

Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für datenintensive Dienste – bwDATA Phase I (2013-2014)

Vorgelegt von den Leitern der Rechen- bzw. Informationszentren der Universitäten des Landes Baden-Württemberg (ALWR-BW) in Abstimmung mit dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg

Zur Begutachtung durch die Deutsche Forschungsgemeinschaft (DFG)

Version 1.0, 01.11.2012

Inhalt

1	Zur Einordnung dieses Umsetzungskonzeptes.....	3
2	Ausgangslage	5
2.1	Netzinfrastruktur: BelWü	5
2.2	IT-Basisdienst für Backup von Arbeitsplätzen und Servern: Arbeitskreis TSM	5
2.3	Informationsdienste: Integriertes Bibliothekssystem Baden-Württemberg (IBS BW).....	6
2.4	Große wissenschaftliche Datenmengen: Large Scale Data Facility	6
2.5	Laufende Landesprojekte	7
3	Kernkonzept bwDATA.....	8
3.1	Erfassung Ist-Zustand und Prognose	8
3.2	Analyse und abgeleitetes Vorgehen.....	9
3.3	Kostenabschätzung.....	12
4	Maßnahmen im Zeitraum 2013-2014	13
4.1	Beschaffungsmaßnahmen	13
4.2	Projektmaßnahmen.....	13
5	Governance-Strukturen und Compliance-Aspekte	16
6	Zusammenfassung, zeitliche Gliederung und Bezug zu weiteren Landesaktivitäten	17
7	Referenzen	18

Autoren

Die Mitglieder des Arbeitskreises der Leiter der wissenschaftlichen Rechen- und Informationszentren von Baden-Württemberg (ALWR-BW):

Prof. Dr. Gerhard Schneider, Universität Freiburg

Prof. Michael Hebgen, Universität Heidelberg

Karl-Wilhelm Horstmann, Universität Hohenheim

Prof. Dr. Hannes Hartenstein, Karlsruher Institut für Technologie

Prof. Dr. Marcel Waldvogel, Universität Konstanz

Dr. Peter Leinen, Universität Mannheim

Prof. Dr.-Ing. Dr. h.c. Dr. h.c. Michael Resch, HLRS, Stuttgart

Prof. Dr. Thomas Walter, Universität Tübingen

Prof. Dr. Hans Peter Großmann, Universität Ulm

Peter Castellaz, Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg

sowie Jörn Beutner (S), Dr. Thomas Bönisch (S), Rolf Bogus (HD), Dr. Steffen Bücheler (HO), Mathias Feiler (HO), Dr. Ulrich Hahn (TÜ), Dr. Marc Hemberger (HD), Dr. Heinz Kredel (MA), Thomas Nau (UL), Dr. Sebastian Rieger (KA), Klaus Scheibenberger (KA), Jos van Wezel (KA).

1 Zur Einordnung dieses Umsetzungskonzeptes

Der Bedarf an Kapazität für die Speicherung von Daten hat in den vergangenen Jahren stark zugenommen. Im wissenschaftlichen Umfeld wird diese Entwicklung beispielsweise durch umfangreiche hochauflösende Messdaten und zunehmend leistungsfähigere Sensorik unterstützt. Nahezu alle Wissenschaftsbereiche arbeiten „daten-getrieben“ und tragen zum exponentiellen Wachstum der zu speichernden Datenmenge bei [1]. Dieser Bedarfszuwachs ist jedoch nicht auf den Bereich der Forschung beschränkt, da neue Dienste sowohl im Bereich der Lehre als auch der Verwaltung ebenfalls erhebliche zusätzliche Datenmengen produzieren. Zu nennen sind hier etwa die Aufzeichnung von Vorlesungen im Rahmen von E-Learning Aktivitäten, der Einsatz höchst auflösender Bilder im Rahmen der medizinischen Ausbildung oder auch elektronische Studierenden- und Personalakten. Der dabei zu erwartende Speicherbedarf steigt schneller als dies durch die mit neuen Speichertechnologien verfügbare höhere Kapazität kompensiert werden kann. Die Universitäten stehen somit vor der Herausforderung, den Betrieb der Datenspeicherung nachhaltig zu gewährleisten und für eine Unterstützung der Wissenschaft auszubauen. Eine Förderung des konsequenten Ausbaus benötigt daher ein Konzept für die Bündelung von Speicherressourcen sowie für eine föderative Erbringung und gemeinsame Nutzung der daten- und somit speicherintensiven Dienste, im Folgenden als *Speicherdienste* bezeichnet, an landesweit verteilten Standorten, um die Diskrepanz zwischen benötigter und finanzierbarer Kapazität nachhaltig zu minimieren. Das vorliegende Umsetzungskonzept bildet den Rahmen für die Kooperation sowie Konsolidierung und Fortentwicklung bei der Erbringung und Nutzung von datenintensiven Diensten an den baden-württembergischen Universitäten. Es beschreibt Maßnahmen und Projekte, die in den Jahren 2013 und 2014 angestoßen werden sollen, sowie ein Vorgehensmodell zur weiteren Ausgestaltung der daten- und somit speicherintensiven Dienste in den nachfolgenden Jahren. Das hier entwickelte Konzept (Phase I) konzentriert sich zunächst auf die neun Universitäten des Landes einschließlich der vier dort angesiedelten Medizinischen Fakultäten, es wird im nächsten Schritt auf weitere Hochschularten und Einrichtungen (Landesbibliotheken, Archive) erweitert werden (Phase II); der Klinikbetrieb kann in Phase II jeweils in standortbezogenen Kooperationen an den vier Standorten berücksichtigt werden.

Die Speicherkapazität für datenintensive Dienste wird in Rechenzentren in Form von Speichersystemen mit unterschiedlichen Charakteristiken und Zugriffsverfahren realisiert. Nutzer der darauf aufsetzenden/dadurch unterstützten Dienste sind z.B. Wissenschaftler, Studierende oder Einrichtungen. Ausgehend von den Speicherbedarfen allgemeiner IT- und Informationsdienste sowie vom Bedarf für das Speichern großer wissenschaftlicher Daten („Big Data in Science“) werden anhand geeigneter Speichersysteme und -technologien Lösungen für eine landesweite Realisierung beschrieben. Um die Eignung der betrachteten Lösungen für die Konsolidierung und den Ausbau von Speicherdiensten zu überprüfen, werden im Konzept Bewertungsaspekte definiert, die es ermöglichen, die technische und qualitative Realisierbarkeit in Bezug auf die Nutzung und Erbringung der Dienste (vgl. Tabelle 1) an landesweit verteilten Standorten zu evaluieren. Hierbei wird einerseits die Entfernung der Dienste zu den Nutzern und andererseits die Entfernung der Dienste zu den Speichersystemen betrachtet. Abbildung 1 veran-

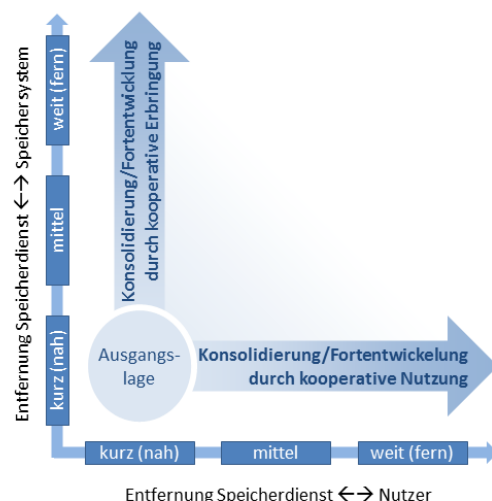


Abbildung 1: Vorgehensmodell für die Realisierung

den. Hierbei wird einerseits die Entfernung der Dienste zu den Nutzern und andererseits die Entfernung der Dienste zu den Speichersystemen betrachtet. Abbildung 1 veran-

schaulich dies. Zukünftig soll hierbei, ausgehend von den bestehenden Anforderungen und Diensten, eine zunehmende Verlagerung einzelner Datenspeicherdienste an wenige Standorte im Land in „Daten-Kompetenzzentren“ und in „Data Life Cycle Labs“¹ erfolgen. Hierfür gelten die Grundsätze, dass **Performance-Speicher lokal zur datengenerierenden Einheit gestellt wird, wenn der Datentransport an eine weiter entfernte Stelle technisch oder wirtschaftlich nicht sinnvoll ist. In diesem Fall soll grundsätzlich eine landesweite Nutzung vorgesehen werden, entweder allgemein oder für eine spezielle „Community“.** In allen anderen Fällen soll die Speicherung nur an wenigen Stellen im Land kooperativ erfolgen. Hierbei kommt der LSDF am KIT eine Musterrolle bei Backup und Archivierung großer wissenschaftlicher Daten zu, ebenso entsprechen die TSM-Verbünde diesem Grundsatz bezogen auf Backup und Archivierung von Arbeitsplatzrechnern und Servern.

