Projektantrag im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS)

LIS-Förderprogramm oder Ausschreibung: Werkzeuge und Verfahren

Prof. Dr. Klaus Tochtermann, Kiel	Prof. Dr. Wilhelm Hasselbring, Kiel
Prof. Dr. Hans-Joachim Bungartz, München	Prof. Dr. Arndt Bode, München
Prof. Dr. Wolfgang E. Nagel, Dresden	Dr. Christian Grimm, Berlin
Jochem Pattloch, Berlin	

Beschreibung des Vorhabens

GeRDI: Generic Research Data Infrastructure

1 Ausgangslage und eigene Vorarbeiten

Ausgangslage

Der vorliegende Projektantrag befasst sich mit der Entwicklung einer *Generic Research Data Infrastructure* (GeRDI), in deren ersten Phase drei Pilot-Datenzentren für das Management von Forschungsdaten aufgebaut werden. Diese Pilotsysteme sind so miteinander vernetzt, dass Disziplingrenzen überwunden und Forschungsdaten sowohl innerhalb einer Disziplin als auch aus verschiedenen Disziplinen miteinander kombiniert und so multidisziplinär genutzt werden können. Parallel zu einer zweiten Projektphase (die nicht Gegenstand dieses Antrags ist) soll die entwickelte Lösung dann gegebenenfalls in Deutschland breit ausgerollt werden und kann somit, falls entsprechende Fördermechanismen eingerichtet werden, Modellcharakter für eine zukünftige *German Research Data Infrastructure* haben.
Die *Generic Research Data Infrastructure* hat den Anspruch, den sogenannten *Long Tail* zu bedienen – also primär universitäre Nutzer bzw. Nutzer mit moderaten Volumina und ohne großen bereits existierenden Organisationsgrad der zugehörigen Fachcommunities. Es geht also nicht um Großnutzer in bereits organisierten Communities, also Big Data, so wie es die *Helmholtz Data Initiative* anstrebt. Da es insbesondere in den Diskussionen zur *European Open Science Cloud* noch keine klare Position zur Berücksichtigung des Themas *Citizen Science* sowie den dabei entstehenden Daten gibt, wird *Citizen Science* im Rahmen von *GeRDI* ebenfalls nicht näher behandelt.
Insgesamt kann mit *GeRDI* ein wesentlicher und modellhafter Beitrag dazu geleistet werden, dass Hochschulen und außeruniversitäre Forschungseinrichtungen die Aufgabe der Bereitstellung von Forschungsdaten aktiv annehmen können. In weiterer Folge kann dies beispielsweise für Bibliotheken der Hochschulen und Forschungseinrichtungen neue Aufgabenfelder eröffnen, da sie neben der Versorgung mit wissenschaftlicher Literatur auch eine tragende Rolle in der Versorgung mit Forschungsdaten einnehmen können.
Mit den angestrebten Ergebnissen leistet das Projekt eine wichtige Ergänzung der deutschen eInfrastructure-Landschaft und stellt die Anschlussfähigkeit des deutsches Wissenschaftsystems an aktuelle europäische und internationale Entwicklungen (*European Open Science Cloud*) sicher.
Der nachfolgende Abschnitt skizziert die Ausgangslage bezüglich der Themenfelder eInfrastructures für Forschungsdaten in Deutschland und Management von Forschungsdaten.

eInfrastructures für Forschungsdaten in Deutschland
Betrachtet man die derzeitige eInfrastructure-Landschaft in Deutschland, so erkennt man, dass über das Deutsche Forschungsnetz (DFN) leistungsfähige Kommunikationsnetze flächendeckend für das deutsche Wissenschaftssystem vorhanden sind. Auch im Bereich des wissenschaftlichen Hoch- bzw. Höchstleistungsrechnens gibt es mit der *Gauß-Allianz* und dem *Gauss Centre for Supercomputing* (GCS) ein gut organisiertes, mehrstufiges System an Hoch- bzw.

Höchstleistungsrechnern sowie einschlägige Nutzungskompetenz. Ein vergleichbares flächendeckendes, miteinander verbundenes System für das Management von und den Zugang zu Forschungsdaten über Forschungsdaten-Repositorien, insbesondere an den Hochschulen, fehlt derzeit jedoch. Vielmehr ist die deutsche Landschaft hier immer noch sehr fragmentiert und in hohem Maße disziplinär geprägt (vgl. nächster Abschnitt).

Auch in Europa fehlt ein solches disziplinübergreifendes und vernetztes System. Die Europäische Kommission hat dieses Defizit jedoch für die europäische eInfrastructure-Landschaft erkannt und eine *High Level Expert Group European Open Science Cloud* ins Leben gerufen. In dem zum Zeitpunkt dieser Antragstellung noch unveröffentlichten Bericht[1] der HLEG werden insbesondere die nachfolgenden Herausforderungen für ein europaweit koordiniertes Forschungsdatenmanagement genannt. Der vorliegende Antrag greift genau diese Anforderungen auf:

- *The major challenge is not the size of data per se, but in particular complex data and analytics across domains.*
- *The fragmentation (even now that the ESFRI scheme is highly successful) between domains causes repetitive and isolated solutions.*
- *The ever larger distributed data sets increasingly do not move (for sheer size or for privacy reasons) and centralised HPC is therefore insufficient to support the critically federated and distributed meta-analysis.*

Hintergrund ist die Idee der Europäischen Kommission, eine *European Open Science Cloud* umzusetzen, die drei Ebenen umfasst: Auf einem Data Layer wird die technische Infrastruktur (Storage etc.) angeboten, ein Service Layer bietet generische Dienste für das Management von Forschungsdaten, und über einen Governance Layer werden disziplinspezifische Rahmenbedingungen (z.B. Datenschutz) und Funktionalitäten angeboten. Die nachfolgende Abbildung aus einem Vortrag[2] von Jean-Claude Burgelman, Head of Unit A6 “Science Policy, Foresight and Data” innerhalb der *Generaldirektion Research and Innovation*, illustriert diese Idee:

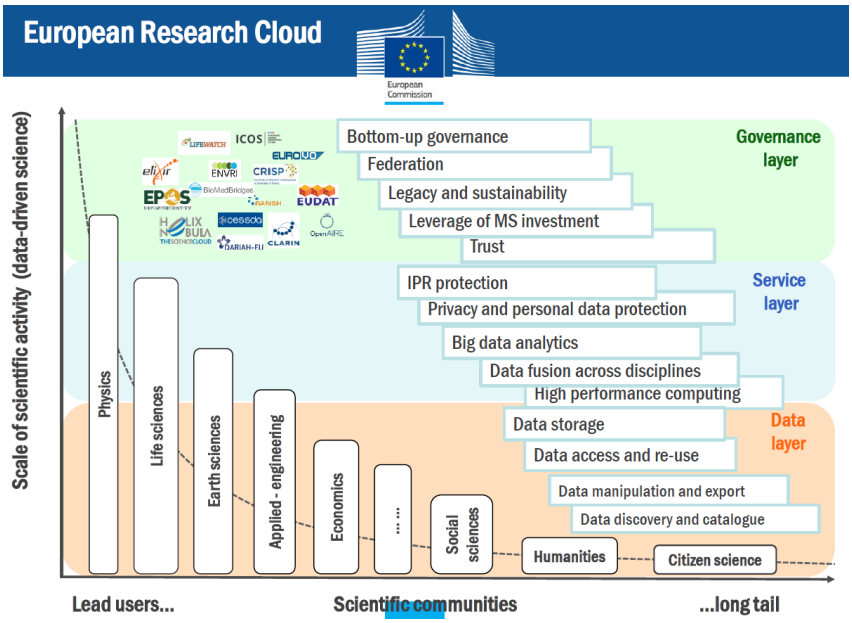

Abb. 1: Aktuelle Idee einer *European Open Science Cloud*

Mit dieser Idee möchte die Europäische Kommission dem insbesondere mit Horizon 2020 angestoßenen Trend zur multidisziplinären Forschung mehr Nachdruck verleihen. Übertragen auf das Forschungsdatenmanagement bedeutet multidisziplinäre Forschung auch, dass zukünftig erhöhte Anforderungen an die Kombination von Forschungsdaten aus unterschiedlichen

[1] Der Bericht der High Level Expert Gruppe, liegt K. Tochtermann als Vertreter für Deutschland in dieser Gruppe vor.
[2] Folien siehe International Science 2.0 Conference 2015, http://www.science20-conference.eu

Disziplinen gestellt werden. Beispielsweise benötigt die Verhaltensökonomik, die sich mit menschlichem Verhalten in wirtschaftlichen Situationen befasst, sowohl Daten aus der Psychologie als auch aus den Wirtschaftswissenschaften.

Management von Forschungsdaten

Auf nationaler, europäischer und internationaler Ebene werden seit geraumer Zeit erhebliche Anstrengungen unternommen, um das Thema Management von Forschungsdaten voranzutreiben. Die Aktivitäten umfassen disziplinabhängige sowie disziplinübergreifende Projekte zur Entwicklung von Forschungsdaten-Infrastrukturen bzw. zur Verknüpfung von Publikationen mit Forschungsdaten. Darüber hinaus gibt es auch Initiativen zum Nachweis von Forschungsdatenrepositorien bzw. zur Entwicklung von Handlungsempfehlungen und Standards. Die wahrscheinlich wichtigsten dieser Aktivitäten werden nachfolgend kurz skizziert.

Zu den wichtigen **nationalen Projekten und Forschungsdatenzentren** gehört sicherlich das Projekt *Rahmenbedingungen einer disziplinübergreifenden Forschungsdaten-Infrastruktur* (*Radieschen*), an dem auch die TU Dresden beteiligt war. Es untersuchte die Anforderungen an generische Komponenten einer Infrastruktur und deren Vernetzung mit disziplinspezifischen Bestandteilen. Der Abschlussbericht von *Radieschen* kommt u.a. zu dem Ergebnis, dass es in Zukunft immer wichtiger wird, professionelle Dienste zum Management von Forschungsdaten zur Verfügung zu haben und zu nutzen, wobei einige Datenmanagement- und Datenverarbeitungsschritte spezifisch für eine Fachdisziplin seien, für eine Vielzahl an Diensten jedoch übergreifende Lösungen entwickelt werden, die allerdings eine langfristige Perspektive und Nachhaltigkeit erst noch nachweisen müssen (Radieschen 2013:18)[3]. Die wesentlichen Fortschritte für das nachhaltige Forschungsdatenmanagement werden in den nächsten Jahren laut *Radieschen* immer noch innerhalb einer Fachdisziplin erzielt werden (Radieschen 2013:20).
Als disziplinabhängige sowie disziplinübergreifende Projekte sind die DFG-geförderten Projekte *RADAR* und *MASi* sowie das von der Leibniz-Gemeinschaft geförderte Projekt *SowiDataNe*t interessant. *RADAR*[4] – *Research Data Repositorium* – hat sich zum Ziel gesetzt, eine generische Forschungsdateninfrastruktur für die Archivierung und Publikation von Forschungsdaten in „Small Sciences“ zu etablieren und dort das oft noch fehlende Forschungsdatenmanagement zu unterstützen. *MASi*[5] – *Metadatenmanagement für Angewandte Wissenschaften*, mit der TU Dresden als Koordinator – bietet disziplinübergreifenden Anwendungsfällen mit unterschiedlichsten Datenmengen ein fortschrittliches Forschungsdatenrepositorium, welches integrativ den gesamten Datenlebenszyklus – von der Datenerzeugung über die Analyse bis zur Archivierung – abdeckt. Einen im Vergleich zu *RADAR* ähnlichen, allerdings disziplinspezifischen Ansatz verfolgt das von der Leibniz-Gemeinschaft geförderte Projekt *SowiDataNet*[6], an dem auch die ZBW beteiligt ist. Ziel des Projekts ist der Aufbau eines Forschungsdatenverbundes für quantitative Daten aus den Sozial- und Wirtschaftswissenschaften. Kern des Verbunds wird eine webbasierte, eigenständige Infrastruktur sein, die eine niederschwellige Selbstarchivierung, Dokumentation und Distribution von Forschungsdaten ermöglicht und sich dabei am konkreten Bedarf der Scientific Community orientiert. Dahinter steht die Überzeugung, dass der Umgang mit Forschungsdaten hochgradig disziplinspezifisch ist und Infrastrukturlösungen deshalb in enger Kooperation mit den Fachwissenschaftlern entwickelt werden müssen, um erfolgreich zu sein.
Neben den Forschungsdatenrepositorien beschäftigen sich verschiedene Projekte mit publikationsbezogenen Forschungsdaten, d.h. mit der Verknüpfung wissenschaftlicher Artikel mit den ihnen zugrunde liegenden Forschungsdaten. Im Projekt *InFoLIS*[7] wurden z.B. die im Datenbestandskatalog der GESIS verwalteten Forschungsdaten des Datenarchivs mit den Aufsätzen und Büchern verknüpft, die die Universitätsbibliothek Mannheim in ihrem Recherchesystem präsentiert. Ebenfalls in den Kontext publikationsbezogener Forschungsdaten fällt das Projekt *European Data Watch Extended* (*EDaWaX*). Inhaltlich beschäftigt sich *EDaWaX* mit dem Management von

[3] Projekt *Radieschen* Report „Synthese“. Entspricht dem Report D6.3 „Abschlussbericht des Projekts und Roadmap für eine Infrastruktur für Forschungsdaten in Deutschland“ nach Projektantrag. 30.04.2013, S.31
[4] http://www.radar-projekt.org/display/RD/Home
[5] http://tu-dresden.de/die_tu_dresden/zentrale_einrichtungen/zih/forschung/projekte/masi
[6] http://www.zbw.eu/de/ueber-uns/arbeitsschwerpunkte/forschungsdatenmanagement/sowidatanet/
[7] http://www.gesis.org/forschung/drittmittelprojekte/archiv/infolis/

Forschungsdaten in wirtschaftswissenschaftlichen Fachzeitschriften. Hintergrund ist, dass zahlreiche Beiträge in Fachzeitschriften auf der Auswertung von Daten und Statistiken beruhen. Die in der Publikation postulierten Ergebnisse solcher Auswertungen können jedoch nicht ohne die dafür verwendeten Forschungsdaten sowie den Berechnungsweg (Syntax) geprüft oder repliziert werden. Im DFG-geförderten Projekt *PubFlow* wird ein workflowbasiertes Publikationsframework für Forschungsdaten konzipiert, prototypisch realisiert und durch das Datenmanagementteam am GEOMAR (Helmholtz-Zentrum für Ozeanforschung Kiel) genutzt und evaluiert. Konkret wird der Datenfluss von Ozeanbeobachtungsinstrumenten in das Forschungsdatensystem *Pangaea*[8] durch Workflow-Technologie unterstützt.
Zudem gibt es in Deutschland mittlerweile zahlreiche institutionelle Forschungsdatenzentren, wie z.B. die ca. 30 vom *Rat für Sozial- und Wirtschaftsdaten* (*RatSWD*) akkreditierten Datenzentren[9]. Die Max-Planck-Gesellschaft betreibt im Rahmen ihrer digitalen Bibliothek gemeinsam mit dem FIZ Karlsruhe die *eSciDoc*[10]-Infrastruktur, die als Repositorium für unterschiedliche Forschungsdaten in der MPG eingesetzt wird. Auch Helmholtz-Zentren betreiben signifikante Forschungsdatenzentren, wie das Datenrepositorium *Pangaea*, das *World Data Center for Remote Sensing of the Atmosphere*, das *World Stress Map Project* oder das Projekt *Large-Scale Data Management and Analysis* (*LSDMA*). Schließlich seien die zwei vom BMBF geförderte Big Data-Kompetenzzentren *Competence Center for Scalable Data Services* (*ScaDS*) und *Berlin Big Data Center* (*BBDC*) genannt. Während das *BBDC* automatisch skalierbare Technologien entwickeln wird, die riesige heterogene Datenmengen organisieren und daraus intelligent Informationen gewinnen, wird das *ScaDS* Dresden/Leipzig durch den eher serviceorientierten Ansatz bei gleichzeitiger Forschung für Big Data ein Portfolio von Big Data-Lösungen für die Wissenschaft und die Industrie erforschen, entwickeln und zugänglich machen. Ein besonderer Schwerpunkt dieses Kompetenzzentrums, der auch diesem Projekt zugutekommen wird, ist die Forschung im Bereich effizienter Big Data-Architekturen.
Schließlich befindet sich derzeit die *Helmholtz Data Initiative* in Vorbereitung. Diese soll in 2017 starten und wird dabei Großnutzer (Big Data) aus bereits gut organisierten Communities als Zielgruppe haben. Im Zuge der Antragstellung fanden bereits Abstimmungsgespräche mit dem Koordinator, Prof. Streit vom KIT, statt. Hier wurde vereinbart, durch regelmäßige Workshops beider Projekte, die gegenseitige Anschlussfähigkeit sicherzustellen.

