Anzeige von Kategorienhierarchien

Anhang A.5. Definition von Zugriffsrechten

In Kapitel 5.2.3 wird erwähnt, dass die Datenverwaltungswerkzeuge Möglichkeiten zur Spezifikation von Zugriffsrechten bieten müssen. Dies könnte wie folgt implementiert werden:

- Über das Kontextmenü aufrufbarer Editor zur Anzeige und Vergabe von eigenen und gruppenweiten Zugriffsrechten auf Dateien oder Kategorien
- Beachtung von Zugriffsrechten bei der Darstellung von Kategorienhierarchien durch das Nicht-Anzeigen von Dateien und Kategorien, die für einen Nutzer nicht sichtbar sein dürfen
- Beachtung von Zugriffsrechten durch das Verhindern von Modifikationen oder Löschungen in Kategorienhierarchien

Anhang A.6. Umsetzung der Dateisuche

Die Suche nach Dateien, so wie sie in Kapitel 5.2.3 als Teil der Datenverwaltungswerkzeuge erwähnt wurde, könnte wie folgt umgesetzt werden:

- Über ein Menü aufrufbare Suchmaske für eine globale Suche, welche die Angabe von Bedingungen in Form von Keywords oder Key-Value-Paaren, ggf. unter Beachtung einfacher regulärer Ausdrücke, erlaubt. Die Ergebnisse sollen als Liste unter Angabe der verschiedenen Kategorienhierarchien, in der die Dateien, Kategorien oder virtuellen Datensätze eingeordnet wurden, angezeigt werden, sofern der Benutzer Zugriff auf diese und mindestens eine darüberliegende Kategorienhierarchie hat. Hierbei soll die Spezifikation von Suchbedingungen auf Dateiverknüpfungen möglich sein und existierende Verknüpfungen in der Ergebnisliste angezeigt werden.
- Über das Kontextmenü einer Kategorie aufrufbare Suchmaske für eine eingeschränkte Suche, welche die Angabe von Bedingungen in Form von Keywords oder Key-Value-Paaren, ggf. unter Beachtung einfacher regulärer Ausdrücke, erlaubt. Die Ergebnisse sollen als Teilhierarchie der ausgewählten Kategorienhierarchie dargestellt werden. Auch hierbei soll die Spezifikation von Suchbedingungen auf Dateiverknüpfungen möglich sein und existierende Verknüpfungen in der Ergebnisliste angezeigt werden.

Anhang A.7. Aufruf der Datenbearbeitungswerkzeuge

Die Datenbearbeitungswerkzeuge sollen so einfach wie möglich für eine existierende Datei aufgerufen werden können. Dabei sollte sich an existierenden Systemen orientiert werden. Mögliche Umsetzungen wären:

- Öffnen einer Datei mit einem Standard-Editor durch Doppelklick auf die Datei
- Freie Auswahl eines speziellen Editors durch einen Menüpunkt im Kontextmenü

- Anpassung des Standard-Editors entweder anhand von Häufigkeiten seiner Auswahl und/oder auf Basis von Konfigurationseinstellungen des Nutzers (z.B. über Kontextmenü)
- Möglichkeiten zur Formatkonvertierung und -validierung über Menüpunkte im Kontextmenü

Anhang A.8. Funktionalitäten eines Metadateneditors

Die Aufgabe eines Metadateneditors besteht darin, dem Forscher effektive Möglichkeiten zum Anzeigen und Bearbeiten der mit einer Datei verknüpften Metadaten zu bieten. Wichtig ist hierbei, dass die Einfachheit und Intuitivität zu jeder Zeit gegeben sein muss. Sollten z.B. Metadatenstrukturen in XML vorliegen, ist es wenig sinnvoll einen einfachen XML-Editor zur Verfügung zu stellen. Dem Forscher sind nämlich die zu erstellenden XML Strukturen nicht unbedingt bekannt. Genauso kann nicht erwartet werden, dass der Forscher die Konzepte von XML kennt. Daher sollten bei der Entwicklung eines Metadateneditors folgende Aspekte Berücksichtigung finden:

- Der Metadateneditor dient sowohl dem Anzeigen als auch dem Bearbeiten von Metadaten.
- Es sollten mit Label versehene Eingabefelder für die einzelnen Metadatenelemente existieren. Die gewählten Label sollten selbsterklärend sein und gegebenenfalls durch Hilfefunktionen oder Hover-Effekte ergänzt werden, welche eine genauere Beschreibung der Semantik und der Syntax eines Metadatenelements und dessen Wertes liefern. Die Eingabefelder selbst sollten entweder nur valide Eingaben erlauben oder von Listen und Comboboxen Gebrauch machen.
- Die Metadatenelemente sollten in logische Gruppen zusammengefasst und dargestellt werden. Die Namen der Gruppen und entsprechende Erläuterungen sollten ähnlich wie bei den Metadatenelementen vorhanden sein. Die Gruppen werden dabei nicht notwendiger Weise anhand der zugrundeliegenden Datenstruktur, sondern sinnvoll zusammengestellt.
- Aus Gründen der Übersicht werden gegebenenfalls nicht immer alle Eingabefelder angezeigt. Stattdessen sollten dem Benutzer nur die relevantesten Eingabefelder direkt zugänglich sein. Alle anderen sind entweder über Scrollen der Anzeige oder dedizierte Dialoge zugreifbar. Auch bei den Dialogen gelten die Prinzipien der Selbsterklärung.
- Der Editor sollte während der Bearbeitung genau zeigen, welche Werte zwar verändert aber noch nicht gespeichert wurden. Beim Verlassen des Editors sollte der Benutzer auf nicht gespeicherte Änderungen hingewiesen und um entsprechende Entscheidung zum weiteren Vorgehen gebeten werden.
- Sollte es durch parallele Bearbeitung von Metadaten einer Datei zu Speicherungskonflikten kommen, so sollten dieser vom Editor erkannt und intelligent (ggf. mit Nutzerinteraktion) gelöst werden.
- Zur Kontrolle aber auch zum Austausch sollten Metadatenstrukturen in gängige Formate, z.B. XML, exportiert werden können.

Anhang A.9. Dialoge zum Ausführen von Berechnungen

Um statistische Berechnungen aus einer virtuellen Arbeitsumgebung heraus anzustarten, wird ein einfacher und leicht zu bedienender Dialog benötigt. Dieser könnte z.B. über das Kontextmenü einer Syntaxdarstellung (Datenverwaltungswerkzeuge oder Syntaxeditor) aufrufbar sein. Der Dialog sollte folgende Eigenschaften aufweisen:

- Darstellung, welche Syntax mit welchen Ein- und Ausgangsdaten verarbeitet wird

- Möglichkeiten zur Einstellung relevanter Parameter
- Möglichkeiten zur manuellen Auswahl von Ein- und Ausgabedaten z.B. anhand eines Dateiauswahldialogs
- Möglichkeiten zur Auswahl zwischen lokaler oder entfernter Ausführung der Verarbeitung inklusive einer ausführlichen Beschreibung
- Möglichkeit zur Eingabe einer E-Mailadresse zur automatischen Benachrichtigung
- Automatisches Speichern vorgenommener Einstellungen zur wiederholten Ausführung der Berechnungen
- Buttons zum Starten der Berechnung, zum Zurücksetzen der vorgenommenen Einstellungen und zum Abbruch des Dialogs

Nachdem eine Berechnung angestartet wurde, sollte das System eine Anzeige zu aktuellen Verarbeitungsstatus öffnen. Diese Anzeige könnte folgendes Erscheinungsbild haben:

- Darstellung einer Liste mit Einzelaufträgen
- Zu jedem Auftrag werden grundlegende Informationen, wie die ausgeführte Syntax, die Startzeit, die vermutete Endzeit o.ä. angezeigt
- Über ein Kontextmenü oder mit Doppelklick sollten die Detailinformationen für einen Auftrag angezeigt werden können. Diese Anzeige sollte ähnlich dem Dialog zur Erstellung von Aufträgen sein. Hieraus sollte auch ein Abbrechen der Verarbeitung initiiert werden können.

Anhang A.10. Kollaborativer Syntaxeditor

Ein kollaborativer Syntaxeditor soll mehreren räumlich von einander getrennten Personen das gleichzeitige Bearbeiten von Syntaxdateien erlauben. Dazu sollten die folgenden Funktionen zur Verfügung stehen:

- Aufruf des Editors ähnlich eines normalen Syntaxeditors
- Darstellung der Syntax ähnlich der eines normalen Syntaxeditors, wenn möglich mit Highlighting von Keywords
- Bearbeiten der Syntax ähnlich eines normalen Syntaxeditors, wenn möglich mit Autovervollständigen und sonstigen Funktionen
- Möglichkeit zum Einladen weiterer Projektteilnehmer zur Bearbeitung entweder durch eine einfache Listenauswahl oder per entsprechendem Dialog
- Sperren der Bearbeitung durch andere Teilnehmer, sofern ein Teilnehmer gerade bearbeitet. Die Sperrung muss für die Teilnehmer eindeutig ersichtlich sein und auch angeben, welcher der Teilnehmer gerade bearbeitet
- Synchronisation des Scrollens in einer Syntaxdatei so dass alle Teilnehmer immer das gleiche sehen.