In Bezug auf die Speichersysteme lassen sich hierbei die in Abbildung 2 dargestellten Charakteristika in Anlehnung an „Tiered-Storage“-Konzepte (vgl. Hierarchisches Speichermanagement – HSM) unterscheiden. Hierbei nimmt die Wartezeit (Latenz) beim Zugriff auf die Daten ausgehend von verfügbaren Performance-Medien (vgl. Festplatten) zu Nearline-Medien (z.B. Tapes) zu. Die in der Abbildung dargestellten Charakteristika bilden die groben Kategorien für die Einordnung von landesweiten Speichersystemen an den baden-württembergischen Universitäten. Speicherdienste können hierbei

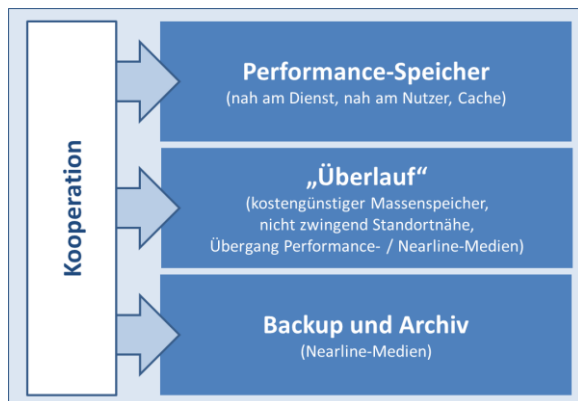


Abbildung 2: Abbildung auf Speicherebenen und -systeme

auf verschiedenen bzw. mehreren der skizzierten Ebenen realisiert werden. Durch Konsolidierung und Kooperation können Skaleneffekte bei gemeinsamen Beschaffungen sowie eine Reduktion der Betriebskosten erzielt werden. Hinsichtlich der Anforderungen lassen sich hier neben Kapazität auch weitere Charakteristika wie IOPS (Input/Output Operations per Second), Latenz, Durchsatz und WAN-fähige (Wide Area Network) Zugriffsprotokolle differenzieren, die bei der Evaluierung der möglichen Entfernung der Dienste vom Nutzer bzw. den Speichersystemen betrachtet werden müssen. Hierbei wird auf die Ergebnisse bestehender Landesprojekte aufgebaut.

Eine aktuelle Bestandsaufnahme der an den Universitäten im Land Baden-Württemberg vorhandenen Speichersysteme, bestehender Kooperationen sowie der für den Betrieb notwendigen Mittel in den Abschnitten 2 und 3 bildet die Entscheidungsgrundlage für den in Abschnitt 4 dargestellten Maßnahmenkatalog. Dort werden kurz- und mittelfristige Maßnahmen vorgestellt, die in einem Stufenkonzept das in Abbildung 1 gezeigte Vorgehensmodell im Hinblick auf die Vision kooperativer und konsolidierter Speicherdienste abbilden. In Abschnitt 5 werden die hierfür erforderlichen standortübergreifenden Governance-Strukturen beschrieben, die für den Betrieb der Speicherdienste und für die wissenschaftsadäquate Aussteuerung der Dienste und Ressourcen erforderlich sind. Abschnitt 6 stellt die zeitliche Planung für das Umsetzungskonzept dar.

¹ Der Begriff entstammt der „Large Scale Data Management and Analysis (LSDMA)“-Portfolioerweiterung der Helmholtz Gemeinschaft [2] und wird entsprechend verwendet.

2 Ausgangslage

Datenhaltung und Dienstleistung an den Universitäten Baden-Württembergs erfolgen derzeit in den meisten Fällen standortbezogen, Anbieter und Nutzer befinden sich also in der Regel im selben lokalen Netz. Im Zuge der fortschreitenden Kooperation zeigt sich die Notwendigkeit, verstärkt föderative Strukturen zur Erhöhung der Effizienz und Vereinfachung der Nutzung des Dienstangebots zu etablieren. Die nachfolgenden Abschnitte geben eine Übersicht zu den bereits aufgebauten und auszubauenden Kooperationen sowie zu den im Land realisierten Speicherdiensten (siehe auch [3]).

2.1 Netzinfrastuktur: BelWü

Leistungsfähige und hochverfügbare Netze sind die Grundlage jeglicher IT-Services-Entwicklung, insbesondere dort, wo IT-Dienste verteilt erbracht werden sollen. Die Bedeutung der Vernetzung wurde vom Land Baden-Württemberg sehr frühzeitig erkannt und mit dem Aufbau und ständigen Ausbau des BelWü (Baden-Württembergs extended LAN) seit 1987 nachhaltig umgesetzt. Mit der Einrichtung eines landesweiten Hochschulnetzes, welches im Kern alle Universitäten redundant mit derzeit 10Gbit/s verbindet, ist es dem Land möglich, auch alle anderen Landeshochschulen (Pädagogische Hochschulen, Hochschulen für angewandte Wissenschaften, die Duale Hochschule, Musik- und Kunsthochschulen, öffentliche Bibliotheken sowie andere Forschungseinrichtungen) mit einzubeziehen.

Fazit: Verfügbarkeit und Bandbreite des BelWü sind die Basis für eine föderative Dienstleistung.

Beteiligte Einrichtungen: die neun Landesuniversitäten, sechs Pädagogische Hochschulen, 25 Hochschulen für angewandte Wissenschaften, die Duale Hochschule mit acht Standorten, Schulen, Bibliotheken und weitere wissenschaftliche Einrichtungen.

2.2 IT-Basisdienst für Backup von Arbeitsplätzen und Servern: Arbeitskreis TSM

Basierend auf dem leistungsfähigen und hochverfügbaren Landeshochschulnetz BelWü erbringen vier Verbünde (Heidelberg-Mannheim-Freiburg, Ulm-Tübingen-Konstanz, Stuttgart-Hohenheim, KIT mit den pädagogischen Hochschulen des Landes) seit weit mehr als 10 Jahren kooperativ Dienstleistungen im Bereich Backup. Bei diesem konsortialen Versorgungsmodell ist besonders hervorzuheben, dass die Basisdienstleistungen sich nicht nur auf den jeweils regionalen Kundenstamm, einschließlich vieler Hochschulen, beschränken, sondern auch in einem landesweiten Kontext eng verflochten sind und damit voll und ganz dem gemeinsam erarbeiteten kooperativen Versorgungsmodell entsprechen. Durch die Nutzung einheitlicher Software wurden bereits in der Vergangenheit erhebliche Synergieeffekte erzielt. Diese wurden auch in der mehrfach positiven Begutachtung durch die DFG sowie die daran gekoppelte Förderung durch Bund und Land bestätigt.

Die landesweite Backup-Struktur bietet auch die Spiegelung von Teilbereichen der anfallenden Daten im Rahmen eines gemeinsamen Disaster-Recovery-Szenarios an und nutzt die sich aus der einheitlichen Software ergebenden Vorteile effektiv und effizient aus, was sich auch in der sehr hohen Akzeptanz dieser Lösung zeigt. Derzeit werden etwa 8.000 Arbeitsplatzrechner und 1.300 Server in den Regionalverbänden gesichert. Die Planung und Fortentwicklung im Rahmen dieses Umsetzungskonzeptes obliegt einer landesweiten Arbeitsgruppe, der ALWR-BW fungiert als Steuerungsgremium.

Fazit: Durch Kooperation und geographische Verteilung wurde die Basis für eine effektive, verbesserte Nutzung des Dienstes geschaffen, die im Rahmen des Umsetzungskonzeptes fortgeschrieben wird.