Auch auf **europäischer und internationaler Ebene** werden zurzeit vor allem Forschungsdatenrepositorien aufgebaut. Exemplarisch sei auf *ZENODO*[11] verwiesen, das im Rahmen des Forschungsrahmenprogramms 7 (FP7) der Europäischen Union am CERN entwickelt wurde. *ZENODO* ermöglicht es Forschenden, ihre Forschungsergebnisse (Daten und Publikationen, aber auch Quellcode und Multimedia-Materialien) über ein webbasiertes Repositorium zu teilen, wenn diese nicht bereits Speicher- und Publikationsmöglichkeiten über ein institutionelles oder disziplinäres Repositorium haben. Darüber hinaus gibt es zahlreiche europäische Forschungsdatenrepositorien, die disziplinspezifisch aufgestellt sind. Anstelle der Erwähnung einzelner solcher Repositorien sei auf die *Registry of Research Data Repositories* verwiesen (vgl. nächster Abschnitt). Auf europäischer Ebene bekommt das Thema Forschungsdaten über die jüngst eingesetzte High Level Expert Group *European Open Science Cloud* zusätzlichen Auftrieb. Diese Gruppe soll Empfehlungen für die Entwicklung einer europäischen Infrastruktur für Forschungsdaten erarbeiten. Weiterhin sind etwa die Datenrepositorien *Dryad*, *Figshare* und *EUDAT* zu erwähnen. *Dryad*[12] ist ein universelles Datenrepository, das mit Mitteln der NSF in den USA gestartet wurde und derzeit als Non-Profit-Organisation betrieben wird. Es setzt sich zum Ziel, die abgelegten Forschungsdaten auffindbar, nachnutzbar und zitierbar zu machen. Bemerkenswert ist auch das von Macmillan Publishers betriebene Open Access-Repository *Figshare*[13], das allerdings keinen ausschließlichen Fokus auf Forschungsdaten hat, sondern gleichermaßen als Repository für wissenschaftliche Bilder und Videos auftritt. Erwähnenswert ist,

[8] http://www.pangaea.de
[9] http://www.ratswd.de/forschungsdaten/fdz
[10] http://www.escidoc.org
[11] https://zenodo.org/
[12] http://datadryad.org/
[13] http://figshare.com/

dass *Figshare* Gründungsmitglied der *Reproducibility Initiative*[14] ist, die sich zum Ziel setzt, die Reproduzierbarkeit von Forschungsdaten zu fördern. *EUDAT*[15] ist schließlich eine pan-europäische Infrastruktur, die Dienste für Forschungsdaten, aber auch Training und Beratung zum Thema anbietet.

Verzeichnisse zur Beschreibung von Forschungsdaten-Repositorien bieten das initial von der DFG finanzierte *Registry of Research Data Repositories* (re3data.org)[16] bzw. das international aufgestellte Portal *Databib*[17]. Beide verfolgen das Ziel, Forschenden einen schnellen Überblick über relevante Repositorien zur Speicherung und Auffindbarkeit von Forschungsdaten zu liefern. Beide Nachweissysteme werden verschmolzen und unter der Schirmherrschaft des international aufgestellten *DataCite*[18] weiter betrieben. *DataCite* ist ein internationales Konsortium (mit der ZBW als Mitglied), das sich u.a. zum Ziel setzt, einen einfachen Zugang zu wissenschaftlichen Forschungsdaten zu ermöglichen. Abschließend sei auf internationaler Ebene noch die *Research Data Alliance*[19] (*RDA*) genannt. Die *RDA* ist ein internationaler Zusammenschluss von derzeit mehr als 3000 Forschenden aus mehr als 100 Ländern. Die *RDA* zielt auf die Entwicklung und Förderung der Akzeptanz von Forschungsdateninfrastrukturen ab. Dies umfasst u.a. die Entwicklung von Standards, Handlungsempfehlungen und Policies sowie die Entwicklung von Ideen zur besseren Auffindbarkeit und Nachnutzung von Forschungsdaten.

Eigene Vorarbeiten

ZBW – Leibniz-Informationszentrum Wirtschaft
Die ZBW – Leibniz-Informationszentrum Wirtschaft ist die weltweit größte Spezialbibliothek für Wirtschaftswissenschaften. Als nationale Informationsinfrastruktureinrichtung arbeitet die ZBW seit vielen Jahren auf dem Gebiet des elektronischen Publizierens. So betreibt sie das fachliche Repository *EconStor*, das über 90.000 Publikationen wirtschaftswissenschaftlicher Fakultäten, Institute und Fachgesellschaften frei verfügbar als Volltext bereitstellt. Zudem betreibt die ZBW das wirtschaftswissenschaftliche Fachportal *EconBiz*, in dem über 9 Millionen Literaturnachweise aus unterschiedlichsten Quellen nachgewiesen werden; pro Jahr wird *EconBiz* von ca. 3,6 Millionen „unique visitors“ genutzt. Im Jahr 2014 wurden die digitalen Volltexte der ZBW (lizensiert und open access) ca. 5,4 Millionen Mal aus aller Welt heruntergeladen.

Der Bereich Forschungsdatenmanagement bildet bereits seit mehreren Jahren einen wichtigen Innovationsschwerpunkt. Dies lässt sich u.a. anhand der folgenden Aktivitäten veranschaulichen:

- Im von der DFG finanzierten Projekt *IJEMD* (*International Journal of Economic Micro-Data – An Open Access DataJournal*) wird die Technologie für den Betrieb eines Peer Review Data Journal für wirtschaftswissenschaftliche Mikrodaten entwickelt.
- Im DFG-Projekt *European Data Watch Extended* (*EDaWaX*) arbeitet sie seit 2011 an der Entwicklung einer Software, die auf die Einbeziehung und Präsentation von publikations-bezogenen Forschungsdaten in wirtschaftswissenschaftlichen Fachzeitschriften zielt.
- Die ZBW hat im Rahmen des EU-Projekts *Economists Online* unter anderem ein Datenarchiv auf Basis der Software Dataverse als Pilotanwendung aufgebaut, um Datensupplemente empirischer Forschungspublikationen standardisiert aufzubereiten.
- Die ZBW ist Mitglied der *DataCite*-Initiative, die die Standardisierung der Zitation von Daten fördert. Gemeinsam mit GESIS betreibt sie die Forschungsdatenregistratur *da|ra*, die mittels der Vergabe von DOI-Namen die Voraussetzungen für eine dauerhafte Identifizierung, Lokalisierung und verlässliche Zitierbarkeit von Forschungsdaten schafft.

[14] http://validation.scienceexchange.com
[15] http://eudat.eu
[16] http://www.re3data.org/
[17] http://databib.org/about.php
[18] https://www.datacite.org/
[19] https://rd-alliance.org/

- In einem Drittmittelprojekt der Leibniz-Gemeinschaft entwickelt die ZBW zusammen mit dem GESIS sowie mit zwei Leibniz-Wirtschaftsforschungsinstituten (WZB, DIW) die Forschungsdatenplattform *SowiDataNet*[20]. Ziel des Projekts ist der Betrieb eines Datenrepositories speziell für kleinere Forschungsvorhaben und -gruppen an Instituten, die keine eigene Infrastruktur betreiben können oder wollen.
- In dem DFG-Projekt *Digitale Reichsstatistik* schließlich wurde ein Verfahren zur Retrodigitalisierung von Statistikdaten des ehemaligen Deutschen Reiches entwickelt mit der Zielstellung, die Daten in maschinell weiterverarbeitbarer Form bereitzustellen.

Neben dem Thema Forschungsdatenmanagement betreibt die ZBW aktiv Forschung im Bereich Science 2.0. Sie koordiniert den multidisziplinären Leibniz-Forschungsverbund *Science 2.0*, in dem Phänomene im Zusammenhang der Digitalisierung der Wissenschaft erforscht werden. Über diese Expertise ist die ZBW in zahlreiche forschungspolitische Arbeitsgruppen zum Thema eingebunden (so ist z.B. der Direktor der ZBW das einzige Mitglied aus Deutschland in der High Level Expert Group *European Open Science Cloud* der EC sowie Mitglied im *Rat für Informationsinfrastrukturen* der Gemeinsamen Wissenschaftskonferenz).

CAU – Christian-Albrechts-Universität zu Kiel
Die Arbeitsgruppe *Software Engineering* der CAU – Christian-Albrechts-Universität zu Kiel – befasst sich in Forschung und Lehre mit dem Software Engineering für parallele und verteilte Systeme. Neben der Grundlagenforschung etwa in DFG-Projekten engagiert sich die Arbeitsgruppe besonders für den Technologietransfer, sowohl in die softwareentwickelnde Industrie als auch in andere Wissenschaftsdisziplinen, in denen softwaregestützte Forschung betrieben wird. Der Kompetenzverbund *Software Systems Engineering* (*KoSSE*) bündelt die Softwaretechnik-Kompetenzen der Informatik-Fachbereiche an den Universitäten in Kiel und Lübeck mit IT-Unternehmen in Schleswig-Holstein. Prof. Hasselbring ist Sprecher der Kieler *KoSSE*-Projekte. Im Bereich Forschungsdatenmanagement sei auf die folgenden Aktivitäten hingewiesen:

- Die CAU konzipiert und realisiert im DFG-geförderten Projekt *PubFlow* die workflowbasierte Verknüpfung der Forschungsdatenerhebung (insbesondere Sensordatenerfassung), der Datenverarbeitung (durch mathematische Simulationen) und der Datenarchivierung und -publikation. Dieses Pilotsystem wird durch das Datenmanagementteam am GEOMAR (Helmholtz-Zentrum für Ozeanforschung Kiel) genutzt und evaluiert. Die ZBW ist assoziierter Partner in diesem Projekt.
- Im Exzellenzcluster *Ozean der Zukunft* ist Prof. Hasselbring einer der Principal Investigators und für die Koordination des Forschungsbereichs R10 *Ocean Observations* mitverantwortlich. Hier geht es insbesondere um das Management von Ozeanbeobachtungsdaten.
- In der Helmholtz Research School *Ocean System Science and Technology* (*HOSST*) ist Prof. Hasselbring einer der Principal Investigators und u.a. Erstbetreuer einer interdisziplinären Dissertation zu neuen Softwaretechnik-Ansätzen zur datenbasierten Simulation von Fischpopulationen.
- Für das durch die durch die Helmholtz-Gemeinschaft geförderte *Marine Network for Integrated Data Access* (*MaNIDA*) ist die CAU assoziierter Projektpartner und Prof. Hasselbring Mitglied im Steuerungsausschuss.
- In der *D-Grid-Initiative* koordinierte Prof. Hasselbring die Verbundprojekte *BIS-Grid* (*Grid-basierte Integration und Orchestrierung*) und *WISENT* (*Wissensnetz Energiemeteorologie*).

Neben dem Thema Forschungsdatenmanagement befasst sich die Arbeitsgruppe als Kernkompetenz im Software Engineering insbesondere mit Software-Architektur-Entwurf, Betrieb und Monitoring komplexer verteilter Systeme, Integration verteilter Systeme sowie Middleware für Grid- und Cloud-basierte Systeme.

[20] https://sowidatanet.de/

LRZ - Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften
Das Leibniz-Rechenzentrum (LRZ) der Bayerischen Akademie der Wissenschaften ist Mitglied des *Gauss Centre for Supercomputing* (GCS). Es ist gemeinsames Rechenzentrum der Ludwig-Maximilians-Universität München, der Technischen Universität München sowie der Bayerischen Akademie der Wissenschaften. Das LRZ verfügt über modernste, redundant versorgte Rechnerräume sowie eigene Schulungs- und Kursräume.

Das LRZ ist mit seinen 180 Mitarbeitern Kompetenzzentrum für Datenkommunikationsnetze, Kompetenzzentrum für technisch-wissenschaftliches Hochleistungsrechnen und Grid-Computing, Zentrum für Langzeitarchivierung sowie ein Zentrum für virtuelle Realität und Visualisierung von großen Datenmengen. Es betreibt das Münchner Wissenschaftsnetz, umfangreiche Platten- und automatisierte Magnetband-Speicher zur Sicherung großer Datenmengen, eine große Server-Hosting-Lösung für die Münchner Universitäten sowie zentrale NAS-Speicherlösungen, mehrere Rechen-Cluster sowie den europäischen Tier-0-Rechner *SuperMUC*.

Mehr als 50 LRZ-Wissenschaftler kooperieren derzeit mit nationalen oder internationalen Projekten in folgenden projektrelevanten Forschungsschwerpunkten: IT-Service-Management und IT-Sicherheit, mandantenfähige zentrale Verzeichnisdienste, E-Learning- und Multimedia-Umgebungen, Grid- und Cloud-Technologien, automatische Provisionierung und Abrechnung von virtuellen Serverumgebungen, Hochgeschwindigkeitsnetze, Digitalisierung und Langzeitarchivierung, energieeffizienter Betrieb von Data Centern, flexible und adaptive Betriebsmodi für wissenschaftliche Rechenressourcen sowie effiziente numerische Programmierung und parallele Ein- und Ausgabe.

Im Bereich der Geisteswissenschaften hat das LRZ mit dem Kooperationspartner Bayerische Staatsbibliothek (BSB) langjährige Erfahrungen in der Herstellung, Verarbeitung, Bereitstellung und der digitalen Langzeitarchivierung von großen Mengen an Metadaten, Digitalisaten und OCR-Volltexten aufbauen können, etwa anhand der Massendigitalisierungsprojekte *VD16 digital* (*Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. Jahrhunderts*) und der Public Private Partnership der BSB mit Google zur Digitalisierung von urheberrechtsfreien Büchern.

Das LRZ hostet darüber hinaus
- die Serviceinstanzen für den Bibliotheksverbund Bayern[21], dem regionalen Zusammenschluss von über 150 Bibliotheken unterschiedlicher Größenordnungen und Fachorientierungen in Bayern;
- *Bavarikon*[22], das Portal zur Kunst, Kultur und Landeskunde des Freistaats Bayern;
- das *Bibliothekarische Archivierungs- und Bereitstellungssystem* (*BABS*). Es vereint eine heterogene organisatorische und technische Infrastruktur für die Langzeitarchivierung und Bereitstellung von elektronischen Publikationen unterschiedlicher Art;
- den Langzeitarchivierungsservice auf Basis von Rosetta im Auftrag der BSB für die bayernweite Nutzung;
- und viele weitere Services des Münchner Digitalisierungszentrums an der BSB[23].

Als Institut der Bayerischen Akademie der Wissenschaften (BAdW) beteiligt sich das LRZ aktiv am „Münchner Zentrum für digitale Geisteswissenschaften“[24], mit den Schwerpunkten Servicehosting und Visualisierung. Das Zentrum vernetzt wichtige Interessens- und Fachgruppen im Münchner Raum, bindet Nachwuchswissenschaftler ein und vermittelt die lokalen Aktivitäten im Bereich der digitalen Geisteswissenschaften auch an Interessenten über München hinaus.