Beteiligte Einrichtungen: die neun Landesuniversitäten sowie weitere Hochschulen des Landes, organisiert in vier regionalen Verbänden

2.3 Informationsdienste: Integriertes Bibliothekssystem Baden-Württemberg (IBS|BW)

Wegweisend für kooperative, landesweite Lösungen ist das Betriebskonzept für ein einheitliches integriertes Bibliothekssystem des Landes. Hierfür wurde ein Bibliotheksverwaltungssystem beschafft, welches von allen wissenschaftlichen Bibliotheken in Baden-Württemberg eingesetzt werden kann. Zunächst werden 54 wissenschaftliche Bibliotheken eingebunden: fünf Universitätsbibliotheken, die beiden Landesbibliotheken sowie die Bibliotheken der Landeshochschulen. Das Betriebskonzept beinhaltet die zentrale, hochverfügbare Bereitstellung der notwendigen technischen Infrastruktur am Zentrum für Datenverarbeitung der Universität Tübingen (ZDV). Die Betreuung der bibliothekarischen Anwendungen wird vom Bibliotheks-Service-Zentrum (BSZ) in Konstanz geleistet. Die Datenhaltung für das IBS|BW erfolgt am Dienststandort.

Fazit: Mit IBS|BW konnte eine komplexe, verteilte und heterogene Architektur vereinheitlicht und mit zentralem Betrieb ein föderativer Dienst aufgebaut werden.

Beteiligte Einrichtungen: Fünf Landesuniversitäten (FR, S, HO, TÜ, UL), die beiden Landesbibliotheken und 46 wissenschaftliche Bibliotheken weiterer Hochschulen.

2.4 Große wissenschaftliche Datenmengen: Large Scale Data Facility

Seit 2009 ist am KIT und an der Universität Heidelberg die Large Scale Data Facility (LSDF) im Aufbau [4]. Mit der LSDF hat das Land Baden-Württemberg den Ausgangspunkt für die Realisierung eines Landeskonzepts zur Speicherung unstrukturierter wissenschaftlicher Daten für die Hochschulen des Landes gesetzt. Entsprechend der Antragstellung [4] bietet die LSDF Überlaufspeicher und Speicher für Archivierung von wissenschaftlichen Daten für eine Aufbewahrungszeit von 5 bis 10 Jahren. Dabei wird der Speicher an der Universität Heidelberg für Daten der Systembiologie und Genomforschung eingesetzt, die an der Universität Heidelberg und mit ihr kooperierenden Instituten anfallen. Der Speicher am KIT bedient als Überlaufspeicher und Archiv die baden-württembergischen Hochschulen sowie als „Performance-Speicher“ einige ausgewählte Institute. Die LSDF bietet Speicher im Petabyte-Bereich und ist durch eingebundene Rechnersysteme ideal für die Online-Verarbeitung von wissenschaftlichen Messdaten geeignet. Eine Anbindung an die vorgesehenen Rechnersysteme bwUniCluster, bwForCluster und ForHLR ist, wie von der DFG bereits empfohlen, in Vorbereitung. Hierzu wird die erforderliche BelWü Infrastruktur auf 100Gb/s aufgerüstet. An der Universität Heidelberg sind bereits 1,3 PB Daten der Systembiologie und Genom-Forschung gespeichert. Am KIT werden Daten der Hochschulen und wissenschaftlichen Institute in Baden-Württemberg gespeichert und auf Bandsysteme gesichert. Dabei bringt das KIT sein Wissen und seine Erfahrung aus dem Betrieb des WLCG Tier-1 Daten- und Analysezentrums in Karlsruhe (GridKa) in den Betrieb der LSDF ein.

Viele wissenschaftliche Institute haben einen dringenden Bedarf an der Speicherung und dem Management großer Datenmengen. Dies trifft zukünftig auch auf andere Landeseinrichtungen wie Bibliotheken und Archive zu.

Fazit: Die LSDF bildet insbesondere für wissenschaftliche Messdaten aber auch als Basis des Landesprojekts bwLSDF für Hochschulen und Landeseinrichtungen eine ideale Grundlage, um die föderative Speicherung von Daten zu unterstützen. Die direkte Verbindung der LSDF mit den zukünftigen Clustern (bwUniCluster, bwForCluster, ForHLR) verstärkt die Attraktivität der jeweiligen Ressourcen.

Beteiligte Einrichtungen: KIT und Universität Heidelberg als Betreiber, Universitäten und Hochschulen sowie weitere Landeseinrichtungen als Nutzer

2.5 Laufende Landesprojekte

Die Kooperation der Universitäten bei der Realisierung landesweiter Speicher-Lösungen und -Dienste kann bereits auf mehrere Projekte, die zu diesem Zweck gemeinsam geplant und mit finanzieller Unterstützung des Landes umgesetzt wurden, aufbauen. Diese Projekte adressieren unterschiedliche Anforderungen an die Datenhaltung.

bwFLA: Unter der Federführung des Rechenzentrums der Universität Freiburg entstehen im Projekt bwFLA (Funktionale Langzeitarchivierung) gemeinsam mit dem BSZ, dem Projekt KüdKA / HFG Karlsruhe, dem Landesarchiv Baden-Württemberg (LA-BW) und dem kiz der Universität Ulm Lösungen für die Langzeitarchivierung von Daten. Das Projekt hat das Ziel, die praktischen Arbeitsabläufe (*Workflows*) zur Archivierung von (komplexen) digitalen Objekten zu erweitern, um ein funktionsfähiges System zur funktionalen digitalen Langzeitarchivierung zu entwickeln. Dabei konzentriert sich das Vorhaben auf die Unterstützung der Arbeitsabläufe der Nutzer bei der Einstellung der Daten (*Ingest*) und auf Methoden zur langfristigen Zugriffssicherung (*Access*).

bwLSDF: Im Projekt bwLSDF wird eine Erweiterung der Large Scale Data Facility des KIT zur Datenspeicherung für die baden-württembergischen Hochschulen entwickelt und exemplarisch umgesetzt. Schwerpunkte sind eine Erweiterung um flexible Zugriffsverfahren (Block- und File-Zugriff, Backup, Archiv) auf den Überlaufspeicher sowie ein Konzept für die Realisierung allgemeiner Speicherbedarfe (z.B. File-Services). Beide Erweiterungen ermöglichen eine für die Nutzer transparente und ubiquitäre Bereitstellung von Speicher-Ressourcen über das bestehende Landeshochschulnetz.

bwIDM: Eine wesentliche Anforderung für alle landesweiten Dienste stellt die Integration geeigneter Zugriffskontrollen dar, die im Rahmen des Projekts bwIDM basierend auf den existierenden Identitätsmanagement-Systemen der Hochschulen in Baden-Württemberg föderativ umgesetzt wird. Das Ziel des federführend vom KIT unter Beteiligung aller Universitäten des Landes bearbeiteten Projekts bwIDM besteht darin, auf den Ebenen der Authentifizierung und Autorisierung Grundlagen für standortübergreifende und ortsunabhängige Dienstnutzung zu legen. Durch das Landeskonzept bwIDM wird ein nahtloser Zugriff auf über das Land verteilt angebotene Ressourcen und Dienste aus einem lokalen Kontext heraus ermöglicht werden [5].

bwGRiD Ergänzende Maßnahmen: Dieses Projekt stellt die Fortentwicklung des bwGRiD-Projekts dar, bei dem alle neun baden-württembergischen Universitätsrechenzentren erfolgreich und eng zusammen arbeiten, um die Transformation zu dem in [7] ausgeführten bwHPC-Konzept durchzuführen. Als Filesystem wird derzeit an allen Standorten Lustre verwendet, wobei es am KIT eine zentrale Instanz für Backup-Zwecke gibt. Im bwGRiD sind seit Betriebsbeginn viele Terabyte (TB) an Nutzerdaten erzeugt worden. Im Rahmen des Projektes bwGRiD e.M. widmet sich ein eigenes Arbeitspaket den Werkzeugen zum Datenmanagement.

3 Kernkonzept bwDATA

In den folgenden Abschnitten wird eine Analyse der existierenden und zukünftig erforderlichen Kapazitäten und Anforderungen für den Betrieb der Speicherdienste für die baden-württembergischen Universitäten durchgeführt. Um den Betrieb nicht nur an steigende Anforderungen aus Forschung und Lehre anzupassen, sondern auch Synergien für Effizienzsteigerungen und Kostensenkung zu nutzen, sind neue kooperative Landesspeicherdienste erforderlich, die im Abschnitt 3.2 erläutert werden.

3.1 Erfassung Ist-Zustand und Prognose

Als Basis für die Ermittlung der Anforderungen an die Datenhaltung an den Universitäten des Landes Baden-Württemberg wurde im Rahmen der Erstellung des Umsetzungskonzepts eine Erhebung des IST-Zustands durchgeführt. Für eine Stärkung und bessere Nutzung der föderativen Strukturen ist eine dienstbezogene Betrachtung erforderlich. Um eine Identifizierung von zukünftig kooperativ zu erbringenden Speicherdiensten zu ermöglichen, wurden die Ergebnisse für Dienstkategorien in Tabelle 1 zusammengefasst. Für jede Kategorie wurden die aktuell vorhandene Kapazität, die momentane Eignung für die föderative Erbringung sowie die Investitions- und Betriebskosten erhoben. Für den Zeitraum bis 2017 wird neben einer Schätzung des notwendigen Kapazitätszuwachses eine Abschätzung der zukünftigen Entwicklung für die einrichtungsübergreifende kooperative Erbringung/Nutzung der Dienste angegeben.

Dienste / Kategorien	Kooperative Nutzung / Erbringung des Diensts möglich?		Kapazität (in TB)	Kapazitätszuwachs Bis 2017	Kosten €/TB/Jahr
	2012 (IST-Zustand)	2017 (Prognose)	2012 (Schätzung)	2017 (Schätzung)	2012 (Schätzung)
E-Mail und Groupware	nein	ja	500	200%	600
Fileservices	teilweise	ja	1.150	300%	400
Informationsdienste	teilweise	ja	150	300%	400
Datenbanken	teilweise	teilweise	250	100%	600
Virtualisierung	nein	ja	400	300%	600
Wiss. Anwendungen	ja	ja	20.200	1.000%	400
Backup	ja	ja	15.750	500%	150
Archiv	ja	ja	11.600	500%	150

Tabelle 1: Ergebnisse der Ermittlung des IST-Zustands 06/2012, Angaben beruhen zum Teil auf Schätzungen und Erwartungen für die technologische Entwicklung

Für den E-Mail-Dienst kann aufgrund der niedrigen Durchsatzanforderungen der Nutzer ein dezentrales Nutzungskonzept zentralisierter Dienste erarbeitet werden. Das gleiche gilt für landesweite Backup und Archiv-Lösungen, die zusätzlich auch von der räumlichen Verteilung der Standorte und der damit verbundenen dezentralen Speicherung profitieren. Darüber hinaus stellen Backup und Archiv sowie insbesondere auch die wissenschaftlichen Anwendungen den größten Anteil der zu realisierenden Speicherkapazität. Daher sind große Skaleneffekte und Synergien bei einer landesweit kooperativen Diensterbringung zu erreichen. Für die Speicherung von E-Learning Inhalten ermöglicht die Entwicklung neuer landesweiter Konzepte für Informationsdienste eine nachhaltige Konsolidierung.

Im Bereich der Fileservices ist dies in Teilbereichen ebenfalls möglich, was bereits im Projekt bwLSDF adressiert wird. Applikationen, die aufgrund hoher I/O-Last starke Anforderungen an die Latenzen zum nachgelagerten Speichersystem aufweisen (vgl. Datenbanken), werden auf absehbare Zeit nahe am Speichersystem realisiert bleiben müssen.

Die in diesem Abschnitt aufgezeigte Bestandsaufnahme verdeutlicht die heterogenen Anforderungen an nachhaltige Datenspeicherdienste. Vor dem Hintergrund des fortwährenden technologischen Fortschritts sowie den zu erwartenden neuen Anforderungen an die Speicher-Infrastruktur der Universitäten ist eine kontinuierliche bzw. erneute Erfassung der Dienste und deren charakteristischer Anforderungen während der Umsetzung erforderlich.

3.2 Analyse und abgeleitetes Vorgehen

Ein Ausbau der Kapazität und Leistungsfähigkeit der Speicherdienste stützt sich ab auf Konsolidierung, Skaleneffekte, neue technologische Entwicklungen sowie abnehmende Preise pro TB. Die im Kapitel 1 erläuterte zunehmende Nachfrage erfordert jedoch weitaus umfangreichere Erweiterungen und einen erhöhten Mitteleinsatz (siehe auch [6]). Um für diese Erweiterungen einen effizienten Mitteleinsatz zu gewährleisten, sind Speicherdienste, wie in Tabelle 1 beschrieben, erforderlich, die z.B. durch kooperative Speicherbereitstellung zukünftig zusätzliche Skaleneffekte ermöglichen. Im Rahmen von bwDATA sollen solche Speicherdienste, die gemeinsam von mehreren Universitäten bereitgestellt bzw. genutzt werden, zusätzlich gefördert werden und so zukünftige Konsolidierungen ermöglichen. Die Prüfung der Wirtschaftlichkeit der Speicherdienste gewährleistet dabei eine effiziente Mittelverwendung.

Im Folgenden werden auf Basis der in Tabelle 1 genannten Daten die verschiedenen Dienstekategorien hinsichtlich der zukünftigen Ausprägung bewertet und ein Vorgehensvorschlag unterbreitet.

E-Mail und Groupware: Dienst und Speichersystem eignen sich sehr gut für eine kooperative Erbringung und Nutzung des Dienstes. Besondere Herausforderungen bei der Einführung gemeinsamer E-Mail-Dienste bestehen allerdings darin, die deutlich unterschiedlichen Gegebenheiten und Gepflogenheiten der beteiligten Institutionen zu vereinen. Besonders tritt dies in der Heterogenität bezüglich des lokal genutzten Identitätsmanagements, der verwendeten Serversoftware, im Ablageformat sowie im Zentralisierungsgrad hervor. Deshalb soll zunächst im Rahmen eines Projektes untersucht werden, in welcher Weise E-Mail-Dienste sinnvoll kooperativ für die Universitäten des Landes erbracht werden können.

File Services: Um den hohen wissenschaftlichen Standard in Baden-Württemberg zu erhalten und weiter auszubauen, wird eine noch aktivere Kooperation zwischen den verschiedenen Forschungseinrichtungen notwendig werden. Die gemeinsame Arbeit an Dokumenten, der schnelle Austausch von Daten sowie die rasche und flexible Bereitstellung projektbezogener Speichercontainer nehmen dabei einen hohen Stellenwert ein. Zukünftig werden die vorhandenen klassischen Ansätze (z.B. E-Mail-Anhänge oder Netzlaufwerke) von Entwicklungen abgelöst, welche sich die neuen Cloud-Paradigmen zunutze machen, um so ein Speicher-Agglomerat zu etablieren und über die Landeseinrichtungen aufzuspannen. Somit wird der im Land vorhandene Speicher integriert, dezentral verwaltet und standortübergreifend genutzt. Dafür notwendige Technologien wie beispielsweise Cloud Data Management Interface (CDMI), Object Storage, REST (Representational State Transfer) und Sync&Share (vgl. Enterprise DropBox-Alternativen) werden aktuell in bwLSDF verglichen und in Pilotdiensten evaluiert.

Informationsdienste: Landesweit wurden in den vergangenen Jahren an den Universitäten Lösungen für die Verbesserung der Qualität der Lehre realisiert. Hierbei zählen zu den Innovationen auch E-Learning- sowie Streaming-Systeme, die die Leistungsfähigkeit der Hochschulen nachhaltig stärken. Insbesondere die Speicherung von aufgezeichneten Lehrveranstaltungen für deren Nachbereitung durch die Studierenden („on-demand streaming“) sowie die Verteilung der Lehrinhalte an andere Kooperationsstandorte umfasst ein immer größeres Datenvolumen. Die zunehmende Verfügbarkeit von Aufnahmegegeräten mit hohen Auflösungen unterstreicht diesen Trend zusätzlich. Zusätzlich werden die Lehrveranstaltungen in den E-Learning- und Campus-Management-Systemen mit weiteren Inhalten angereichert und durch klassische Web-Seiten und Kollaborationsplattformen ergänzt, die ebenfalls zur Erhöhung des Datenvolumens der Informationsdienste beitragen.