[21] http://www.bib-bvb.de/
[22] http://www.bavarikon.de/
[23] http://www.digital-collections.de
[24] http://dhmuc.hypotheses.org/

Um die Zusammenarbeit mit herausragenden Wissenschaftlergruppen im Münchner Raum auszubauen, hat das LRZ vor einigen Jahren die *Partnerschaftsinitiative Computational Sciences* ins Leben gerufen. Diese fokussiert sich insbesondere auf den Bereich der Umweltwissenschaften und intensiviert durch Diskussion neuer Anforderungen und gemeinsame Forschungsprojekte die Zusammenarbeit zwischen Wissenschaftlern und dem Rechenzentrum. Gerade für die Umweltwissenschaften spielt die Thematik des interdisziplinären Datenaustausches eine zentrale Rolle.

Als Mitglied von GCS ist das LRZ darüber hinaus aktiv am Betrieb der *PRACE-Forschungs-infrastruktur*[25] involviert, mit den Schwerpunkten Servicehosting, Anwendungsoptimierung, Anwendertraining, Marktanalyse sowie energieeffizienter Betrieb von Rechenzentrums-infrastrukturen.

TUD- Technische Universität Dresden
Das Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH) ist eine zentrale wissenschaftliche Einrichtung der Technischen Universität Dresden mit begleitenden Forschungen im vollen Spektrum der Aufgabengebiete. Gleichzeitig ist es als HPC-Zentrum Sachsens Mitglied in der *Gauß-Allianz*. Im Mai 2015 wurde ein neuer Maschinenraum im 1. Bauabschnitt des Lehmann-Zentrums (LZR) in Dienst genommen, der den energieeffizienten und sicheren Betrieb von Rechner und Speicherressourcen auf insgesamt 1250 qm – mit redundanter Stromversorgung und redundanter Kühlung – sicherstellt. Den Nutzern stehen mit dem ebenfalls im Mai 2015 installierten Hochleistungsrechner-Speicherkomplex (HRSK-II) insgesamt ca. 45.000 Prozessor-kerne zur Verfügung. Der Gesamtkomplex wurde bereits 2011 für das so genannte „datenintensive Rechnen“ konzipiert, ein Antrag zur Erweiterung des HPC-Systems im Hinblick auf notwendige Komponenten zur deutschlandweiten Versorgung von Big Data-Anwendungen im Kontext von *ScaDS* Dresden/Leipzig wurde beim Wissenschaftsrat gestellt und dort positiv bewertet.

Die Methodenforschung des ZIH ist eng mit den Anforderungen der Anwenderwissenschaften verknüpft. Damit Hochleistungsrechner ihrer Funktion als Forschungswerkzeug gerecht werden können, sind eine funktionierende Infrastruktur zum Umgang mit diesen Datenmengen sowie geeignete Werkzeuge zur Unterstützung der Anwender unabdingbar. Die Anwender sind zumeist Wissenschaftlerinnen und Wissenschaftler aus verschiedenen Gebieten, u.a. aus den Natur-wissenschaften, dem Ingenieurwesen oder den Lebenswissenschaften, die Leipziger Kollegen liefern vermehrt Aufgabenstellungen aus den E-Humanities und dem Business-Data-Umfeld.

Weitreichende Erfahrungen bei der Entwicklung von Methoden zur effizienten Verwaltung großer Datenmengen hat das ZIH u.a. bereits im Bereich Grid-Computing gesammelt. Hier ist es seit Jahren kontinuierlich an verschiedenen Forschungsprojekten (BMBF, EU) zum Teil federführend am Entwurf und der Implementierung von Datenmanagement-Software beteiligt. So entwickelt das ZIH im Rahmen der deutschen *D-Grid Initiative*, aber auch innerhalb des europäischen Projektes *Chemomentum* für Wissenschaftler (z.B. für Mediziner, Chemiker, Bioinformatiker) die Möglichkeit, ihre Forschungsdaten mit Hilfe verschiedener Grid-Technologien zu speichern und zu verwalten. Dabei wurde sowohl Unterstützung bei der Erarbeitung von Konzepten als auch der softwareseitigen Umsetzung geleistet. Als Ergebnis erfolgreich abgeschlossener Projekte stehen den Anwendern Datencontainer auf Basis von UNICORE, iRods und dCache zur Verfügung. Im Projekt *WisNetGrid* wurde eine generische Infrastruktur für „Wissen“ entwickelt, die Dienste zur Integration und Vernetzung von Daten und Metadaten in einen gemeinsamen auch Community-übergreifenden „Wissensraum“ bereitstellt. Methoden aus dem Umfeld des *SIOX*-Projektes, des 100/400 GBit-Testbeds und die Forschung an Self Defined Networks helfen, Muster in Daten-zugriffen zu erkennen und Daten effizient zwischen verschiedenen Standorten bzw. Verarbeitungs-schritten zu bewegen. Auf dem Gebiet der Performance-Analyse für parallele Anwendungen besitzt das ZIH eine weltweit anerkannte Expertise. Einschlägige Software-Werkzeuge, die am ZIH oder in Kooperation mit anderen Forschungsgruppen entwickelt werden, wie z.B. *Vampir* und *Score-P* sind auf vielen der größten HPC-Systeme weltweit im Einsatz, um die Performance und

[25] http://www.prace-ri.eu/

Skalierbarkeit von Anwendungsprogrammen zu steigern. Sie ermöglichen auch die Analyse des spezifischen I/O-Verhaltens, um Engpässe bereits im Ansatz zu erkennen. Für die Software-Werkzeuge selbst konnte bereits eine Skalierbarkeit bis zu mehreren hunderttausend Prozessen demonstriert werden

DFN-Verein - Verein zur Förderung eines Deutschen Forschungsnetzes e. V.
Der Verein zur Förderung eines Deutschen Forschungsnetzes e. V. (DFN-Verein) ist die zentrale Einrichtung der Wissenschaft in Deutschland für Entwicklung und Betrieb ihrer eigenen Kommunikationsinfrastruktur, dem Deutschen Forschungsnetz. Mit 342 institutionellen Mitgliedern (Stand April 2016) engagiert sich die überwiegende Mehrzahl der deutschen Hochschulen und Forschungseinrichtungen sowie forschungsnahen Unternehmen der gewerblichen Wirtschaft im DFN-Verein. Mit Tagungen und Workshops trägt der DFN-Verein zur Weiterbildung und zum Informationsaustausch seiner Anwender bei. Darüber hinaus stehen in mehreren Kompetenz-zentren Ansprechpartner bereit, um in wichtigen Fragen zur Nutzung der Dienste gezielte Hilfestellungen zu geben. Über das Deutsche Forschungsnetz nutzen derzeit Einrichtungen an 608 Standorten (Stand April 2016) Kommunikationsdienste für die Wissenschaft. Die zur Erbringung der Dienste erforderliche Kommunikationsinfrastruktur betreibt der DFN-Verein in Eigenregie. Das Dienstangebot umfasst u. a. Dienste der Konnektivität (*DFNInternet* und *DFN-VPN*), Mobilität (*DFNRoaming* und *eduroam*), Kollaboration und interpersonellen Kommunikation (*DFNVC* und *DFNFernsprechen*) sowie der Sicherheit (*DFN-PKI* und weitere Dienste des *DFN-CERT*).

Angesichts der absehbar wachsenden Abstützung zukünftiger IT-Infrastrukturen auf zuverlässige, sichere und hoch performante Kommunikationsdienste haben die Mitglieder die strategische Ausrichtung des DFN-Vereins erweitert. Ergänzend zu seiner Rolle als Betreiber des nationalen Kommunikationsnetzes für die Wissenschaft und Organisator dessen internationaler Einbettung positioniert sich der DFN-Verein auch als „Enabler von netzgestützten F&L-Prozessen“. Entsprechende Tätigkeitsfelder wurden in dem zuletzt 2013 aktualisierten „Rahmenprogramm der Entwicklungen des DFN-Vereins“ festgeschrieben. So schafft die *Authentifizierung- und Autorisierungs-Infrastruktur* (*AAI*) des Dienstes DFN-AAI das notwendige Vertrauensverhältnis sowie einen organisatorisch-technischen Rahmen für den Austausch von Benutzerinformationen zwischen nutzenden Einrichtungen und Anbietern von Ressourcen. Durch die Einbettung der DFN-AAI in den internationalen Dienst *eduGAIN* ist sichergestellt, dass dieser Austausch auch mit globalen Partnern erfolgen kann.

Mit der *DFN-Cloud* wurde ein Modell für föderierte Dienste im DFN-Verein entwickelt. Verschiedene wissenschaftliche Einrichtungen in Deutschland haben bereits große Erfahrung mit der Bereitstellung von Cloud-Diensten. Der DFN-Verein schafft für diese Einrichtungen ein vertragliches Rahmenwerk, in dem diese Cloud-Dienste von allen Teilnehmern am DFN genutzt werden können. In entsprechenden Forschungsvorhaben bringt der DFN-Verein die beteiligten Partner zusammen, um die *DFN-Cloud* zu nutzen, zu erproben und ständig weiterzuentwickeln.

Das Deutsche Forschungsnetz wird laufend um neue und innovative Dienste ergänzt. Zur Sicherung deren Einbettung in den Verbund weltweiter Forschungsnetze engagiert sich der DFN-Verein seit Jahren in international geförderten Projekten, vorrangig von der Europäischen Kommission. Aus dem aktuell laufenden 8. Rahmenprogramm Horizon 2020 sind zwei Projekte hervorzuheben:

- Der DFN-Verein ist mit 38 weiteren Europäischen Forschungsnetzen Partner in dem Projekt *GN4-1 Research and Education Networking – GÉANT*. Über dieses Projekt fördert die Europäische Kommission den Betrieb und den Ausbau des Europäischen Verbindungsnetzes *GÉANT* zwischen den nationalen Forschungsnetzen in den beteiligten Ländern. Neben der Bereitstellung dieser paneuropäischen Netzinfrastruktur werden in dem Projekt auch Technologien und Infrastrukturen für Dienste u. a. in den Bereichen Trust&Identity sowie Clouds erprobt und für den flächendeckenden Einsatz pilotiert.

- Der DFN-Verein ist Partner im Europäischen Projekt *Authentication and Authorization für Research and Collaboration* (*AARC*). Ziel des Projekts ist die Harmonisierung der Dienste zur Authentifizierung und Autorisierung von Datenzugriffen zwischen verschiedenen Communities. Hierbei vereint *AARC* die Anforderungen der Anwender in großen Europäischen Forschungsprojekten wie *ELIXIR*, *EUDAT*, *DARIAH* oder *CLARIN* mit den Ressourcen der etablierten eInfrastructures *GÉANT*, *EGI* und *PRACE*.
- Der DFN-Verein ist assoziiertes Mitglied in der *Gauß-Allianz*. Der DFN-Verein ist Mitglied in der *GÉANT Association*, ein Geschäftsführer ist Chairman des Board of Directors. Der DFN-Verein ist Mitglied im *Shibboleth Consortium* und gewählter Vertreter im Consortium Board.

1.1 Projektbezogene Publikationen

1.1.1 Veröffentlichte Arbeiten aus Publikationsorganen mit wissenschaftlicher Qualitätssicherung, Buchveröffentlichungen sowie bereits zur Veröffentlichung angenommene, aber noch nicht veröffentlichte Arbeiten

A. Latif; A. Scherp; K. Tochtermann (2015): LOD for Library Science: Benefits of Applying Linked Open Data in the Digital Library Setting; German Journal on Artificial Intelligence, Springer, 2015, ISSN (Online) 1610-1987, ISSN (Print), http://link.springer.com/article/10.1007/s13218-015-0420-x

A. Hajra; V. Radevski; K. Tochtermann (2015): Author Profile Enrichment for Cross-linking Digital Libraries. Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries, LNCS Springer, Poznan, Poland. 14.-18.9.2015, pp 124-136.

K. Tochtermann (2014): How Science 2.0 will impact on Scientific Libraries. In: it - Information Technology. Band 56, Heft 5, S. 224–229, ISSN (Online) 2196-7032, ISSN (Print) 1611-2776, DOI: 10.1515/itit-2014-1050, September 2014.

W. Hasselbring (2015): Formalization of Federated Schema Architectural Style Variability. In: Journal of Software Engineering and Applications, 8(02). pp. 72-92.

P. Brauer, A. Czerniak, W. Hasselbring (2014): Start Smart and Finish Wise: The Kiel Marine Science Provenance-Aware Data Management Approach. In: 6th USENIX Workshop on the Theory and Practice of Provenance (TaPP 2014), June 2014.

P. Brauer, W. Hasselbring (2013): PubFlow: a scientific data publication framework for marine science. International Conference on Marine Data and Information Systems, Sept. 2013.

B. Peherstorfer, C. Kowitz, D. Pflüger, H.-J. Bungartz (2015): Selected Recent Applications of Sparse Grids; Numerical Mathematics: Theory, Methods, & Applications 8(1), pp. 47-77.

B. Peherstorfer, P. Gomez, H.-J. Bungartz (2015): Reduced Models for Sparse Grid Discretizations of the Multi-Asset Black-Scholes Equations; Adv. Computat. Math., Springer, (accepted).

B. Peherstorfer, D. Butnaru, K. Willcox, H.-J. Bungartz (2014): Localized Discrete Empirical Interpolation Method; SIAM J. Sci. Comp. 36(1), pp. A168-A192, SIAM, 2014.

L. Hämmerle, A. Harding, W. Pempe (2015): DARIAH Integration with eduGAIN – Technical Architecture. Joint Project Report GN3plus and DARIAH-EU, April 2015.

L. Hämmerle, W. Pempe (2014): Enabling Users: Options for Joining eduGAIN. Project Report GN3plus, March 2014.

A. Bode, R. Borgeest (2010): Informationsmanagement in Hochschulen; ISBN 978-3-642-04719-0466 pp., Springer Verlag Heidelberg 2010.

A. Bode, C. Trinitis (2009): Computer Hardware Development as a Basis for Numerical Simulation In: E.H. Hirschel et al (eds.): 100 Vol of Notes on Num. Fluid Mech., NNFM 100, pp 473-480, Springer Verlag Berlin, ISBN 978-3-540-70904-9, 2009.

A. Bode, V. Magliaris, D. Oleson, R. Saracco, P. Tindemans, Z. Turk, P. Veiga, M. Sharpe (2011): Knowledge without Borders, GEANT 2020 as the European Communications Commons Report of the GEANT Expert Group; European Commission, Information Society and Media, 53 pp. ISBN 978-92-79-21036-5

J Krüger, R. Grunzke, S. Gesing, S. Breuers, A. Brinkmann, L. de la Garza, O. Kohlbacher, M. Kruse, W.E Nagel, L. Packschies, R. Müller-Pfefferkorn, P. Schäfer, C. Schärfe, T. Steinke, T. Schlemmer, K.D. Warzecha, A. Zink, S. Herres-Pawlis (2014) The MoSGrid science gateway–a complete solution for molecular simulations, in: Journal of Chemical Theory and Computation 2014/10, 2232-2245.

T. Ilsche, J. Schuchart, J. Cope, D. Kimpe, T. Jones, A. Knüpfer, K. Iskra, R. Ross, W.E. Nagel, S. Poole, S. (2012) Enabling event tracing at leadership-class scale through I/O forwarding middleware, In: Proceedings of the 21st international symposium on High-Performance Parallel and Distributed Computing, 49-60.