Virtualisierung: Die Virtualisierung von Servern der Rechenzentren bildet an vielen Hochschulen die Basis für eine hochverfügbare und effiziente Bereitstellung von serverbasierten Diensten. Zur Erhöhung der Verfügbarkeit und der Realisierung eines Disaster-Recovery wird die Virtualisierungsumgebung häufig an zwei Standorten betrieben oder als ein Cluster über die Standorte gespannt. Über die Bereitstellung von virtuellen Servern für den rechenzentrumsinternen Betrieb hinaus wird auch eine rasche Provisionierung von virtuellen Servern für die Fakultäten und zentralen Einrichtungen ermöglicht. Mit der zunehmenden Virtualisierung und damit einhergehendem Grad der Konzentration von Systemen steigen die Anforderungen an eine leistungsfähige und hochverfügbare Storage-Infrastruktur, die auch eine einfache Datenmigration zwischen Standorten erlaubt.

Wissenschaftliche Anwendungen: Die Generierung und Verarbeitung der typischerweise großskaligen Daten aus den wissenschaftlichen Anwendungen stellen besonders hohe Anforderungen an Durchsatz und Qualität der Infrastruktur. Daher wird die Datenhaltung, bspw. für Simulation und bestimmte Experimente, auch zukünftig in unmittelbarer Nähe zur Datenquelle bzw. zu den Datenverarbeitungssystemen stattfinden müssen. Unabhängig davon sind für wissenschaftliche Anwendungen standortübergreifende Szenarien tragfähig, zum Beispiel Überlauf- und Archivspeicher. Der Zugang zu dem gemeinsamen Speicher soll für die Einrichtungen transparent und von den verschiedenen Standorten weitgehend automatisiert möglich sein. Dies ist auch aus einem wirtschaftlichen Betrachtungswinkel heraus erforderlich. Die für die langfristige Speicherung wissenschaftlicher Masendaten notwendige Zugriffsqualität (Durchsatz und Latenz) ist in der Regel deutlich geringer und damit kostengünstiger realisierbar als die zur Datenverarbeitung notwendige Zugriffsqualität. Eine Besonderheit bildet der Überlaufspeicher in HPC-Umgebungen, der einen höheren Durchsatz als es technisch und vor allem wirtschaftlich im WAN möglich ist, benötigt und daher in der Regel am Standort realisiert werden muss. Entsprechend dem vorgelegten bwHPC-Konzept [7] können somit entsprechende Kompetenzzentren entstehen. Eine Festlegung auf gemeinsame Standards und Schnittstellen für den Zugriff ist hierbei neben der landesweiten Verwendung des Überlaufspeichers auch für das Archiv zur langfristigen Auslagerung von Daten erforderlich. Die Archivierung verlangt eine sichere Aufbewahrung der Daten über einen langen Zeitraum. Typischerweise bedingt dies die Lagerung von mehreren Kopien auf unterschiedlichen Medien an unterschiedlichen Standorten mit großer räumlicher Distanz. Hierfür wird die Realisierung von wenigen Archivierungszentren im Land angestrebt.

Backup: Der hochperformante und redundante Ausbau des Landeshochschulnetzes BelWü erlaubt es Dienstleistungen im Bereich Backup noch stärker zu konsolidieren als dies in der Vergangenheit der Fall war. Da jedoch auch in diesem Bereich Teile der Dienstleistung sinnvoller vor Ort erbracht werden, sind neue Betriebsmodelle zwingend. Exemplarisch ist etwa das Remote-Management, also der Betrieb von Servern außerhalb der eigenen Einrichtungsgrenzen, zu nennen.

Archiv: An den Universitäten existieren unterschiedliche Lösungen für die Archivierung z.B. von sich nicht länger im regelmäßigen Zugriff befindlichen Daten. Die Archivierungslösungen sind in der Regel auf Massendaten ausgerichtet und erlauben eine kostengünstige langfristige Speicherung und hohe Durchsatzraten. Die Zugriffszeiten, die beim späteren Lesen der Daten aus dem Archiv in Kauf genommen werden müssen, sind jedoch im Vergleich zu den zuvor genannten Diensten hoch. An einigen Standorten sind für die Archivierung bereits redundante Systeme realisiert worden, die die Daten beispielsweise an getrennten Standorten mit separaten Archiv-Systemen (Tape-Libraries) lagern. Teilweise sind bereits Kooperationen zwischen Universitäten hinsichtlich der Realisierung eines gegenseitigen Redundanzstandorts für die Archivdaten etabliert worden. Der konsequente weitere Ausbau dieser Kooperationen ermöglicht eine weitere Effizienzsteigerung sowohl in Bezug auf die standortübergreifend erforderlichen Kosten für den Betrieb als auch für die erzielte Ausfallsicherheit.

Zusammenfassung Kernkonzept bwDATA: auf Basis der genannten Überlegungen wird folgendes Vorgehen in diesem Umsetzungskonzept für den Zeitraum 2013/2014 vorgeschlagen:

- Kurzfristige Maßnahmen zur Förderung von Speicherdienstleistung mit landesweiter Nutzung:
In einem ersten Schritt sollen die Skaleneffekte der großen Speicherkapazitäten wissenschaftlicher Anwendungen sowie bei Backup und Archiv weiter ausgenutzt werden. Durch die geförderte Realisierung von Speicherdienstleistung mit landesweiter Nutzung lässt sich zukünftig für alle Universitäten des Landes, z.B. durch die genannten Skaleneffekte und die Bildung von Kernkompetenzen resp. Konsolidierung, eine effizientere Mittelausnutzung erzielen. Hierbei ist nicht nur ein Management der Speicherdienste sondern auch der gespeicherten Daten erforderlich. Dieses muss in Form eines Data Life Cycles betrachtet werden, der von der Entstehung der Daten, über deren Verarbeitung bis hin zur Archivierung und Löschung reicht.
- Förderung von Entwicklungsprojekten für kooperative Speicherdienste:
In den anderen genannten Kategorien sind die Möglichkeiten der Konsolidierung nicht kurzfristig erreichbar. Daher wird in landesweiten Entwicklungsprojekten die mögliche Effizienzsteigerung durch Kooperationsmodelle evaluiert. Deren Ergebnisse fließen dann direkt in das ständig fortgeschriebene Umsetzungskonzept ein und garantieren eine Anpassung an technologische Entwicklungen. Nur so kann auch langfristig der Bedarf aus dem wissenschaftlichen Umfeld effizient gedeckt werden. So ist etwa die Bildung von Kompetenzzentren, die sich mit unterschiedlichen Aspekten des gesamten Data Life Cycle beschäftigen, angestrebt. Es wird zudem erwartet, dass kooperative Lösungen für Speicherdienste den Einsatz neuer Speichermedien wie SSDs unterstützen und hoch priorisierten Anforderungen aus dem Bereich des Energie- und Kosten-Managements entsprechen können.
- Einbezug weiterer Hochschularten und Landeseinrichtungen für Bedarfsfeststellung und Maßnahmenvorschläge hinsichtlich Speicherdiensten (Phase II):
Die Erfassung von Anforderungen und möglichen Betriebs- und Nutzungsszenarien wie in Tabelle 1 für die Landesuniversitäten dargelegt, muss bezüglich der Bedürfnisse und Sichten weiterer Hochschularten und Landeseinrichtungen fortgeschrieben werden. Hierzu wird im Dialog deren Bedarf quantitativ und qualitativ erfasst und in die Fortschreibung des Landeskonzeptes ab 2014 integriert.

Dabei werden folgende Grundsätze beachtet:

- **Performance-Speicher wird in der Nähe der datengenerierenden Einheit (Messgerät, Hochleistungsrechner etc.) installiert, wenn der Datentransport an eine weiter entfernte Stelle technisch oder wirtschaftlich nicht sinnvoll ist. In diesem Fall soll grundsätzlich eine landesweite Nutzung vorgesehen werden, entweder allgemein oder für eine spezielle Nutzergruppe.**
- **In allen anderen Fällen soll die Speicherung nur an wenigen Stellen im Land kooperativ erfolgen. Hierbei kommt der LSDF am KIT eine Musterrolle bei Backup und Archivierung großer wissenschaftlicher Daten zu, ebenso entsprechen die TSM-Verbünde diesem Grundsatz bezogen auf Backup und Archivierung von Standard-Arbeitsplatzrechnern und Servern.**

3.3 Kostenabschätzung

In den letzten Jahren stieg die Datendichte der magnetischen Datenträger getrieben von der technologischen Entwicklung jährlich um 30% bis 40%. In Bezug auf die Speichersystemkosten, die nicht alleine die Kosten der Datenträger umfasst, ist daher eine Reduzierung des Kaufpreises pro TB um 50% nach 4 bis 6 Jahre zu erwarten. Die Reduzierung bezieht sich jedoch ausschließlich auf die Kapazität. Je nach Leistungsanforderungen (z.B. Durchsatz, Zugriffszeit) entstehen beim Kauf und Betrieb der Speichersysteme höhere Kosten.