1.1.2 Andere Veröffentlichungen

1.1.3 Patente

1.1.3.1 Angemeldet

entfällt

1.1.3.2 Erteilt

entfällt

2 Ziele und Arbeitsprogramm

2.1 Voraussichtliche Gesamtdauer des Projekts

36 Monate

2.2 Ziele

Das geplante Vorhaben erstreckt sich insgesamt über sechs Jahre, wobei der vorliegende Antrag die Finanzierung für die ersten drei Projektjahre umfasst.
Im Rahmen der ersten Phase des Projektes werden die Gesamtarchitektur sowie ausgewählte Piloten für eine virtuelle *Generic Research Data Infrastructure* (*GeRDI*) konzipiert, umgesetzt und evaluiert. Für diesen Zweck werden drei physische und über ein föderiertes System miteinander vernetzte Forschungsdatensysteme an drei Standorten eingerichtet. Diese Standorte haben unterschiedliche institutionelle Anbindung und werden für dieses Projekt unterschiedliche Einrichtungstypen repräsentieren. Mehrere als Anlage beigefügte "Letters of Intent" aus den anvisierten Fachcommunities belegen die Notwendigkeit und das Interesse der Wissenschaftler an der Mitarbeit an und Umsetzung von *GeRDI*. Zudem sind in dem Projekt eigene Ressourcen für das Community-Management vorgesehen (vgl. TP 3.4).
Der DFN-Verein als Partner repräsentiert keine Fachdisziplin und wird auch kein Pilotzentrum einrichten. Als ein wichtiger Vertrauenspartner im deutschen Wissenschaftssystem verfügt der DFN-Verein aber über Basistechnologien, die schon im breiten Einsatz sind und für *GeRDI*genutzt werden sollen. Hinsichtlich ihrer Eigenschaften als Einrichtungstypen lässt sich Folgendes zu den Einrichtungen sagen: Mit der TU Dresden ist eine Universität eingebunden, die eines der zwei vom BMBF geförderten Big Data-Kompetenzzentren in Deutschland betreibt. Mit dem LRZ München beteiligt sich ein bestehendes Höchstleistungsrechenzentrum am Projekt, über das insbesondere die Verbindung zum *Gauss Centre for Supercomputing* (GCS) und zur *Gauß-Allianz* sowie zu deren Anstrengungen im Bereich Management von Forschungsdaten sichergestellt ist. Die CAU repräsentiert die Bedürfnisse einer Universität hinsichtlich des Managements von Forschungsdaten, die aus einem Exzellenzcluster in Zusammenarbeit mit einem großen außeruniversitären Institut (GEOMAR) heraus entstehen. Mit der ZBW ist eine außeruniversitäre Informationsinfrastruktureinrichtung eingebunden, über die demonstriert wird, wie sich ein neues Aufgabenspektrum für Bibliotheken aus dem Projekt heraus entwickeln kann. Die ZBW hat zudem engste Verbindungen zu den zentralen Datenzentren in den Wirtschaftswissenschaften (*Sozio-ökonomisches Panel* und *Rat für Sozial- und Wirtschaftsdaten*).
Schließlich war ein Kriterium für die Zusammensetzung des Konsortiums die bestmögliche und möglichst vielfältige Sicherstellung der Anschlussfähigkeit an europäische Entwicklungen. Hierzu sei noch einmal auf die Partnerbeschreibungen verwiesen.

GeRDI besteht im Wesentlichen aus den folgenden Komponenten, die in den anschließenden Arbeitspaketen ausführlicher beschrieben werden (vgl. Abb. 2):

(1) **Speichermanagement**, wobei hier unterschiedliche Methoden erprobt werden sollen, sowie **Hardware** zum Speichern der Forschungsdaten.
(2) **Generische Dienste** für das Management von Forschungsdaten. Hier werden die Dienste angesiedelt, die disziplinunabhängig und für alle Forschungsdatensysteme (FDS) gleich sind (z.B. speichern, auslesen, suchen). Die generischen Dienste bieten auch elementare Mechanismen für Datensicherheit, Datenschutz etc. sowie – sofern möglich – den Zugang zu Forschungsdatenzentren, die außerhalb von *GeRDI* liegen.
(3) **Schnittstelle für spezifische disziplinäre Dienste**: Innerhalb einer Disziplin kann es zu unterschiedlichen Anforderungen kommen, für die disziplinäre Dienste erforderlich sind. Da es unrealistisch ist, das ganze Spektrum möglicher disziplinärer Dienste abzudecken, werden standardisierte Schnittstellen geschaffen, die das Andocken von vielfältigen – auch extern entwickelten – Diensten an *GeRDI* erlauben.

(4) **Disziplinäre Dienste**, die die Funktionalitäten für einzelne Disziplinen abbilden. Im Rahmen des Projekts werden diese entsprechend zuvor erstellter Fallstudien entwickelt. Hierunter fallen auch solche Dienste, die je nach Disziplin z.B. Aspekte der Sicherstellung von Privatsphäre, Datenschutz, Datensicherheit etc. abbilden.

(5) **Middleware-Komponente** für die Kommunikation und den Datenaustausch zwischen verschiedenen Forschungsdatenrepositorien innerhalb der *GeRDI* (z.B. Weiterleiten von Suchanfragen, Zusammenführen der Suchergebnisse aus den unterschiedlichen Repositorien). Es wird ein föderierter Ansatz zur Datenaufbewahrung gewählt, daher kommt der Entwicklung eines Kommunikationsmodells besondere Bedeutung zu. Hierbei liegt ein besonderer Grad der Komplexität in der vertikalen Kommunikation über alle Ebenen des Layermodells. Aber auch die horizontale Kommunikation zwischen Datenrepositorien ist sicherzustellen.

Abbildung 2 zeigt zudem auf der rechten Seite den wichtigen Aspekt, dass neben universitären Zentren auch die Anbindung von bereits bestehenden Forschungsdatenzentren exemplarisch umgesetzt wird. Diese können neben ihren existierenden disziplinären Diensten auch die generischen Dienste bzw. die Forschungsdatensysteme von *GeRDI* nutzen, sofern sie die in *GeRDI* eingesetzten etablierten Standards berücksichtigen.

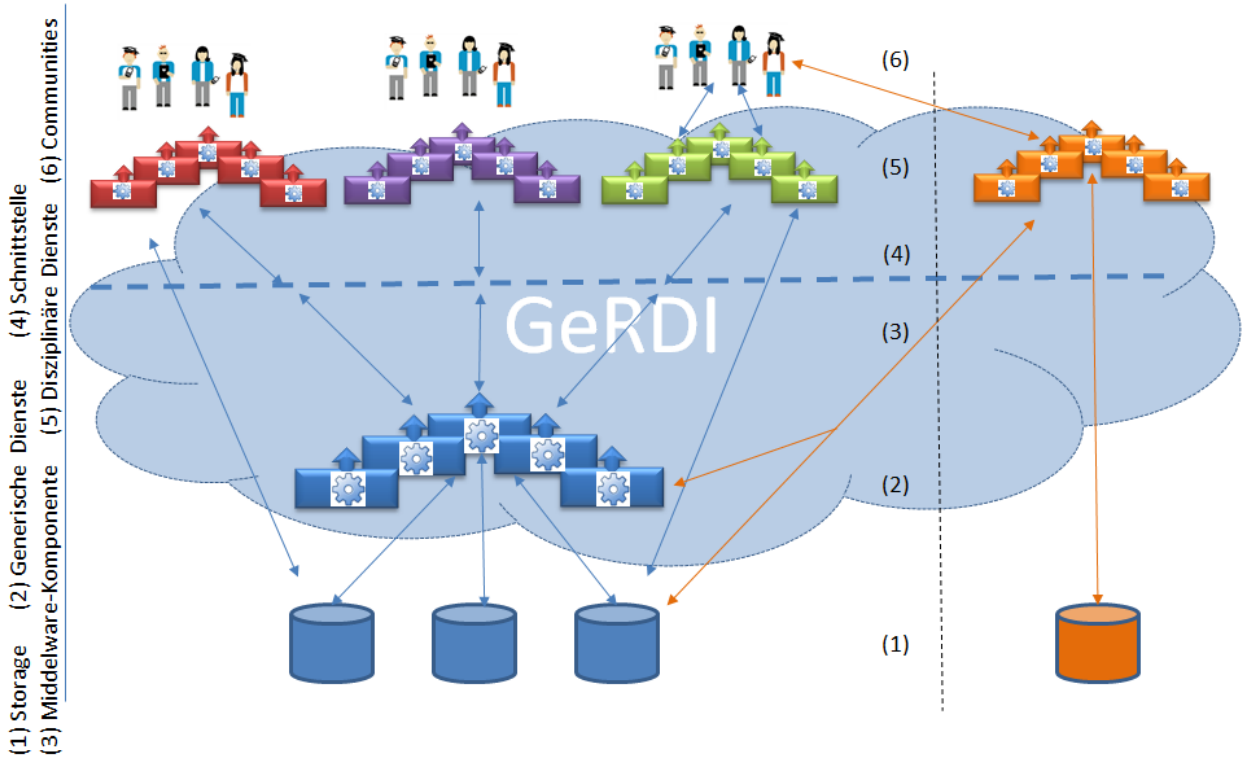

Abb. 2: Idee der *Generic Research Data Infrastructure GeRDI*

Für einen Proof-of-Concept werden an den drei Standorten Dresden, Kiel und München Pilotzentren aufgesetzt und anhand der genannten Disziplinen entlang zuvor festgelegter Fallstudien evaluiert. Die Fallstudien werden dabei gleichermaßen folgende Ausprägungen berücksichtigen:

1. Anbindung von *GeRDI* an universitäre Rechenzentren
2. Anbindung von *GeRDI* an Höchstleistungsrechenzentren
3. Einbindung von bestehenden Forschungsdatenzentren in *GeRDI*
4. Disziplinenübergreifender Zugriff und Nutzung von Forschungsdaten mittels *GeRDI*

Schließlich werden als weiteres Ergebnis dieser Projektphase eine Vorgehensweise sowie Empfehlungen entwickelt, wie diese modellhafte Infrastruktur in einer zweiten Projektphase für das

gesamte Wissenschaftssystem in Deutschland flächendeckend ausgerollt werden kann. Dieses Ergebnis beinhaltet ein in Projektphase 1 entwickeltes Betriebsmodell, über das der nachhaltige Betrieb der Infrastruktur sichergestellt werden kann.
Das Projekt verfolgt grundsätzlich den Ansatz, dass – wo immer möglich – bereits bestehende Softwarekomponenten nachgenutzt und ggf. auf die projektspezifischen Anforderungen hin weiterentwickelt werden.

2.3 Arbeitsprogramm und Umsetzung

Das Arbeitsprogramm besteht aus fünf Arbeitspaketen (AP), von denen die Arbeitspakete 1-4 jeweils Teilpakete (TP) haben.

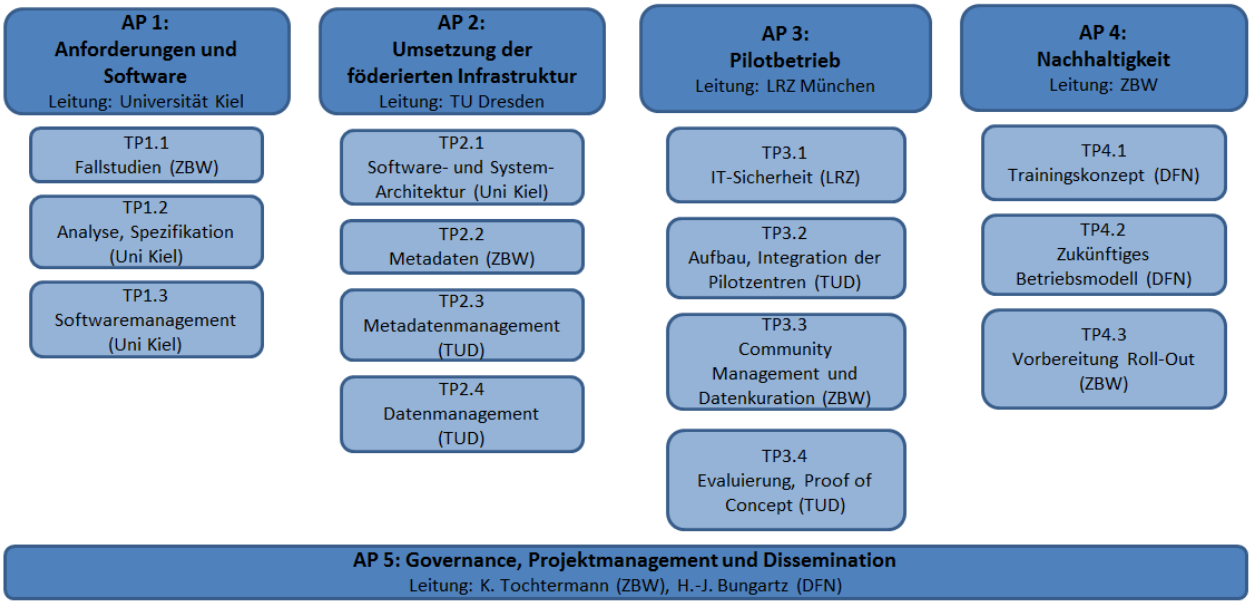

Abb. 3: Arbeitspakete

Die Arbeitspakete werden nachfolgend näher beschrieben.

AP1: Anforderungen und Software (Leitung: CAU)

Dieses Arbeitspaket befasst sich mit der inhaltlichen Auseinandersetzung von möglichen Einsatzszenarien von *GeRDI*. Für diesen Zweck entwickelt TP 1.1 entsprechende Fallstudien, die in allen Phasen des Projekts Berücksichtigung finden werden und zudem die Grundlage für die Anforderungsspezifikation bilden. Auf Basis der Fallstudien werden in TP 1.2 die Anforderungen analysiert und spezifiziert. TP 1.3 befasst sich mit dem Management aller Softwareentwicklungen, die im Rahmen von *GeRDI* vorgenommen werden.

TP 1.1 Fallstudien (Leitung: ZBW)

In diesem TP werden Fallstudien für die Erprobung und das Proof-of-Concept der drei Pilotzentren entwickelt. Diese Fallstudien werden in enger Abstimmung mit den entsprechenden Fach-Communities aus den Anwendungsdomänen des Projekts entwickelt. Ausgangspunkt wird die Sammlung von Use Cases der *RDA-Initiative*[26] bilden. Auf Basis von Interviews mit den Forschenden aus der jeweiligen Fach-Community bzw. auf Basis von Beobachtungen von deren Arbeitsabläufen werden Personas abgeleitet. Für diese Personas werden unterschiedliche Nutzungsszenarien durchgespielt und als Fallstudien dokumentiert. Damit ist eine Grundlage geschaffen, um im nachfolgenden TP 1.2 die Softwareanalyse und Anforderungsspezifikation durchzuführen.

[26] https://rd-alliance.org/use-cases.html-3

Um das System in seiner ganzen Komplexität evaluieren zu können (vgl. TP 3.3), sind die Fallstudien derart, dass sie gleichermaßen generische und disziplinäre Dienste benötigen. Das Arbeitspaket ist ein Querschnittspaket, das in allen AP und TP Berücksichtigung findet. Dadurch wird die Ausrichtung aller Entwicklungsarbeiten an praxisrelevanten Fragestellungen sichergestellt.

TP 1.2 Analyse, Spezifikation (Leitung: CAU)

In diesem TP wird die erste Anforderungsanalyse, die im Rahmen des Projektantrags durchgeführt wurde, fortgeschrieben und hin zu einer Anforderungsspezifikation entwickelt, in der die Anforderungen ausführlich dokumentiert sind. Für diesen Zweck wird zwischen technischen Anforderungen sowie Nutzungsanforderungen unterschieden. Diese Anforderungen dienen als Grundlage für die Evaluierungen in TP 3.4.

Technische Anforderungen betreffen

- Speichereffizienz
- Kommunikationsperformance zwischen den FDS
- die optimale Nutzung verfügbarer Netzressourcen zum Datentransfer
- die Skalierbarkeit für hohe Datenvolumina
- Möglichkeiten zur Anbindung zu bereits bestehenden Datenzentren
- Aspekte der digitalen Langzeitarchivierung von Forschungsdaten
- die Nutzung von technischen Standards
- Erfordernisse zur Zertifizierung über ein Datensiegel, wie das *Data Seal of Approval*[27]

Nutzungsanforderungen betreffen

- Stakeholder-Perspektiven, die die unterschiedlichen Interessenlagen deutlich machen. So werden Drittmittelgeber, Verlage, Forschende, Infrastruktureinrichtungen oder Nachnutzer von Forschungsdaten einen unterschiedlichen Blickwinkel auf die Anforderungen haben, die bei der Konzeption und Umsetzung einer solchen Infrastruktur zu berücksichtigen sind;
- Anforderungen an generische Funktionalitäten wie Speichern und Auslesen von Forschungsdaten;
- die Schnittstelle und Grenzen zwischen generischen Funktionalitäten und disziplinspezifischen Funktionalitäten;
- den Übergang von der Erhebung der Forschungsdaten zur Ablage in ein Forschungsdatensystem FDS;
- den Übergang *GeRDI* zu Rechenzentren, die auf den Daten „rechnen“, sie weiteren Analysen unterziehen und ggf. neu generierte Forschungsdaten zurückspielen;
- die Verbindung zu bestehenden Datenzentren.