Aktuell werden ca. 10 PB der in Tabelle 1 gezeigten 20,2 PB für wissenschaftliche Anwendungen innerhalb kooperativer Landesprojekte (vgl. HLRS, LSDF Überlaufspeicher) genutzt. Bei einem angenommenen Zuwachs von 60% im kommenden Jahr ergibt sich eine erforderliche Erhöhung der Kapazität um 6 PB. Bei einem Wachstum von 60% p.a. kann, wie in Tabelle 1, für 2017 die 10-fache Kapazität für die Speicherung wissenschaftlicher Anwendungsdaten angenommen werden. Bei den landesweiten Backup Lösungen ist der Anteil an kooperativ realisierten Lösungen (vgl. TSM Verbünde) bereits höher und liegt bei ca. 10 PB. Der Kapazitätzuwachs liegt beim Backup bei ca. 43% p.a.

Bei 400 €/TB/Jahr ergeben sich für die prognostizierten 6 PB für wissenschaftlichen Anwendungen zusätzliche Kosten von ca. 2,4 Mio € für kooperative landesweite Speicherdienste. Für den Backup Speicherdienst kann entsprechend bei 150 €/TB/Jahr von ca. 645.000 € ausgegangen werden.

Wie in Tabelle 1 dargestellt lassen sich aus der Ist-Analyse sowie aus den zukünftigen Anforderungen weitere kooperative Speicherdienste ableiten. Ausgehend von einem zukünftigen kooperativen Anteil von annähernd 100% bei E-Mail, Fileservices, Informationsdiensten und Archiv (für die teilweise bereits Kooperationskonzepte existieren) sowie ca. 50% bei Datenbanken und Virtualisierung, ergibt sich für die Gesamtkosten bei kooperativen Datenspeicherdiensten inkl. der oben erwähnten 3,045 Mio € für wissenschaftliche Anwendungen und Backup ein Mittelbedarf von mindestens 4,1 Mio € p.a. Etwa 250.000 € verbleiben als Kosten bei dem „nicht-kooperativen Anteil“ für Speicherdienste bei Datenbanken und Virtualisierung.

Die prognostizierten Kosten beziehen sich, wie schon die Schätzungen in Tabelle 1, ausschließlich auf die Kapazität und deren Zuwachs in 2013. Abhängig von Performance-Anforderungen der Anwender und Einrichtungen können die Kosten deutlich höher ausfallen. Zusätzlich beziehen sich die Kosten aktuell nur auf den Bedarf an den Universitäten des Landes Baden-Württemberg. Mittelfristig gehört die Erweiterung des Landeskonzeptes auf weitere Einrichtungen aus dem Zuständigkeitsbereich des MWK zu den Zielen. In erster Linie ist hier der Einbezug der Hochschulen für Angewandte Wissenschaften zu nennen, aber auch weitere Einrichtungen wie die Landesbibliotheken und Archive. Gerade kleinere Einrichtungen werden durch die Partizipation an der föderativen Diensterbringung in ihrer Servicequalität gestärkt.

4 Maßnahmen im Zeitraum 2013-2014

4.1 Beschaffungsmaßnahmen

Die Umsetzung des vorliegenden Landeskonzeptes wird sich mit dem Übergang zur föderativen Dienstleistung über einen längeren Zeitraum erstrecken. Zu Beginn sind zwei konkrete Ausbauschaffungen für die Umsetzung des Landeskonzeptes geplant, die in den folgenden Abschnitten näher erläutert werden.

4.1.1 Wissenschaftliche Anwendungen

In einem ersten Schritt muss für die wissenschaftlichen Anwendungen der HPC-Überlaufspeicher in Stuttgart (HLRS) ausgebaut werden. Das derzeitige System in Stuttgart, ein HSM auf Basis HPSS, ist weder in seiner Kapazität noch in seiner Leistungsfähigkeit den aktuellen Anforderungen gewachsen, die insbesondere aufgrund der neuen HPC-Umgebung stark gestiegen sind. Das Beschaffungskonzept integriert diesen Ausbau mit der Erweiterung der bestehenden TSM-Umgebung, die kooperativ vom TSM-Regionalverbund der Universitäten Hohenheim und Stuttgart betrieben wird. Hierbei kann durch die entstehenden Synergien erhebliches Einsparpotenzial genutzt werden. Zudem wird der Bereich Ausbau Archiv kooperativ mit dem KIT durchgeführt. Das erforderliche Gesamtvolumen für den entsprechenden Ausbau sowie die Realisierung kooperativer Speicherdienste im Umfeld der wissenschaftlichen Anwendungen umfasst, wie im zugehörigen Antrag nach Artikel 143c GG ausgeführt, 2,5 Mio € mit einer angestrebten Förderung von 1,25 Mio € aus zentralen Mitteln des Landes. Antragsteller ist die Universität Stuttgart, beteiligt ist das KIT.

4.1.2 Backup

Stellvertretend für den TSM-Regionalverbund der Universitäten Konstanz, Tübingen und Ulm beantragt das Kommunikations- und Informationszentrum der Universität Ulm (kiz) Mittel zur Fortschreibung und insbesondere zur notwendigen Erweiterung der erforderlichen Backup-Kapazitäten. Der kurzfristige Handlungsbedarf leitet sich direkt aus dem stetig wachsenden Datenaufkommen sowie insbesondere auch aus wirtschaftlichen Gesichtspunkten ab, wie im beiliegenden Antrag dargelegt. Der Einsatz aktueller Band- und Schranktechnologien derselben Evolutionsreihe bietet neben deutlich geringeren Kosten pro Gigabyte weitere wesentliche Vorteile. Die vielfach höhere Packungsdichte erlaubt einen Ausbau mit Medien über viele Jahre hinweg und die Verwendung vergleichbarer Bandtechnologien stellt eine einfache Migration im Betrieb und ohne Zutun der Nutzer sicher. Folgerichtig nutzt das ZDV der Universität Tübingen diese Synergien aus und wird sukzessive die gesamte Backup Dienstleistung aus Ulm beziehen. Damit wird mittelfristig die gesamte Backup Infrastruktur der Universitäten Konstanz, Tübingen und Ulm zentral vom kiz betreut und fortgeschrieben, wobei der jeweilige Endnutzer-Support am Standort erbracht wird. Eine Erweiterung des Verbundes, etwa um die lokalen Hochschulen, ist konzeptionell ohne Probleme möglich. Das erforderliche Gesamtvolumen für den entsprechenden Ausbau sowie für die Realisierung kooperativer Speicherdienste umfasst, wie im zugehörigen Antrag nach Art. 143c GG ausgeführt, 0,6 Mio € mit einer angestrebten Förderung von 0,3 Mio € aus zentralen Mitteln des Landes. Antragssteller ist die Universität Ulm.

4.2 Projektmaßnahmen

Neben den zwei in Abschnitt 4.1 genannten Maßnahmen haben die Universitäten des Landes weitere gemeinsame Vorhaben zur Realisierung landesweiter Speicherdienste geplant.

1. Gemeinsame technische Konzeption der Servervirtualisierung, welche eine Migration virtueller Systeme zwischen verschiedenen Standorten erlaubt: Kooperation der Universitäten Mannheim, Freiburg, Hohenheim, Stuttgart und dem KIT.
2. Gemeinsame Konzeption für die föderative Erbringung des E-Mail-Dienstes an den Universitäten Hohenheim, Freiburg, Heidelberg, Stuttgart und dem KIT.
3. Gemeinsame Konzeption für Archivdienst, Preservation und Curation (KIT, Stuttgart).
4. Gemeinsame Konzeption für die Behandlung des „Data Life Cycle“ wissenschaftlicher Daten in der föderativen Speicherinfrastruktur (alle, Federführung KIT)

Die damit einhergehenden Projekte werden in den nachfolgenden Abschnitten beschrieben.

4.2.1 Virtualisierung

In Rahmen eines Projekts sollen Möglichkeiten evaluiert werden, wie die Virtualisierungsumgebungen rechenzentrumsübergreifend zu koppeln sind. Von dieser Maßnahme werden deutliche Vorteile durch eine übergreifende Nutzung des vergrößerten gemeinsamen Resource-Pools von virtuellen Servern erwartet, insbesondere die kurzfristige Bereitstellung von Serverkapazitäten. Auch Wartungsarbeiten an einzelnen Server-Standorten lassen sich so leichter kompensieren. Bei einer Einbeziehung einer standortübergreifenden Migration von virtuellen Servern lassen sich auch flexible Lastausgleichszenarien realisieren, die zu einer besseren Auslastung von Server-Ressourcen führen können.

Neben der Server-Virtualisierungstechnik wird die Untersuchung von weitverkehrstauglichen Speicher-Virtualisierungstechnologien zur Kopplung der Storage-Backends eine zentrale Aufgabe des Projekts werden. Diese Aufgabe setzt ein leistungsfähiges Netzwerk (WAN) und entsprechend konfigurierte lokale Netzwerke (VLAN, Firewall) voraus, deren Entwicklung damit ein integraler Bestandteil des Projekts sein muss. Der Einsatz einer gemeinsamen Speichervirtualisierung könnte die Basis für eine solche standortübergreifende Infrastruktur sein, die auch das Backup und ein Disaster Recovery von Servern einbezieht.

Die Universitäten sind derzeit in Verhandlungen mit der Firma VMware über den Abschluss eines Rahmenvertrags zur landesweiten Lizenzierung der Produkte zur Servervirtualisierung. Eine abgestimmte Beschaffung von Speicherinfrastruktur und ergänzender, speicherspezifischer Lösungen mit einer einheitlichen Schnittstelle zur Serverinfrastruktur sind weitere wichtige Bausteine zum Aufbau einer landesweiten verfügbaren Virtualisierungsumgebung. Bei der notwendigen Einführung eines standortübergreifenden Rechte- und Rollenkonzeptes wird auf die Ergebnisse von bwIDM zurückgegriffen werden können. Die damit erzielte Infrastruktur ist dazu prädestiniert, weitere Dienstleistungen wie Software as a Service und Desktop-Virtualisierung aufzubauen.

4.2.2 E-Mail

In einem Projekt soll zunächst die technische Machbarkeit, beispielsweise die Trennung von Dienst und Speicher, geprüft werden. Darauf aufbauend sind die organisatorische Umsetzbarkeit und die wirtschaftliche Machbarkeit zu untersuchen, wobei rechtlichen Rahmenbedingungen in diesem Kontext eine große Bedeutung zukommt. Besondere Beachtung wird auf folgende Aspekte gelegt:

- Archivierung von E-Mails inklusive Attachments
- Groupware (z.B. gemeinsame Kalender, Adressbücher, Dokumentenverwaltung, Aufgaben)

- Heterogene Klienten (besonders auch „mobile devices“ bzw. „mobile access“) müssen für E-Mail und Groupware-Zugriff unterstützt werden
- E-Mail-Policies, z.B. Spam- und Virenabwehr.

In dem skizzierten Projekt ist auch die Evaluierung von Diensten, die bereits von BelWü und dem DFN-Verein angeboten werden, einzubeziehen. Die Möglichkeiten eines weitergehenden Einkaufs dieser Dienstleistung wird ebenfalls betrachtet werden.

4.2.3 Archivierung

In der zweiten Phase des Konzepts bwDATA, welche auch die Erweiterung über den Kreis der Universitäten hinaus beinhaltet, soll ein Projekt zur Planung von Archivkapazitäten initiiert werden. Neben der Planung von Archivkapazitäten an einigen Standorten soll die Implementierung einer Bitstream Preservation und einer Data Curation in dem Projekt einen zweiten Schwerpunkt bilden. Für die betriebliche Umsetzung sind wenige spezialisierte Archivierungszentren mit großer räumlicher Distanz im Land für die Speicherung von mehreren Kopien auf unterschiedlichen Medien geplant. Hierbei kann auf die enge Abstimmung mit den schon laufenden Landesprojekten bwFLA und bwIDM gesetzt werden, um alle notwendigen Aspekte abzudecken, die für die Archivierung von digitalen Objekten über Zeiträume von mindestens 10 Jahren erforderlich sind. Damit sollen explizit die Forderungen der DFG zur Archivierung von wissenschaftlichen Daten erfüllt werden [8].

Da die Herausforderungen im Bereich der Langzeitarchivierung nicht allein von den universitären Rechenzentren geschultert werden können, ist eine enge Zusammenarbeit mit den Bibliotheken und dem Landesarchiv unabdingbar. Aus diesem Grund sollen die Einrichtungen schon in der Planungsphase eingebunden werden. Der Projektstart ist für das erste Quartal 2014 geplant.

4.2.4 Wissenschaftliche Anwendungen: Data Life Cycle Management

Daten aus Messungen von Experimenten und aus Simulationen durchgeführt auf Hochleistungsrechnern stellen ein besonders wertvolles Gut dar, da sie in der Regel nicht einfach wieder erzeugt werden können. In naher Zukunft kommt insbesondere dem Erkenntnisgewinn durch semantische Auswertungsmethoden eine immer größere Bedeutung zu. Um effizient Informationen aus den Daten zu gewinnen und damit wissenschafts- und gesellschaftsrelevante Schlussfolgerungen ziehen zu können, muss die Datenspeicherung verknüpft werden mit einem System zur Registrierung und Verwaltung der Metadaten. Automatische Analyse, Transport und Archivierung von großen Datenmengen und das Wiederauffinden der Daten ist nur möglich, wenn ausgehend von der Erzeugung der Daten der vollständige „Data Life Cycle“ festgehalten und in die Datenmanagementumgebung eingebunden wird. Verschiedene nationale und internationale Initiativen entwickeln dazu derzeit Methoden für den universellen Zugriff auf unterschiedliche Datenobjekte über Wissenschaftsdisziplinen hinweg.

In diesem Projekt sollen diese Erkenntnisse für den Aufbau einer universellen Dateninformations-Infrastruktur angewandt werden. In dem Projekt sind u.a. Arbeiten an einer „globalen“ Datenregistrierung, Unterstützung verschiedener Container-Formate sowie Schnittstellen zu Ressourcen vorgesehen. Die Anbindung verschiedener Wissenschaftsdisziplinen ist Teil dieses Projekts, in dem auch die Erkenntnisse aus bwIDM, bwFLA, bwLSDF und aus dem skizzierten Archivierungsprojekt einen direkten Einsatz finden.

5 Governance-Strukturen und Compliance-Aspekte

Das entwickelte Umsetzungskonzept bedarf in seiner Etablierung einer regelmäßigen und kontinuierlichen Überprüfung und Fortentwicklung. Hierfür ist ein geregelter Diskurs mit den Nutzern und weiteren Bedarfsträgern vorgesehen. Das Umsetzungskonzept setzt dabei auf drei Säulen:

- Etablierte Nutzer-/Aufsichtsgremien der Standorte für allgemeine IT- und Informationsdienste
- Ein neu einzusetzender landesweiter Nutzausschuss für wissenschaftliche Daten und für das zugehörige Data Life Cycle Management. Hinsichtlich der Behandlung großer wissenschaftlicher Daten soll bei mindestens jährlich stattfindenden Workshops der Nutzervertreter und der Betreiber eine Bedarfserhebung erstellt werden, welche die Basis für die konkreten Planungen der Betreiber bildet. Der ALWR-BW fungiert bereits als Steuerkreis für LSDF und übernimmt die Koordination der Fortentwicklung innerhalb des Umsetzungskonzepts für die Phase I. Hierfür wird aus den Ergebnissen der Workshops ein Bericht zu den Fortschritten von bwDATA erstellt. Für die Workshops wird auch eine Beteiligung von Landesprojekten, die zum Fortschritt der Umsetzung von bwDATA beitragen, sowie Vertretern von vergleichbaren Konzepten in anderen Bundesländern angestrebt, um den Austausch und die Weiterentwicklung von bwDATA voranzutreiben.
- Einbezug der Bibliotheken sowie der Hochschulen und weiterer Landeseinrichtungen bei der Entwicklung des Umsetzungskonzepts Phase II und bei der Ausprägung von Landesprojekten; die vier medizinischen Fakultäten sind nach Landeshochschulgesetz als Teil der Universitäten bereits in der ersten Phase am Konzept bwDATA beteiligt; die Universitätskliniken als autonome, getrennte Einrichtungen können im bwDATA-Konzept in Phase II in standortbezogenen Kooperationen an den jeweiligen Standorten einbezogen werden.
- Für neue, föderativ erbrachte Dienste wird basierend auf den vorhandenen Erfahrungen und Methoden ein Modell für eine Kostenverrechnung erarbeitet, welches neben der Verteilung der Beschaffungskosten insbesondere auch die Betriebskosten berücksichtigt. Hierbei sind insbesondere beim kooperativen Betrieb der Dienste auch Strukturen für die standortübergreifende Planung von Personalressourcen (z.B. in Form von Teams) zu realisieren.