Schließlich soll in diesem Arbeitspaket auch der ökonomische Nutzen des Vorhabens für das wissenschaftliche Gesamtsystem in Deutschland untersucht und bewertet werden. In der Anforderungsanalyse sind ferner bereits vorhandene Datenmanagementpläne zu berücksichtigen, wie sie z.B. für das *WissGrid*-Projekt[28] entwickelt wurden oder seitens Forschungsförderern wie der DFG[29] mittlerweile standardmäßig eingefordert werden.

Konkret sollen die Anforderungen u.a. mit der folgenden Technik erhoben und dokumentiert werden: „Behavior Driven Development"[30] ist eine Technik der Softwareentwicklung, welche die Zusammenarbeit zwischen Qualitätsmanagement und Anforderungsanalyse in Softwareentwicklungsprojekten stärkt. Beim Behavior Driven Development werden während der Anforderungsanalyse die Aufgaben, Ziele und Ergebnisse der Software derart festgehalten, dass diese später automatisiert auf ihre korrekte Implementierung getestet werden können (siehe TP

[27] http://datasealofapproval.org
[28] http://www.wissgrid.de/publikationen/Leitfaden_Data-Management-WissGrid.pdf
[29] http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/richtlinien_forschungsdaten_biodiversitaetsforschung.pdf
[30] https://dannorth.net/introducing-bdd/

1.3 zum Softwaremanagement). Ein wichtiger Aspekt ist dabei, dass die Anforderungen gemeinsam mit den (potenziellen) Nutzern der *GeRDI*-Infrastruktur in Nutzer-Workshops erhoben werden.

TP 1.3 Softwaremanagement (Leitung: CAU)

Für *GeRDI* wird in großem Umfang Software entwickelt, installiert, konfiguriert und in Rechenzentren betrieben. Sowohl für die Entwicklung als auch für den Betrieb dieser Software bestehen sehr hohe Qualitätsanforderungen.

In diesem TP werden dazu organisatorische, methodische und technische Aspekte der Softwareentwicklung in *GeRDI* koordiniert und unterstützt. Im Sinne eines „Continuous Software Engineering“ soll erreicht werden, dass die entwickelte Software schon während der Projektdurchführung eine kontinuierliche hohe Qualität erreicht.

In *GeRDI* soll Open-Source-Software entwickelt werden. Diese muss mit einem geeigneten Versionsmanagement-System verwaltet werden. Die konsistente und effiziente Nutzung eines entsprechenden Dienstes wie z.B. *GitHub*[31] muss durch dieses TP koordiniert werden.

Für die Priorisierung und Koordination von Anforderungen (siehe AP1), Entwicklungsaufgaben und Fehlerberichten ist es sinnvoll ein leistungsfähiges Aufgabenmanagementsystem wie *JIRA*[32] einzurichten und zu koordinieren. *JIRA* kann für Open-Source-Software kostenfrei genutzt werden. Es bietet sich an, die Anforderungsdokumente kollaborativ in einem Web-basierten System zu verwalten. Dazu bietet es sich an, *Confluence*[33] zu nutzen, was sehr gut mit *JIRA* integriert ist.

Kontinuierlich kann die Qualität nur durch automatisierte Mechanismen erreicht werden:

- Schon während der Entwicklung sollen parallel zur eigentlichen Funktionalität automatisierte Unit- und Integrationstests implementiert werden. Insbesondere soll hier auch testgetriebene Softwareentwicklung genutzt werden.
- Diese automatisierten Unit- und Integrationstests können dann auf einem „Continuous Integration“ Server regelmäßig als Regressionstests überprüft werden. Technisch bietet es sich an hierzu den *Jenkins-Server*[34] zu nutzen. Wie in AP1 dargestellt wurde, sollen die Anforderungen u.a. mit der Technik „Behavior Driven Development“ erhoben und dokumentiert werden.
- Zusätzlich zur *Continuous Integration* soll es durch ein *Continuous Deployment* ermöglicht werden, Softwarekomponenten, die bis dahin alle automatisierten Tests erfolgreich bestanden haben, möglichst automatisiert auf den jeweiligen Zielsystemen zu installieren. Hierdurch wird ein schnelles Testen durch die Nutzer ermöglicht und gleichzeitig können weitere automatisierte Tests (z.B. Lasttests) durchgeführt werden.

Diese Maßnahmen dienen insbesondere auch dazu, frühzeitig Feedback von den Nutzern erhalten zu können.

Weitergehende Qualitätssicherungsmaßnahmen können Tests auf Fehlertoleranz durch Fehlerinjektion betreffen (kontinuierliche Resilienz) sowie automatisierte Security-Checks. Zur Sicherstellung einer ausreichenden Performance können automatisierte Tests zum Regressions-Benchmarking in die *Continuous Integration* integriert werden.

Generell soll durch obige Methoden mit automatisierten Qualitätssicherungsmaßnahmen ein frühzeitiges Aufdecken von Problemen ermöglicht werden, damit diese sich gar nicht erst weitgehend auswirken können.

[31] https://github.com/
[32] https://de.atlassian.com/software/jira/
[33] https://de.atlassian.com/software/confluence/
[34] https://jenkins-ci.org/

Meilensteine von AP1:

Meilenstein	Bezeichnung	Zeitpunkt	Verantwortlich
1.a	Anforderungsspezifikation – erster Entwurf	M12	CAU
1.b	Finale Anforderungsspezifikation für die Fallstudien	M24	CAU
1.c	Evaluation der Erfüllung der Anforderungen	M36	CAU

AP2: Umsetzung der föderierten Infrastruktur (Leitung: TU Dresden)

Die praktische Umsetzung des föderierten Systems für Forschungsdaten auf der Basis vorhandener Systeme wird in diesem Arbeitspaket realisiert. Zunächst steht der Entwurf einer Gesamtarchitektur aus Hard- und Software auf dem Programm (TP 2.1). Deren drei wesentliche Aspekte Struktur und Inhalte von Metadaten (TP 2.2), deren Management (TP 2.3) sowie das Datenmanagement (TP 3.4) werden in den weiteren Unterarbeitspaketen realisiert.

TP 2.1 Software- und Systemarchitektur (Leitung: CAU)

In diesem TP wird eine offene Software-Architektur des Gesamtsystems entworfen. Das Gesamtsystem ist dabei ein virtuelles FDS, das unterschiedliche physikalische FDS umfassend integriert und föderiert. Über die Offenheit der Softwarearchitektur wird sichergestellt, dass bereits bestehende FDS anschlussfähig sind, in dem sie über standardisierte Schnittstellen auf Daten aus *GeRDI* zugreifen bzw. ihrerseits zur Verfügung stellen können.

Während in TP 1.2 mit den Anforderungen spezifiziert, *was GeRDI* leisten wird, wird hier erarbeitet wie *GeRDI* auf Architekturebene diese Anforderungen erfüllen wird.

Die Architektur basiert auf Middleware-Komponenten, über die die Kommunikation zwischen physikalischen FDS mittels geeigneter Protokolle für den (Meta)Datenaustausch (z.B. SFTP, rsync, P2P, Replikationsprotokolle, OAI-PMH, SOAP, CDMI, S3) erfolgt. Zudem legt die Software-Architektur die Grenze zwischen generischen Funktionalitäten/Diensten sowie die Schnittstellen zu disziplinspezifischen Funktionalitäten/Diensten eines FDS fest. Die Architektur beinhaltet auch eine Datenarchitektur, die das Management der Forschungsdaten technisch/konzeptionell und inhaltlich abbildet. Hierzu ist auch eine enge Abstimmung mit den TP 2.3 zum Metadatenmanagement und TP 2.4 zum Datenmanagement geplant. Die Schnittstellenspezifikationen der generischen Dienste sind offen, so dass die Anschlussfähigkeit bestehender Forschungsdatenzentren an *GeRDI* gegeben ist.

Ein wichtiger Punkt ist dabei auch die Verfügbarmachung der Daten auf Rechenressourcen in Rechenzentren, um die Daten für Analysen bereitzustellen. Hier sind u.a. HPC-Infrastrukturen zu nennen, die z.B. über die Middleware *UNICORE* – die im europäischen Flagschiff *Human Brain Project* und in der amerikanischen *XSEDE* Forschungsinfrastruktur eingesetzt wird – angebunden werden können und Cloud-Infrastrukturen mit Schnittstellen wie *CDMI* und S3.

Um Nutzern ein Single Sign-On zu ermöglichen, kommen bereits etablierte Dienste zum Einsatz. Ziel soll es dabei sein, eine dem *eduroam* ähnliche Nutzerfreundlichkeit einer organisations-übergreifenden Authentifizierung zu erreichen. Ausgangspunkt ist die Authentifikations- und Autorisierungs-Infrastruktur des DFN-Vereins (DFN-AAI) auf Basis von *Shibboleth*, das weltweit von zahlreichen Einrichtungen für lokales, aber auch nationales AAI eingesetzt wird. Für die europäische Anschlussfähigkeit von *GeRDI* wird die Kompatibilität mit der europäischen Variante von DFN-AAI, *eduGAIN* sichergestellt.

Für eine Infrastruktur wie *GeRDI* ist es wichtig, während der gesamten Entwicklung den späteren Betrieb der Software zu berücksichtigen. Für den Betrieb in Rechenzentren ist es essentiell wichtig die dort betriebenen Softwaresysteme kontinuierlich beobachten zu können (Monitoring). Die

Integration von Monitoring-Frameworks und Monitoring-Agenten wird folglich schon frühzeitig in der Softwarearchitektur berücksichtigt. Dazu sollen bewährte Monitoring-Frameworks wie *Nagios*[35], *Centreon*[36], *Score-P*[37] und *Kieker*[38] von vornherein auf Architekturebene integriert werden.

Für den Betrieb ist es weiterhin wichtig, Fehlertoleranz-Mechanismen in das System zu integrieren. Dazu gehört die Isolation von Fehlerauswirkungen in einer modularen Softwarearchitektur, sowie die Planung redundanter Komponenten, die im Fehlerfall Notfallaufgaben und ggf. einen Notbetrieb gewährleisten können. Eine Aufgabe besteht darin, für den Betrieb abgestufte Dienstleistungsebenen zu identifizieren, die dann auf Architekturebene je nach Kritikalität mit unterschiedlichen Anforderungen für Ausfallsicherheit, Zuverlässigkeit, Performance und Skalierbarkeit in der Architektur integriert werden müssen.

Der Sicherheit gegen Angriffe (Security) und der Datensicherheit gegen Datenverluste kommt in *GeRDI* ebenso eine wichtige Rolle zu, die auch geeignet auf Architekturebene berücksichtigt werden muss. Intrusion Detection Systeme müssen wirksam integriert werden, wie auch die Isolation von Einbruchsauswirkungen in einer modularen Softwarearchitektur. Für die Datensicherheit werden effiziente Datenredundanz- und Replikationstechniken sowie Backup-Mechanismen in der Architektur integriert. Hierzu ist auch eine enge Abstimmung mit dem TP 3.1 zur IT-Sicherheit geplant.

Digitale Langzeitarchivierung von Forschungsdaten ist ein wichtiger Aspekt des Data Life Cycle Managements und wird im Projekt stets mitgedacht, ist aber in der ersten Phase aufgrund der anders gearteten Funktionalitäten und Eigenschaften kein Schwerpunktthema des Projekts. Hier fließen die Erfahrungen der folgenden Partner ein: Die ZBW setzt seit mehr als 2 Jahren das System *Rosetta* der Firma ExLibris im operativen Betrieb für digitale Langzeitarchivierung von Publikationen ein. Das Zentrum für Informationsdienste und Hochleistungsrechnen der TU Dresden betreibt bereits seit mehreren Jahren ein Langzeitarchiv für Forschungsdaten und baut zurzeit im Projekt *OpARA* ein institutionelles Repositorium für Forschungsdaten auf. Das LRZ betreibt die Langzeitarchivierung für das *Bibliothekarische Archivierungs- und Bereitstellungs-system* der BSB.

Zur Dokumentation und Spezifikation der Software-Architektur soll mehrstufig vorgegangen werden:

- Zunächst werden sogenannte arc42-Templates[39] genutzt, die sich in der Praxis sehr gut bewährt haben.
- Zur detaillierteren Architekturbeschreibung wird der Industriestandard UML[40] genutzt, insbesondere Komponentenstrukturdiagramme und Verteilungsdiagramme.
- Für besonders kritische Komponenten wird deren Spezifikation ergänzt durch eine modellbasierte Formalisierung zentraler Aspekte der Architektur.

Die Entwicklung soll nicht rein sequentiell als Wasserfall erfolgen. Die Softwareentwickung in den folgenden TP muss iterativ und inkrementell mit dem Architekturentwurf abgestimmt werden. Zu den jeweiligen Meilensteinen werden die jeweils aktuellen Stände festgehalten und geeignet dokumentiert.

[35] https://www.nagios.org/
[36] https://www.centreon.com/en/
[37] http://www.vi-hps.org/projects/score-p/
[38] http://kieker-monitoring.net/
[39] http://arc42.de/
[40] http://www.uml.org/

TP 2.2 Metadaten (Leitung: ZBW)

Im Wesentlichen befasst sich TP 2.2 mit der Strukturierung der vor allem inhaltlichen Metadaten, um hierdurch eine semantische Integration verteilter, von unterschiedlichen Fachcommunities genutzten FDS zu erreichen. Unter Qualitäts- und Integrationsaspekten werden dabei neben einer formalen, automatischen Validierung und Qualitätsprüfung (etwa mit Blick auf Schemakonformität) auch „data curation" betrieben und systematisiert. Mit Hilfe sog. Mapping-Tools wird gewährleistet, dass vorhandene bzw. neu hinzukommende Datenrepositorien vor allem auf der Ebene der beschreibenden Metadaten in das *GeRDI*-Netzwerk möglichst reibungslos integriert werden können. Hierfür werden Thesauri und kontrollierte Vokabulare genutzt, wobei sich auf der fachübergreifenden Ebene die Subject Headings der *Library of Congress* (*LCSH*) oder die bundesdeutsche *Schlagwortnormdatei* (*SWD*), auf der disziplinspezifischen Ebene der *Standard-Thesaurus Wirtschaft* (*STW*) für ein wirtschaftswissenschaftliches oder *GEMET* für ein geo- oder umweltwissenschaftliches Repository anbieten.

Darüber hinaus werden bereits existierende disziplinunabhängige Standards wie *DataCite* (Metadata Schema v3.1) untersucht und ggf. eingesetzt, ferner die bei den bereits existierenden FDS genutzten Standards einbezogen. Zudem ist die Anschlussfähigkeit an Standards, Projekten und Initiativen, die sich mit der Standardisierung und Interoperabilität von Metadaten befassen, sicherzustellen (z.B. *DDI* (Umfragedaten), *CERA-2* (Simulation im Klimabereich), *MASi*, die Community-spezifischen *Data Life Cycles* aus dem Projekt *Large-Scale Datenmanagement and Analysis* (*LSDMA*), die *RDA*-Metadaten-Arbeitsgruppen, sowie auf nationaler Ebene vor allem das *Kompetenzzentrum Interoperable Metadaten* (*KIM*) des *DINI e.V.*).

Über die inhaltlichen Metadaten soll im Wesentlichen die Integration der verschiedenen Repositorien zu einem auch multidisziplinär durchsuchbaren und nutzbaren Netzwerk erfolgen. Wichtig sind dafür insbesondere Metadaten, die den Datensatz als Ganzes oder einzelne seiner Variablen beschreiben. Dazu gehören auch Provenance-Informationen, die sich z.B. auf die Herkunft, die Entstehungs- und Auswertungsumgebung, ferner auf die Historie und Versionierung von Daten beziehen.

Eine wichtige infrastrukturelle, aber auch datentechnische Maßnahme wird die Einbeziehung bibliothekarischer Normdaten sein, letztere zunächst bezogen auf Personen, Institutionen und Sachgebieten. Obwohl aus dem bibliothekarischen Umfeld diese Normdaten größtenteils bereits vorliegen oder ggf. noch einmal aufbereitet werden müssten, sind ihre Einführung, Nutzung und Pflege im Kontext von Datenrepositorien ein eigenes und auch relativ neues Thema. Dabei sollen hinsichtlich der Einbindung und Pflege von Normdaten im Zuge klassischer bibliothekarischer und dokumentarischer Geschäftsgänge wie der Erschließung eines Datensatzes verschiedene Modelle erprobt werden, z.B. die (de-)zentrale Pflege durch Datenkuratoren oder auch die im Sinne eines Out- bzw. Crowdsourcing stärkere Einbeziehung der Forschungscommunity. Dies praktiziert z.B. der *RePEc Author-Service* im Kontext der Wirtschaftswissenschaften seit Jahren erfolgreich. Letztlich soll durch die Auszeichnung und Erfassung von Forschungs- mit Normdaten die datentechnische Grundlage geschaffen werden, um in einer späteren Phase dieses TP diese Daten zusammen mit anderen beschreibenden Metadaten als *Linked Open Data* zu veröffentlichen und dabei mit den (Norm-)Daten der anderen FDS zu verknüpfen.

Einen weiteren Schwerpunkt bildet die Verknüpfung der Daten mit sonstigen relevanten externen Fach- und Forschungsinformationen. So werden Entwicklungen und Ergebnisse auch aus anderen Projekten herangezogen, um Forschungsdaten z.B. mit Primärpublikationen, aber auch mit Informationen zu Personen, Projekten, Institutionen oder Lehrgebieten weitgehend automatisiert zu verknüpfen.

Im Rahmen des Arbeitspakets werden bereits bestehende Kontakte vor allem zu den speziell mit Metadaten befassten Arbeitsgruppen der *Research Data Alliance* (*RDA*) genutzt, um die von der *RDA* entwickelten Interoperabilitätsstandards einerseits umzusetzen und zu validieren, und andererseits in der Richtung einer stärkeren Anbindung an bibliothekarische Normdaten(dienste)

weiterzuentwickeln. Bei den *RDA-Initiativen* wären insbesondere zu nennen die beiden Arbeitsgruppen zu *Metadata Standards Directory* und *Metadata Standards Catalog*, ferner die übergeordnete *Interest Group* zu Metadaten.

TP 2.3 Metadatenmanagement (Leitung: TU Dresden)

TP 2.3 zielt darauf ab auf der Basis vorhandener Software ein technisches System zur Verwaltung der Metadaten im Rahmen der Gesamtarchitektur (TP 2.1) aufzubauen. Es soll alle Metadaten (inklusive der Community-spezifischen), die in TP 2.2 evaluiert und verifiziert werden, beinhalten. Dazu sollen Schemata für technische, administrative, deskriptive oder andere Metadaten abgebildet werden können.

Die Nutzer werden umfangreiche Suchmöglichkeiten erhalten. Diese Metadatensuche ist eine essentielle Kernkomponente im Hinblick auf die Sichtbarmachung und damit auch den Zugriff auf die in *GeRDI* verwalteten Forschungsdaten. Über ein ansprechendes Nutzerinterface hinaus wird auch eine Such-API zur Verfügung gestellt, über die die Metadaten oder der Suchindex durch dritte weiterverarbeitende Programme und Applikationen – z.B. Meta-Suchmaschinen oder bibliothekarische Nachweissysteme wie Kataloge und Fachportale – abgefragt werden können.

Weitere Fragestellungen zur technischen Realisierung sind die Skalierbarkeit des Systems, Adaptivität an neue Metadaten, Integration in die wissenschaftlichen Arbeitsabläufe und -umgebungen oder die Automatisierung der Metadatenerfassung.

Schließlich gehören zu den durch das *GeRDI*-Netzwerk generierten Metadaten auch Nutzungsdaten und deren Analyse auf zumindest zwei Ebenen: Zum einen werden durch die Zugriffe auf die FDS laufend Statistiken erzeugt, die aggregiert und über einen zentralen Dienst nach Prinzipien des *COUNTER*-Standards ausgewertet werden. Zum zweiten werden die *GeRDI*-Forschungsdaten in Publikationen zitiert, sodass hier eine bibliometrische Analyse bzw. darauf aufsetzende Dienste und Standards für die Impactmessung von Forschungsdaten zu entwickeln sind. Da sich diese, im engeren Sinne wissenschaftliche, Nutzung der Daten nur schwer prognostizieren lässt, werden die hauptsächlichen Arbeiten hierzu allerdings erst in der zweiten Projektphase verortet.

Für die Umsetzung wird eine Evaluierung vorhandener Entwicklungen zu Metadatensystemen durchgeführt. Dies schließt Systeme und Projekte ein, die sich integrativ mit Daten- als auch Metadatenmanagement beschäftigten. Hierzu gehören zum Beispiel *EUDAT*, *INDOGO-DataCloud* oder der *KIT Data Manager*. Ebenso werden die Ergebnisse der *Research Data Alliance* (z.B. bzgl. Metadaten-Katalogen), *LSDMA*, *MASi* und *RADAR* einfließen. Aufgrund der engen logischen Verbindung zwischen Daten- und Metadatenmanagement wird dies in enger Kooperation mit TP 2.4 geschehen.

Einfließen werden auch die Arbeiten, Erfahrungen und Ergebnisse des DFG-Projektes *MASi*, das von der TU Dresden geleitet wird. Es bietet ein integratives Daten- und Metadatenmanagement-system auf Basis des *KIT Data Managers*, das den Umgang mit großen Datenmengen unterschiedlicher Anwender-Communitys erlaubt. Eine grundlegende Eigenschaft von *MASi* ist die hohe Flexibilität in der Anpassbarkeit für unterschiedlichste Anwendungsfälle. Es ist geplant eine *MASi*-Integration mit *EUDAT* für den produktiven Einsatz in *GeRDI* zu evaluieren. *EUDAT* ist als Basistechnologie für das Datenmanagement und übergreifende Metadatenmanagement absehbar, wobei *MASi* insbesondere die Community-spezifischen Teile des Metadatenmanagements abdeckt. Diese Vorgehensweise ergibt einerseits eine hohe Anschlussfähigkeit an vorhandene Dateninfrastrukturen und andererseits eine maximale Flexibilität in Bezug auf Metadaten bei der Anbindung von unterschiedlichsten Anwendungsfällen.

Die Aspekte des Metadatenmanagements werden auf Basis der Software- und Systemarchitektur (TP2.1) und in enger Zusammenarbeit mit TP2.2 Metadaten, TP2.4. Datenmanagement, den beteiligten Communities in TP 3.3 und TP 3.2 Pilotbetrieb umgesetzt.

TP 2.4 Datenmanagement (Leitung: TU Dresden)

Dieses Arbeitspaket setzt sich mit technischen, konzeptionellen und inhaltlichen Fragen zum Management von Forschungsdaten als Teil der in TP 2.1 entworfenen Gesamtarchitektur und unter Berücksichtigung der Fallstudien aus AP1 auseinander. Wichtige Aspekte sind die bereits in TP 2.1 aufgeführten, also u.a. die Verwaltung der verteilten großen Datenmengen, Datenreplikation und deren Verwaltung, Metadatenmanagement, Suchen und Finden von Daten, Datenzugriff, Anbindung an Analyse-Ressourcen und vorausplanend Aspekte zur Archivierung.

Wesentlich für die Umsetzung des Datenmanagements werden auch die Ergebnisse der *RDA*-Arbeitsgruppe "Practical Policy" sein, die aus einer Best-Practice-Sammlung Definitionen von grundlegenden Datenoperationen für Datenmanagementsysteme und -repositorien erstellt hat. Weitere Fragestellungen sind z.B. Datenformate, die eine möglichst hohe Performance beim Zugriff haben, effiziente Speicherplatznutzung und Skalierbarkeit.

Alle Aspekte zu Metadaten sowie deren Management werden dabei in TP 2.2. und TP 2.3 erarbeitet, aber mit dem Datenmanagement integriert. Deshalb wird es zwischen den Unterarbeitspaketen eine enge Zusammenarbeit geben.

In einem ersten Schritt wird evaluiert auf welcher Ebene der Speicherhierarchie und über welche Schnittstellen eine Föderation zum Tragen kommen soll. Dazu werden die bei den Partnern eingesetzten und andere weit verbreitete Speichersysteme betrachtet. Die Partner können dabei auf ihre umfangreichen Erfahrungen als Betreiber von verteilten Dateisystemen, Grid-basierten Systemen (wie *iRODS* oder *dCache*), Objektspeichern u.a. sowie ihre umfassenden Forschungstätigkeiten in Projekten wie *LSDMA, D-Grid, MASi* oder *SIOX* zurückgreifen.

Weiterhin wird – basierend auf den Anforderungen der Architektur – eine Middleware zur Föderation der FDS ausgewählt und wenn nötig erweitert. Dazu wird eine detaillierte Bestandsaufnahme existierender Middleware-Systeme (Open Source) aus Entwicklungen deutscher, europäischer und internationaler Middleware-Initiativen und -Projekte vorgenommen. Hierzu zählen etwa *EUDAT*, *INDOGO-DataCloud*, Ergebnisse der *Research Data Alliance*, *EMI*, *LSDMA*, *MASi* und *RADAR*. Die verschiedenen Systeme werden für die Eignung im Projekt evaluiert, wobei die Abdeckung der oben genannten technischen Anforderungen sowie der Kernprozesse im Vordergrund steht.

Weitere zu berücksichtigende und umzusetzende Aspekte sind unter anderem:

- Authentifizierungs- und Autorisierungsmechanismen für die klare und einfach nutzbare Umsetzung der diesbezüglichen Anwenderanforderungen.
- Für die eindeutige und persistente Referenzierung von Forschungsdaten wird auf persistente Identifikatoren zurückgegriffen. Dazu können z.B. DOIs von *DataCite* für Langzeitdaten oder *EPIC*-Handles für Zwischendaten genutzt werden.
- Das Suchen und Finden von Datensätzen anhand der Metadaten wird über die Integration der Entwicklungen in TP 2.3 Metadatenmanagement realisiert.
- Die Verfügbarmachung der Daten auf HPC- oder Cloud-Ressourcen kann z.B. über die Middleware UNICORE oder über Cloud-Standards (wie CDMI und S3) bewerkstelligt werden. Die notwendigen Methoden und Werkzeuge sind den Nutzern dafür zur Verfügung zu stellen.
- Datensicherung soll durch teilweise Redundanz zwischen den FDS gewährleistet werden. Dabei sollen Gruppen von FDS gefunden werden, die ihre Daten aufgrund von Ähnlichkeiten redundant halten.

- Die Infrastruktur wird sich möglichst nahtlos in Nutzerumgebungen integrieren lassen, was über Standardschnittstellen und -tools realisiert wird. Dies kann dann z.B. auch durch Science Gateways genutzt werden.
- Die *GeRDI*-Infrastruktur soll in das nationale und internationale Umfeld integriert werden. Entsprechende Schnittstellen oder Dienste sollen z.B. das Zusammenspiel mit anderen Speicherdiensten wie den *EUDAT*-Diensten oder Computing-Umgebungen wie *PRACE* gewährleisten. Hier spielen insbesondere offene Standard-Schnittstellen eine zentrale Rolle.

Alle Aspekte des Datenmanagements werden in einem iterativen Prozess auf Basis der Software- und Systemarchitektur (TP 2.1) in enger Zusammenarbeit mit AP3 Pilotbetrieb sowie TP 2.3 Metadatenmanagement umgesetzt.

Meilensteine von AP2:

Meilenstein	Bezeichnung	Zeitpunkt	Verantwortlich
2.a	Gesamtarchitektur – erster Entwurf	M12	CAU
2.b	Erste Version Metadaten – allgemeine und für Beispiel-Communities	M12	ZBW
2.c	Erste Version der Software für das Daten- und Metadatenmanagement	M12	TUD
2.d	Gesamtsystem- und Softwarearchitekturentwurf	M24	CAU
2.e	Zweite Version der Software für das Daten- und Metadatenmanagement	M24	TUD
2.f	Finale Architekturdokumentation mit Berücksichtigung der IT-Sicherheit, Metadatenmanagement sowie Datenmanagement und Aufbau und Integration der Pilotzentren	M36	CAU
2.g	Finale Version und Dokumentation des Daten- und Metadatenmanagement	M36	TUD

AP3: Pilotbetrieb (Leitung: LRZ München)

In diesem Arbeitspaket wird *GeRDI* in einen Pilotbetrieb übernommen. Dazu wird in den drei Pilotzentren die erforderliche Infrastruktur bereitgestellt. Anschließend werden mehrere Benutzer-Communities mit ihren Daten in das System aufgenommen. In Zusammenarbeit mit den Wissenschaftlern wird eine Evaluation vorgenommen und ein Servicemodell für die Datenkuration als zukünftige Serviceleistung eines Datenzentrums entworfen. Ebenfalls zum Pilotbetrieb gehört die Entwicklung und Umsetzung eines gesamtheitlichen IT Sicherheitskonzeptes.

TP 3.1 IT-Sicherheit (Leitung: LRZ München)

Das Arbeitspaket begleitet alle Phasen des Projekts von der Anforderungsanalyse bis zum Pilotbetrieb und entwickelt das IT-Sicherheitskonzept für das virtuelle FDS. Das Sicherheitskonzept wird in den Pilotzentren umgesetzt und anhand der Pilot-Erfahrungen verbessert.

Ergebnis des Arbeitspakets ist ein IT-Sicherheitskonzept, das aus mehreren Dokumenten besteht, die als Teilaufgaben entwickelt werden. Der verwendete Ansatz orientiert sich an der für den IT-Grundschutz vom Bundesamt für Sicherheit in der Informationstechnik (BSI) empfohlenen Vorgehensweise sowie der ISO 27000.

Die "Struktur- und Schutzbedarfsanalyse" geht von den Anforderungen und einer ersten Informationsverbundkonzeption des virtuellen FDS aus und definiert den notwendigen Schutzbedarf der einzelnen Komponenten im Hinblick auf die Schutzziele Integrität, Verfügbarkeit,

Vertraulichkeit, Authentizität und Nicht-Abstreitbarkeit. Ebenso wird der jeweilige Schutzbedarf der gespeicherten Forschungsdaten berücksichtigt.

Im Rahmen der "Maßnahmenplanung" wird eine Liste von Maßnahmen entwickelt, die geeignet sind, diese Schutzziele gemäß der Schutzbedarfsanalyse zu erreichen. Teilweise werden hier die BSI-Grundschutz-Kataloge herangezogen und voraussichtlich wird mindestens in den Bereichen Integrität und Verfügbarkeit auch eine erweiterte Sicherheitsanalyse notwendig werden.

Die Implementierung der Maßnahmen erfolgt als Teilaufgabe "Umsetzung und Evaluation". Einige der Sicherheitsmaßnahmen fließen in das Pflichtenheft des FDS ein und werden dort berücksichtigt. Andere Maßnahmen wiederum werden technisch und organisatorisch im Pilotbetrieb implementiert.