Anträge mit Bezug zu Datenspeicherdiensten und mit angestrebter finanzieller Beteiligung des Landes aus zentralen Mitteln werden künftig nur noch im Rahmen des Konzeptes bwDATA gestellt. Für Anträge nach Art. 143c GG wird eine Eigenbeteiligung des Standorts von mindestens 50% verlangt.

Anforderungen an die Speicherdienste aus Compliance-Sicht sind gegeben durch die Vorgaben zur guten wissenschaftlichen Praxis [8] der DFG (z.B. 10 Jahre Speicherung der Daten) sowie durch gesetzliche Rahmenbedingungen. Dies wird auch die Etablierung landesweiter Richtlinien erfordern.

Zum anderen sind Richtlinien für die IT-Sicherheit der Speicherdienste erforderlich. Diese sind insbesondere für eine sichere kooperative Nutzung der landesweiten Speicherdienste essentiell. Da personenbezogene Daten gespeichert werden können, müssen ferner §7 LDSG (Auftragsverarbeitung) und §9 LDSG (Technische und organisatorische Maßnahmen) beachtet werden. Im Rahmen des laufenden Landesprojektes bwLSDF wird momentan geprüft, wie durch geeignete kryptographische Methoden ein unberechtigter Zugriff auf personenbezogene Daten verhindert werden kann.

Die notwendige Weiterentwicklung des Kostenverrechnungsmodells sowie eines Sicherheits- und Compliance-Konzepts wird in gemeinsamen Projekten unter Mitwirkung der Universitätsverwaltungen umgesetzt.

6 Zusammenfassung, zeitliche Gliederung und Bezug zu weiteren Landesaktivitäten

Die vorgestellten, im Rahmen des Umsetzungskonzeptes bereits konzipierten Maßnahmen sollen aufgrund des gegebenen Handlungsbedarfs zeitlich gegliedert umgesetzt werden. Die Beteiligten haben hierfür den Ablauf nach Tabelle 2 konzipiert, die zur Vollständigkeit auch die Aktivitäten aus anderen Projekten und Maßnahmen darstellt, die mit den Datenspeicherdiensten in Bezug stehen:

	Q4/2012	Q1/2013	Q2/2013	Q3/2013	Q4/2013	Q1/2014	Q2/2014	Q3/2014	Q4/2014
Phase I:									
E-Mail			Projektantrag	Projekt (4.2.2)					
File Services	Laufendes Landesprojekt bwLSDF								
Virtualisierung			Projektantrag	Projekt (4.2.1)					
Wiss. Anw.	Antrag		Realisierung			Projekt (4.2.4)			
Backup	Antrag		Realisierung						
Archiv						Projekt (4.2.3)			
Phase II:									
		Erarbeitung Umsetzungskonzept Phase II Verbunden mit den zugehörigen Projektanträgen							
Bezüge zu:									
BelWü 100G	Ausbau 100 Gigabit zwischen Standorten der LSDF und bwForCluster durch BelWü								
bwHPC			Inbetriebnahme bwUni-Cluster (KIT)		Inbetriebnahme bwFor-Cluster (HD/ MA, UL)		Inbetriebnahme bwFor-Cluster (FR, TÜ)		

Tabelle 2: Zeitplan der Maßnahmen des bwDATA-Umsetzungskonzeptes Phase I sowie der Erarbeitung des Umsetzungskonzeptes Phase II und Realisierungen im Bereich der Netzinfrastruktur und der Baden-Württemberg Cluster.

Von den acht in diesem Dokument betrachteten Diensten bzw. Dienstkategorien (siehe Tabelle 1) sind sechs Dienste durch laufende Projekte und hier vorgeschlagene Maßnahmen adressiert. Alle vorgeschlagenen Maßnahmen haben zum Ziel, unter den in Abschnitt 3.2 genannten Grundsätzen eine Konsolidierung und Fortentwicklung der Datenspeicherdienste insbesondere durch kooperative Dienstleistung bzw. Dienstonutzung zu erzielen.

Die beiden bislang nicht adressierten Dienstkategorien, Datenbanken und Informationsdienste, werden in der Phase II des bwDATA-Umsetzungskonzeptes aufgegriffen.

Mit der Phase II des bwDATA-Umsetzungskonzeptes werden dann Maßnahmen für den Zeitraum 2014-2016 vorgeschlagen, abgesprochen mit Bibliotheken, Hochschulen und weiteren Landeseinrich-

tungen. Die detaillierte Ausgestaltung der Projekte zu großen wissenschaftlichen Datenmengen (4.2.4) sowie zur Archivierung (4.2.3) wird ebenfalls im Rahmen des Umsetzungskonzepts Phase II erarbeitet. Hierzu werden in 2013 zum Ziele einer klar belegbaren Bedarfsanalyse Workshops mit verschiedenen Wissenschaftszweigen und Nutzergruppen durchgeführt. Das Umsetzungskonzept Phase II beschreibt zudem die vorgesehenen Beschaffungen im Zeitraum 2015 bis 2020. Aus heutiger Sicht fallen in diesen Zeitraum die Erneuerung/Erweiterung der Large Scale Data Facility (2015/2016), die Erneuerung bzw. der Ersatz für TSM-Verbünde und Archivierungslösungen sowie die Beschaffungsmaßnahmen, die durch die hier skizzierten Projekte entwickelt werden. Zudem sind Beschaffungen im Rahmen besonders energieeffizienter Datenspeicher anzustreben. Abschließend kann aufgrund der in diesem Konzept erstellen Kostenabschätzung festgehalten werden, dass eine Förderung des Landes von mindestens zwei Millionen Euro p.a. für Datenspeicherdienste notwendig ist bei mindestens gleichbleibendem Engagement der Universitäten im Bereich der Datenspeicherdienste.

7 Referenzen

- [1] Alex Szalay, „Extreme Data-Intensive Computing in Science“, Vortrag im Rahmen des LSDMA-Symposiums, SCC/KIT, 25. September 2012, online via: indico.desy.de/conferenceTimeTable.py?confId=6264#20120925
- [2] Large Scale Data Management and Analysis (LSDMA), <http://www.helmholtz-isdma.de/>
- [3] Arbeitskreis der Leiter der wissenschaftlichen Rechenzentren von Baden-Württemberg (ALWR-BW), Strategiepapier „IT an den Universitäten Baden-Württembergs“, August 2010
- [4] Datenspeicherung im Rahmen eines Landeskonceptes für die Baden-Württembergischen Hochschulen – 2009
- [5] J. Köhler, S. Labitzke, M. Simon, M. Nußbaumer und H. Hartenstein, „FACIUS: An Easy-to-Deploy SAML-based Approach to Federate Non Web-Based Services,“ Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2012.
- [6] European Commission: Technology-Scalable Datacenters, 2011, online via: <http://cordis.europa.eu/fp7/ict/computing/documents/nextgen-dc-final.pdf>
- [7] Arbeitskreis der Leiter der wissenschaftlichen Rechenzentren von Baden-Württemberg (ALWR-BW), Umsetzungskonzept der Universitäten des Landes Baden-Württemberg für das Hochleistungsrechnen, überarbeitete Fassung vom 05.09.2012
- [8] Deutsche Forschungsgemeinschaft: Sicherung guter wissenschaftlicher Praxis, 1998