TP 3.2 Aufbau, Integration der Pilotzentren (Leitung: TU Dresden)

Auf Basis der vorangegangenen Ergebnisse werden drei Pilotzentren an der CAU, der TU Dresden und am LRZ eingerichtet. Im Einzelnen gehören zu diesem Arbeitspaket die Bereitstellung der Infrastruktur, Installation und Konfiguration von ausgewählten, existierenden Softwarepaketen und ggfs. neu entwickelter Middleware. Die Software-Komponenten werden auf ihre Funktionsweise im Zusammenspiel miteinander getestet. Der Maßnahmenkatalog des IT-Sicherheitskonzeptes wird umgesetzt.

Die Arbeiten erfolgen als „rolling integration“, d.h. Verbesserungen, neue Erweiterungen und Entwicklungen werden im Laufe des Projektes direkt einbezogen. Hierzu ist eine enge Zusammenarbeit mit TP 1.3 zum Softwaremanagement geplant.

Die Pilotzentren werden anschließend mit einem anerkannten Datensiegel zertifiziert.

TP 3.3 Community Management, Datenkuration (Leitung: ZBW)

Ziel des Arbeitspakets ist die kontinuierliche Einbindung der Communities in die Entwicklungsarbeiten (z.B. inhaltliche Arbeiten zu Metadaten in TP 2.2, Evaluierung in TP 3.4) sowie die Erarbeitung eines Servicemodells für die Datenkuration. Arbeitspaket 3.3 unterstützt die Forschergruppen an den jeweiligen Pilotzentren bei der Anbindung bestehender Repositorien, der Datenkuration und einer disziplinübergreifenden Datennutzung mit Hilfe von *GeRDI*. Für das Community Management werden besondere Ressourcen von je einer halben Stelle für die von CAU, TU Dresden, LRZ und ZBW betreuten Community vorgesehen, sodass eine explizite Schnittstelle zwischen den Projektpartnern und den Fachcommunities geschaffen ist. Diese Projektmitarbeiter kümmern sich in enger Vor-Ort-Zusammenarbeit mit Wissenschaftlern um deren Anwendungsfälle und erfassen begleitend die Art der benötigten Arbeiten, die Aufwände und auftretende Probleme. Dabei wird insbesondere die disziplinübergreifende Nutzung von Forschungsdaten berücksichtigt.

Das Servicemodell beinhaltet eine definierte Vorgehensweise, Methoden zur Aufwandsabschätzung und Best Practices für die Aufnahme von Daten in *GeRDI* bzw. Forschungsdatenmanagementsysteme allgemein. Auf Basis des erarbeiteten Servicemodells können Datenzentren zukünftig die Erfassung von Forschungsdaten in *GeRDI* ("Datenkuration") als standardisierte, prozessorientierte Dienstleistung für Wissenschaftler anbieten und damit Schwellen zur Nutzung von FDS abbauen.

Zur Verdeutlichung der Vorgehensweise könnte ein konkretes Beispiel aus dem Bereich der Umweltwissenschaften wie folgt aussehen:

- Forschergruppe A (Umweltphysik) führt in den Alpen Messungen zur Deposition von Radionukleiden im Schnee durch und erfasst dabei primär Radioaktivität und nebenbei Wassermengen an mehreren Messstellen

- Forschergruppe B (Permafrost-Forschung) erforscht Wasserabflüsse aus Gestein und interessiert sich für die von A erfassten Wassermengen

Im Rahmen von *GeRDI* würde zunächst der Anwendungsfall erfasst werden (TP 1.1). Forschergruppe A wird dann bei der Datenkuration (Aufnahme der Messdaten in *GeRDI*) und Forschergruppe B bei der Datennutzung (Abruf der Daten aus *GeRDI*) unterstützt.

An jedem der Pilotzentren werden Forschergruppen betreut, die sich aus bestehenden und geplanten Forschungskooperationen rekrutieren und zu denen die Pilotzentren bereits eine erste Arbeitsbeziehung pflegen. Da bereits das Arbeitspaket ein "Servicemodell" simulieren soll, werden die genauen Forschungsgruppen erst während der Laufzeit von *GeRDI* ausgewählt und aufgenommen, damit auch diese Aufwände korrekt erfasst werden. Derzeit sind folgende Kandidaten im Gespräch:

CAU:

- Meereswissenschaften
  - Evaluation im Kontext des Exzellenzclusters „Ozean der Zukunft“, Prof. Dr. Martin Visbeck, Universität Kiel.
- Fischerei
  - Mit der Arbeitsgruppe für Umwelt-, Ressourcen- und Ökologische Ökonomik im Bereich der Fischerei-Simulationen wird eine Verknüpfung von meereswissenschaftlichen Daten mit wirtschaftswissenschaftlichen Daten erprobt, Prof. Dr. Martin Quaas, Universität Kiel.
- *Pangaea - Data Publisher for Earth & Environmental* Science
  Für die Anbindung eines existierenden disziplinären Dienstes wird *Pangaea* angebunden, Dr. Michael Diepenbroek.

LRZ (Schwerpunkt Umweltwissenschaften):

- *ClimEx* (Hydrometeorologie - Hydrologie)
- *WasserZukunftBayern* (Hydrologie - Landnutzung)
- *VAO Alpine Environmental Data Analysis Centre* (disziplinübergreifendes Projekt)
- Rainfall-induced Earthquakes (Meteorologie - Seismologie)
- Auswirkungen des Klimawandels auf die Fauna/Biodiversität (Klima/Meteorologie - Zoologie)
- Gemeinsame Messkampagnen von Umweltphysik und Permafrost-Forschung
- Einfluss des Gletscherschwunds auf die Alpine Hydrologie (Glaziologie - Geographie)

TUD:

- Lebenswissenschaften
  - Prof. Dr. Ingo Röder (Institut für Medizinische Informatik und Biometrie, TU Dresden) – medizinische Bilddatenanalyse, medizinische Systembiologie
  - Prof. Dr. Gene Myers (MPI für Zellbiologie und Genetik) – Entwicklung neuartiger Mikroskope zur Zellanalyse inklusive automatische Bildanalyse
- Digital Humanities
  - Prof. Dr. Gregory Crane (Alexander von Humboldt Professor of Digital Humanities, Universität Leipzig) – Philologie der griechisch-römischen Kultur anhand automatischer Textanalyse und –annotation

ZBW:

- *SOEP* (Sozio-ökonomisches Panel – Wirtschaftswissenschaften)
  - Prof. Dr. Jürgen Schupp – Institut für Soziologie an der FU Berlin und Leiter SOEP.

Ergebnis des Arbeitspakets sind Empfehlungen, ob eine Datenkuration von Forschungsdaten als Servicemodell für Wissenschaftler realistisch umsetzbar ist und wie diese in der Praxis aussehen

könnte. Die Empfehlungen wären auch ein Bestandteilteil des Roll-out-Modells (TP 4.3) für weitere Datenzentren in einer zweiten Phase von *GeRDI*.

TP 3.4 Evaluierung, Proof-of-Concept (Leitung: TU Dresden)

In diesem Arbeitspaket wird die für *GeRDI* integrierte bzw. neu implementierte Software zusammen mit der eingerichteten Infrastruktur im Sinne eines Proof-of-Concept entlang der in TP 1.1 festgelegten Fallstudien sowie anhand der in AP1 festgelegten Anforderungen formal evaluiert. Die Evaluierung bindet die teilnehmenden Wissenschaftler aus den Disziplinen der Fallstudien ein und behandelt die Benutzbarkeit und Benutzungsfreundlichkeit des Systems.

Es findet zudem eine technische Evaluierung statt, die insbesondere die Skalierbarkeit und die technische Performance (Skalierbarkeit, optimierte Speichernutzung, Effizienz der Kommunikation mittels Middleware, Netz-Performance etc.) und Zuverlässigkeit des FDS im Hinblick auf die Umsetzung eines Datenmanagementplans evaluiert. Als Evaluierungsverfahren werden sowohl eine summative Evaluierung (Abgleich zwischen postuliertem und erreichtem Zielzustand) als auch eine formative Evaluierung zum Einsatz kommen (auf Basis von Zwischenevaluationen erfolgen Korrekturen laufender Maßnahmen, um die Wahrscheinlichkeit der Zielerreichung zu erhöhen).

Meilensteine von AP3

Meilenstein	Bezeichnung	Zeitpunkt	Verantwortlich
3.a	Initiales Sicherheitskonzept	M12	LRZ
3.b	Erster Soll/Ist Vergleich der IT-Sicherheitsmaßnahmen	M24	LRZ
3.c	Erste Version von *GeRDI* ist in Pilotzentren eingerichtet	M24	TUD
3.d	Umsetzungsbericht zum IT-Sicherheitskonzept	M36	LRZ
3.e	Alle Pilotzentren sind mit einem Datensiegel zertifiziert	M36	TUD
3.f	Mindestens fünf Communities sind an *GeRDI* angebunden	M36	TUD

AP4: Nachhaltigkeit (Leitung: ZBW)

Dieses Arbeitspaket bereitet die Nachhaltigkeit der erzielten Projektergebnisse vor. Um dieses Ziel zu erreichen, wird in TP 4.1 ein Konzept erarbeitet, das spätere Nutzer befähigt, *GeRDI* für die eigenen Zwecke einzusetzen. TP 4.2 entwickelt ein Konzept für ein tragfähiges Betriebsmodell für den dauerhaften Erhalt der entwickelten Infrastruktur. Schließlich bereitet TP 4.3 das Roll-Out gemeinsam mit den Forschungsfördern DFG und ggf. auch BMBF vor.

TP 4.1 Trainingskonzept (Leitung: DFN-Verein)

In diesem TP wird ein flexibles Trainingskonzept entwickelt, das es Hochschulen und Forschungseinrichtungen bzw. entsprechenden Konsortien ermöglicht, ein FDS aus dem Projekt lokal in ihrem Rechenzentrum einzurichten, die Möglichkeiten zur Datenkuration (TP 3.3) zu nutzen bzw. das bei der Andockung vorhandener FDS an *GeRDI* Unterstützung bietet. Das Trainingskonzept stützt sich auf einen kontinuierlichen Prozess, in dem die Erstellung der erforderlichen Materialien mit einer gleichzeitigen Erprobung bei Anwendern, Betreibern und Entwicklern einhergeht.

Die Materialien werden von den Projektpartnern erstellt und umfassen u.a. allgemeine Informationen, Anleitungen, HowTo‘s oder Best Practices. Durch die frühe Einbindung ausgewählter Anwender in den Pilotzentren erfolgt eine Qualitätskontrolle und damit eine

sukzessive Verbesserung der Materialen bereits in der ersten Phase des Projektes. Hierzu erfolgt eine enge Abstimmung insbesondere mit den Arbeitspaketen TP 3.3. und TP 3.4.

Als Ergebnis steht zum Projektende ein weitgehend offenes Trainingskonzept für das virtuelle FDS bereit, in das zukünftige Anpassungen und neue Inhalte einfach aufgenommen werden können. Darauf aufbauen kann in einer zweiten Projektphase das Training auf neue Anwendergruppen und Betriebsumgebungen erweitert werden, um die allgemeine Nutzbarkeit der Ergebnisse kontinuierlich zu verbessern.

TP 4.2 Zukünftiges Betriebsmodell (Leitung: DFN-Verein)

In diesem TP werden mögliche Betriebsmodelle für den nachhaltigen Betrieb nach Auslaufen der Förderung definiert, untersucht und bewertet. Ein nachhaltiger Betrieb umfasst damit gleichermaßen die Finanzierung des dauerhaften Betriebs, der dauerhaften Pflege der FDS-Software sowie der übergreifenden *GeRDI*-Infrastruktur. Zur Festlegung der hierfür notwendigen Rahmenbedingungen sind Erkenntnisse insbesondere aus TP 3.1, TP 3.2 und TP 1.3 relevant. Das Betriebsmodell adressiert auch die ggf. unterschiedlichen Finanzierungsmodelle, die für den Betrieb eines FDS erforderlich werden.

Das Betriebsmodell sieht folgende Varianten vor, die im Rahmen des TP ausgearbeitet werden:

a) Einrichtung eines FDS als physikalische Umgebung in einem Rechenzentrum einer Hochschule bzw. Forschungseinrichtung für die Nutzung innerhalb der Einrichtung. In diesem Fall müsste die Finanzierung der Hardware im Sinne von Investitionsmitteln für ein FDS ermöglicht werden.
b) Einrichtung eines FDS für eine Community, die gemeinsam ein eingerichtetes FDS an einem Standort nutzen. In diesem Falle sind für den Standort, an dem die Community das FDS einrichtet und betreibt, Investitionsmittel erforderlich. Demgegenüber müsste für weitere nutzende Einrichtungen dieses FDS die Finanzierung zur „Mit-Nutzung“ des Community-FDS im Sinne eines Dienstes ermöglicht werden.
c) Betrieb über den DFN-Verein (DFN-Cloud) im Sinne eines Anbieter-Nachfrager-Modells, auf Basis eines über den DFN-Verein abzuwickelndes Entgeltsystems.
d) Es sind föderierte Betriebsmodelle denkbar. So ist z.B. als vertragliches Rahmenwerk für eine föderierte Infrastruktur einschließlich Regelungen zur Kostenumlage die DFN-Cloud auf ihre Anwendbarkeit zu prüfen.

TP 4.3 Vorbereitung Roll-Out (Leitung: ZBW)

Gesamtziel des Projekts ist, ein virtuelles FDS – auf Basis vorhandener Systeme, die gegebenenfalls erweitert werden – zu entwickeln, das auf zahlreichen physischen Repositorien aufbaut und zudem Modellcharakter für eine nationale Infrastruktur für Forschungsdaten hat. Mit diesem TP wird das Ausrollen der Projektergebnisse für die zweite Projektphase vorbereitet.

Dieses TP erarbeitet Empfehlungen an die Forschungsförderer (DFG und BMBF), wie ein Förderprogramm sowie ein flächendeckender Roll-out in der Phase II umgesetzt und wie dieses vorbereitet werden kann. Die Empfehlungen enthalten zumindest folgende Angaben:

a) Entwicklung von Maßnahmen zur Kommunikation über *GeRDI* und den damit verbundenen Möglichkeiten (z.B. Workshops zu Interessensbekundungen zur Teilnahme an *GeRDI*)
b) Abstimmung und Berücksichtigung anderer Initiativen, wie etwa die Umsetzung einer *European Open Science Cloud* im Rahmen der derzeit vorgesehenen „preparatory phase“ bzw. der *Helmholtz Data Initiative*.
c) Mittelbedarf anhand einer Abschätzung der Anzahl an FDS, die in der Phase 2 eingerichtet werden könnten
d) Festlegen der Anforderungen und Kriterien, die eine Einrichtungen erfüllen muss, um ein FDS zu erhalten; Hierzu zählt bspw. die Verwendung der im Projekt entwickelten Softwarekomponenten, die Nutzung der Metadatenstandards etc.
e) Entwurf eines Ausschreibungstextes/einer Förderlinie.

Meilensteine von AP4:

Meilenstein	Bezeichnung	Zeitpunkt	Verantwortlich
4.a	Erste Version Trainingskonzept	M12	DFN-Verein
4.b	Erprobtes Trainingskonzept mit 1-2 Fachcommunities	M24	DFN-Verein
4.c	Finales Trainingskonzept	M36	DFN-Verein
4.d	Kriterienkatalog für Betriebsmodell	M12	DFN-Verein
4.e	Bewertung verschiedener Betriebsmodelle	M24	DFN-Verein
4.f	Finales Betriebsmodell	M36	DFN-Verein
4.g	Erstes Konzept für Roll-out	M24	ZBW
4.h	Finales Konzept für Roll-out	M36	ZBW

AP5: Governance, Projektmanagement und Dissemination (Leitung: K. Tochtermann, H.-J. Bungartz)

Dieses Arbeitspaket stellt die Governancestruktur für das Projekt und insbesondere die hierfür vorgesehenen Organe *Lenkungsausschuss*, *Nutzerausschuss* sowie *Fachbeirat* (s.u.) vor. Die nachfolgende Abbildung veranschaulicht deren Beziehungen untereinander.

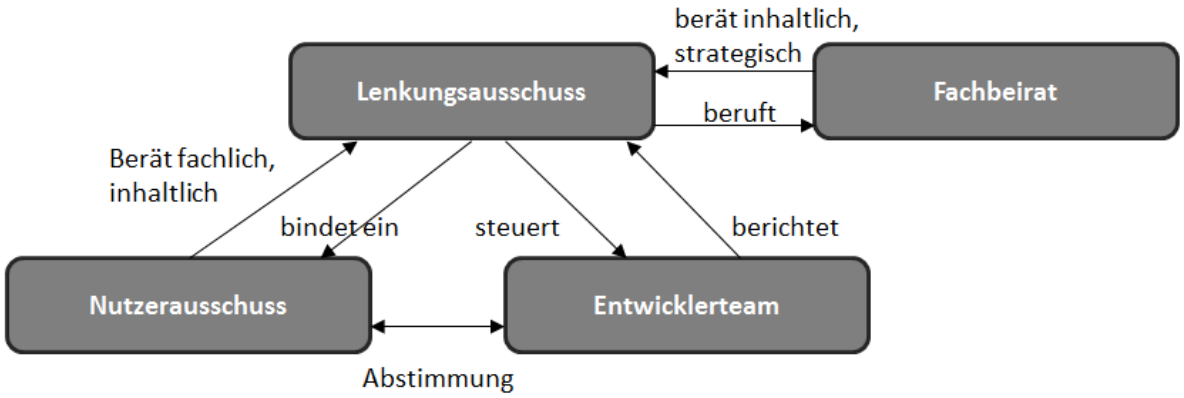

Abb. 4: Governance-Struktur

Dabei werden insbesondere die Aspekte beschrieben, die über die Grundmechanismen für ein erfolgreiches Projektmanagement hinausgehen. Zu den Grundmechanismen gehören etwa die Verantwortlichkeit der AP- bzw. TP-Leitungen, ihre Arbeitspakete technisch, inhaltlich und wissenschaftlich so zu koordinieren, dass Zeit- und Budget-Rahmen eingehalten werden. Zudem tragen die AP- bzw. TP-Leitungen dafür Sorge, dass sie sich regelmäßig untereinander und AP- bzw. TP-übergreifend abstimmen und dass die jeweiligen Nutzergruppen im gebotenen Umfang eingebunden werden.

Die in Abbildung 3 vorgestellte Struktur der Arbeitspakete wurde gestaltet, um eine Governance-Struktur für die bestmögliche Erreichung der folgenden Ziele sicherzustellen:

- Einbindung aller Projektpartner in die Entscheidungsprozesse
- Einbindung der verschiedenen Fachcommunities aus den Pilotzentren in die Entwicklungen
- Klare und effiziente Entscheidungs- und Kommunikationsstrukturen
- Mechanismen, um Konflikten vorzubeugen bzw. sie entlang der Struktur zu lösen
- Management, das hohe Qualität sowie Budget- und Termintreue sicherstellt

Die Governance-Struktur unterscheidet zwischen den folgenden Organen:

Lenkungsausschuss

Der Lenkungsausschuss ist das höchste Entscheidungsgremium des Projekts. In seinen Zuständigkeitsbereich fallen strategische Entscheidungen, Anpassungen des Projektplans, Budgetänderungen und Konfliktlösungen. Die Mitglieder des Lenkungsausschusses müssen das Mandat haben, für ihre Organisation bindende Entscheidungen für das Projekt zu treffen.

Die Mitglieder des Lenkungsausschusses treffen sich quartalsweise. Zudem kann bei Bedarf eine Sitzung auch per Video-Konferenz einberufen werden.

Die Mitglieder des Lenkungsausschuss sind:
- A. Bode für das LRZ München
- H.-J. Bungartz für den DFN-Verein Berlin
- W. Hasselbring für die Universität Kiel
- W. Nagel für die TU Dresden
- K. Tochtermann für die ZBW (Vorsitz)

Zudem nimmt ein Vertreter des Nutzerausschusses (s.u.) als Gast an den Sitzungen des Lenkungsausschusses teil, um auch dort die Interessen und Bedürfnisse der Nutzer zu vertreten.

Nutzerausschuss

Analog zu den von HPC-Zentren bekannten Nutzerausschüssen setzt sich der Nutzerausschuss aus Vertretern der Datenrepositorien und Fachcommunities aus den Pilotzentren zusammen und trifft sich zu festen Terminen zweimal pro Jahr mit den Projektpartnern für den fachlich-inhaltlichen Austausch mit diesen sowie zwischen den Fachcommunities untereinander (siehe Beschreibung im Antrag). Es ist geplant, nicht mehr als 10 Personen für die Mitwirkung im Nutzerausschuss einzuladen.

Fachbeirat

Der Fachbeirat begleitet das Projekt beratend und bringt insbesondere zusätzliche Kompetenzen ein, die nicht im erforderlichen Umfang in das Projekt integriert sein können, für die erfolgreiche Bearbeitung aber relevant sind. Hierzu gehören etwa Kompetenzen aus den Bereichen Recht und Sicherheit, aber auch Kompetenzen aus anderen ähnlich gearteten Projekten, wie etwa EU-Projekten oder der Dateninitiative der HGF. Der Fachbeirat trifft sich maximal zwei Mal pro Jahr und benennt den oder die Vorsitzende/n aus seinen eigenen Reihen. Er wird in seinen Sitzungen über den Fortschritt des Projekts informiert und um Stellungnahme zu ausgewählten Punkten gebeten. Es ist geplant, nicht mehr als 10 Personen für die Mitwirkung in den Fachbeirat einzuladen.

Zudem umfasst das AP5 das interne Projektmanagement, wie Koordination der Arbeitspakete, Qualitätssicherung, Überwachung des inhaltlichen Fortschritts entsprechend des Zeitplans, Einleiten von Änderungsmaßnahmen etc.

Schließlich kümmert sich dieses AP um die Dissemination des Projekts und seiner Ergebnisse hinsichtlich ähnlicher Aktivitäten in Deutschland und in Europa. Als Zielgruppen werden die Wissenschaftspolitik, die Öffentlichkeit und die wissenschaftlichen Forschungs-Communities in Betracht gezogen. Um maximale Transparenz über die Projektaktivitäten sowie Konformität mit europäischen Forschungsinfrastrukturen sicherzustellen, orientiert sich das Projekt an der Europäischen *Charta für den Zugriff auf Forschungsinfrastrukturen*, die im Juni 2015 in ihrer ersten Version veröffentlicht wurde.[41]
Von besonderer Bedeutung und zur Vorbereitung des Roll-outs (TP 4.3) wird in Abstimmung mit der DFG bereits während der Phase I ein Aufruf für Interessensbekundungen gestartet.

[41] http://ec.europa.eu/research/infrastructures/pdf/2015_charterforaccessto-ris.pdf

Anschließend findet ein Workshop statt, in dem die Ideen und Technologien des Projekts vorgestellt werden, aber auch die zu diesem Zeitpunkt vorliegenden Anforderungen, die interessierte Einrichtungen erfüllen müssen, um in Phase II Fördermittel einzuwerben.

Meilensteine von AP5:

Meilenstein	Bezeichnung	Zeitpunkt	Verantwortlich
5.a	Tätigkeitsbericht für Projektjahr 1	M12	ZBW
5.b	Tätigkeitsbericht für Projektjahr 2	M24	ZBW
5.c	Tätigkeitsbericht für Projektjahr 3	M36	ZBW

Ablaufplan

Die Arbeitspakete werden in folgendem zeitlichen Ablauf bearbeitet. Zur besseren Übersichtlichkeit wird der Zeitplan quartalsweise dargestellt.

Jahr/Quartal	1/1	1/2	1/3	1/4	2/1	2/2	2/3	2/4	3/1	3/2	3/3	3/4
TP 1.1												
TP 1.2												
TP 1.3												
TP 2.1												
TP 2.2												
TP 2.3												
TP 2.4												
TP 3.1												
TP 3.2												
TP 3.3												
TP 3.4												
TP 4.1												
TP 4.2												
TP 4.3												
AP5												

Abb. 5: Ablaufplan

2.4 Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

Die im Projekt erzielten Ergebnisse werden kontinuierlich auf einem Projektblog veröffentlicht. Zudem sind Vorträge auf einschlägigen Veranstaltungen sowie Veröffentlichungen in referierten Fachzeitschriften bzw. Tagungsbänden geplant. Diese Publikationen werden open-access, im Falle von referierten Zeitschriften ggf. nach einer Embargozeit, verfügbar sein. Sämtlicher Quellcode, der im Zusammenhang des Projekts entsteht, wird einschließlich seiner Dokumentation als Open Source auf den einschlägigen Plattformen (z.B. *SourceForge*, *GitHub*) bereitgestellt. Die softwaretechnischen Neuerungen werden während des Projekts so dokumentiert bzw. kommuniziert, dass Bugfixing und Änderungen prinzipiell auch durch nicht an dem Projekt beteiligtes Personal durchgeführt werden können. Der nachhaltige Betrieb der entwickelten Lösungen wird in entsprechenden AP vorbereitet und in einer zweiten Projektphase sichergestellt.

4 Finanzierung des Vorhabens

[...]

5 Voraussetzungen für die Durchführung des Vorhabens

5.1 Angaben zur Dienststellung

Prof. Dr. Klaus Tochtermann, Direktor ZBW – Leibniz-Informationszentrum Wirtschaft und Universitätsprofessor am Institut für Informatik an der Universität Kiel.
Prof. Dr. Wilhelm Hasselbring, Universitätsprofessor am Institut für Informatik an der Universität Kiel, Mitglied (Principal Investigator) im Exzellenzcluster Future Ocean.
Prof. Dr. Hans-Joachim Bungartz, Universitätsprofessor in der Fakultät für Informatik an der TU München, Vorsitzender des Vorstands des DFN-Vereins, Mitglied im Direktorium des Leibniz-Rechenzentrums München.
Prof. Dr. Arndt Bode, Universitätsprofessor in der Fakultät für Informatik an der TU München, Vorsitzender des Direktoriums des Leibniz-Rechenzentrums.
Prof. Dr. Wolfgang Nagel, Universitätsprofessor im Institut für Technische Informatik an der TU Dresden, Direktor des Zentrums für Informationsdienste und Hochleistungsrechnen der TU Dresden.
Dr. Christian Grimm, Geschäftsführer DFN-Verein Berlin
Jochem Pattloch, Geschäftsführer DFN-Verein Berlin

5.2 Zusammensetzung der Projektarbeitsgruppe

Dr. Timo Borst (ZBW), Leiter der Abteilung Innovative Informationssysteme und Publikationstechnologien, wird die fachlich-inhaltlichen Arbeiten der ZBW, speziell TP 2.2 koordinieren, und sich in das Projektmanagement (AP 5) einbringen.
Dr. Doreen Siegfried (ZBW), Leiterin Marketing und Öffentlichkeitsarbeit, wird mit ihrer Abteilung die Öffentlichkeitsarbeit für das Projekt unterstützen (AP 5).
Dipl.-Kfm. Olaf Siegert (ZBW), Leiter der Abteilung Publikationsdienste, wird mit seiner Abteilung die Zusammenarbeit mit der Fachcommunity des *SOEP* und mit dem *SOEP* (speziell Anforderungsdefinition, TP 2.2) sowie die Entwicklung der Fallstudien unterstützen (TP 1.1).
Dr. Willi Scholz (ZBW), Wissenschaftspolitischer Referent, wird an den Arbeiten zur Vorbereitung des Roll-Out (TP 4.3) mitwirken.
Dr. Carsten Schirnick (GEOMAR), Leiter Datenmanagementteam am GEOMAR Kiel, soll in TP 1.1 und TP 3.3 für die Fallstudie in den Meereswissenschaften die Schnittstelle zum Kieler Datenmanagementteam darstellen.
Matthias Westphal, Systemadministrator (Landesstelle) in der Arbeitsgruppe Software Engineering der CAU, soll in TP 3.2 die Installation und Konfiguration der Hard- und Software für den Pilotbetrieb unterstützen.
Arnd Plumhoff, Softwareingenieur (Landesstelle) in der Arbeitsgruppe Software Engineering der CAU, soll in TP 1.3 das Softwaremanagement unterstützen.
Dipl.-Kfm. Torsten Sandersfeld (DFN-Verein), Leiter Verwaltung, wird in TP 4.2 zur Bearbeitung betriebswirtschaftlicher Fragestellungen beitragen.

- RA Christine Legner-Koch (DFN-Verein), Justiziarin, wird in TP 4.2 zur Bearbeitung juristischer Fragestellungen beitragen.
- M.A. Tanja Hanauer (LRZ), IT-Sicherheitsverantwortliche / CISSP in der Gruppe IT Infrastruktur und Dienste, begleitet die Erstellung und Umsetzung des IT-Sicherheitskonzeptes (TP 3.1).
- Dr. Ludwig Maier (LRZ), Betriebsspezialist in der Gruppe IT Infrastruktur und Dienste, unterstützt die Inbetriebnahme des Pilotzentrums am LRZ (TP 3.2).
- Dr. Jens Weismüller (LRZ), Projektkoordinator für Forschungskooperationen, hilft die Datenkuration mit den Communities aufzubauen (TP 3.3).
- Dr. Ralph Müller-Pfefferkorn (ZIH/TUD), Leiter der Abteilung „Verteiltes und Datenintensives Rechnen“ wird beim Entwurf der Architektur (TP 2.1) sowie bei Entwurf und Umsetzung des Metadaten- (TP 2.3) und Datenmanagements (TP 2.4) im Umfang von 6 PM mitarbeiten.
- Dr. Michael Kluge (ZIH/TUD) – wissenschaftlicher Mitarbeiter wird sich beim Datenmanagement (TP 2.4), beim Aufbau der Piloten (TP3.2) und der Evaluierung (TP 3.4) mit 6 PM beteiligen.

5.3 Zusammenarbeit mit anderen Institutionen und anderen Wissenschaftlerinnen und Wissenschaftlern

5.3.1 Institutionen oder Wissenschaftlerinnen und Wissenschaftler, mit denen für dieses Vorhaben eine konkrete Vereinbarung besteht

entfällt

5.3.2 Institutionen, Wissenschaftlerinnen und Wissenschaftler, mit denen in den letzten drei Jahren gemeinsame Projekte durchgeführt wurden

- Prof. Dr. York-Sure Vetter, ehemals GESIS, jetzt KIT
- Prof. Dr. Michael Granitzer, Uni Passau
- Prof. Dr. Markus Strohmaier, GESIS, Uni Koblenz-Landau
- Prof. Dr. Thomas Köhler, TU Dresden
- Prof. Dr. Marc Rittberger, DIPF Frankfurt
- Prof. Dr. Gert Wagner, DIW Berlin
- Prof. Dr. Stefanie Lindstaedt, TU Graz
- Prof. Dr. Klaus Pohl, Uni Duisburg-Essen
- Prof. Dr. Ralf Reussner, KIT
- Prof. Dr. Stephan Paul, TU München
- Dr. Klaus Ceynowa, Generaldirektor Bayerische Staatsbibliothek
- Prof. Dr. Karl-Heinz Hoffmann, Präsident der Bayerischen Akademie der Wissenschaften

5.4 Erklärungen zur Erfüllung der Förderbedingungen

Es wird zugesichert, alle Ergebnisse, Publikationen und sonstigen Resultate – soweit möglich – öffentlich, ohne Kosten und Einschränkungen und barrierefrei für jedermann zur Verfügung zu stellen. Der Quellcode der weiterentwickelten Software wird Open Source, einschließlich der dazugehörigen Dokumentation, über den Projektwebseite sowie weitere Stellen für die Allgemeinheit zugänglich gemacht.

5.5 Projektrelevante Zusammenarbeit mit erwerbswirtschaftlichen Unternehmen

entfällt

5.6 Projektrelevante Beteiligungen an erwerbswirtschaftlichen Unternehmen

entfällt

6 Ergänzende Erklärungen

entfällt