One-stop publishing and archiving: Forschungsdaten für Promotionsvorhaben über Repositorien publizieren und archivieren: Eine landesweite Initiative im Rahmen des Projekts bwDataDiss am Beispiel des Karlsruher Instituts für Technologie (KIT)

Tobias Kurze[1], Regine Tobias[2], Matthias Bonn[3]

1,2 Bibliothek, KIT
2 Steinbuch Centre for Computing, KIT

Abstract. Nowadays research relies more and more on data to achieve progress in various scientific domains. To understand and to be able to reproduce results, it is essential that the underlying research data is available to scientists - even after a relatively long time.
bwDataDiss is an effort to provide infrastructure and services for a very specific group of researchers – namely PhD students – to enable them to store and archive their research data and also to make it available to other researchers.

Schlagwörter. Dissertation, Forschungsdaten, Langzeitarchivierung, OpenAccess

Einführung – Hintergründe für das Projekt bwDataDiss

bwDataDiss ist als dreijähriges Projekt, finanziert durch das MWK Baden-Württemberg gestartet. Das Projekt verbindet die Erfahrungen der Bibliothekswelt im Umgang mit Nutzern mit der Expertise der Rechenzentren im Bereich des Aufbaus und Umgang mit großen Speicherinfrastrukturen. Daher sind im Projekt jeweils die Bibliotheken und die Rechenzentren der Universität Freiburg und des Karlsruher Institut für Technologie beteiligt. Die Hauptmotivation für das Projekt bestand in der Errichtung einer Infrastruktur für die Veröffentlichung und den Erhalt von Forschungsdaten sowie die Bereitstellung dieser Infrastruktur in Baden-Württemberg. Durch bwDataDiss soll es auch kleineren Einrichtungen ohne große lokale Rechenzentren vor Ort möglich sein, auf einfachem und schnellem Wege die Repositorien der Bibliotheken um neue Forschungsdatenservices zu erweitern.

Und die Erweiterung der Services für dieses Anwendungsfeld ist auch dringend nötig, denn die Diskussion um die Veröffentlichung und den Erhalt von Forschungsdaten ist derzeit in der Wissenschaftswelt ein großes Thema[1]. In der Denkschrift der DFG zur „Sicherung guter Wissenschaftlicher Praxis“ wird darauf hingewiesen, dass „Primärdaten als Grundlagen für Veröffentli-

1 Siehe zum Beispiel: Rat für Informationsinfrastrukturen: Leistung aus Vielfalt. Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen, 2016.

chungen (…) auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, zehn Jahre lang aufbewahrt werden (sollen).“[2]

Inzwischen ist in der Wissenschaftsgemeinschaft die Bestätigung, dass zugrundeliegende Forschungsdaten für die Überprüfbarkeit von Forschungsergebnissen häufig unverzichtbar sind, fast ein Allgemeinplatz. An vielen Orten formieren sich daher entsprechende Policies an den Hochschulen, die erste Schritte und Herausforderungen im Umgang mit der Thematik regeln.[3] Gleichzeitig wächst der Bedarf an entsprechender Infrastruktur, diese Forschungsdaten langfristig zu archivieren und den Zugang zu ihnen zu ermöglichen. Trotz vielen Diskussionen und Initiativen in den letzten Jahren besteht aber derzeit noch eine große Lücke bei der tatsächlich bereitgestellten Infrastruktur der Bibliotheken. Forschungsdaten sind, wenn überhaupt, in vielen Fällen in disziplinären Fachrepositorien abgelegt, denen es vielfach an nachhaltigen Betriebskonzepten mangelt und die in punkto Verbindlichkeit und Standardisierung häufig erst noch Neuland betreten.[4]

Angesichts der großen Herausforderungen des neuen Serviceumfelds wollte das Wissenschaftsministerium in Baden-Württemberg für bwDataDiss einen konkreten Rahmen spannen, so dass die Umsetzungserfolge für Infrastrukturanbieter leichter zu erreichen sind. Daher knüpfte man an bestehende Workflows und Vorarbeiten an. Ganz konkret handelt es sich hier um Forschungsservices für Nachwuchswissenschaftler: Denn zum einen entstehen im Rahmen von Forschungsprojekten und im Speziellen bei Doktorarbeiten häufig Forschungsdaten und zum anderen spielen Bibliotheken traditionell eine Schlüsselrolle im Dissertationsprozess: Die Pflichtabgabe zur Erlangung des Doktortitels erfolgt an allen Hochschulstandorten anhand der Veröffentlichung über die zugehörige Bibliothek. Die zugrundeliegenden Workflows sind also bereits existent und wurden von der analogen, auf gedruckten Exemplaren basierenden Abgabe in den letzten Jahren annähernd vollständig auf digitale Veröffentlichungsprozesse transformiert.

Allerdings unterscheiden sich eben diese Prozesse von Bibliothek zu Bibliothek im Detail und können auch relativ komplex sein. Um diese Komplexität und Diversität ein Stück weit abzubilden, bringen sowohl die Universitätsbibliothek Freiburg, als auch die KIT-Bibliothek ihr Wissen um promotionsbezogene Publikationsworkflows in bwDataDiss ein.

Und noch ein weiterer Anknüpfungspunkt war für das Projekt relevant: Bibliotheken haben üblicherweise weder die Möglichkeit, große Datenmengen zu speichern, noch Erfahrungen darin, diese Daten auf Langzeitverfügbarkeit zu analysieren. Daher fließt in bwDataDiss die Expertise von zwei großen Rechenzentren mit ein: Das SCC am KIT stellt die IT Infrastruktur und die Systeme zur Verfügung, die es ermöglichen, Forschungsdaten zu speichern und zu archivieren. Das Rechenzentrum der Universität Freiburg liefert Werkzeuge, um Daten auf Archivierbarkeit und Langzeitverfügbarkeit zu untersuchen – in bwDataDiss auch als Charakterisierung bezeichnet.

2 DFG – Deutsche Forschungsgemeinschaft: Sicherung guter wissenschaftlicher Praxis. Denkschrift, Empfehlungen der Kommission zur Selbstkontrolle in der Wissenschaft, Bonn, 2013: http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_1310.pdf, S. 21 f. : zuletzt geprüft am 24.2.2017.

3 Zum Beispiel die Forschungsdaten-Policy am KIT vom 17.10.2016. http://www.rdm.kit.edu/downloads/FDM-Policy_final.pdf, zuletzt geprüft am 24.2.2017.

4 Eine gute Übersicht über Forschungsdatenrepositorien gibt re3data: http://www.re3data.org/.

Aktueller Stand im Projekt

Im letzten Jahr der Projektförderung präsentiert sich nun bwDataDiss als ein Dienst der KIT-Bibliothek, der Hochschulen des Landes Baden-Württemberg beim Aufbau einer lokalen Infrastruktur für die Langzeitarchivierung und Bereitstellung von Forschungsdaten von Promovierenden unterstützt und über ein landesweites Portal präsentiert. Als Voraussetzung für die Nutzung der Services von bwDataDiss muss zwischen der jeweiligen Bibliothek (bzw. Universität) und dem KIT ein Vertrag abgeschlossen werden. Promovierende, die bei einer teilnehmenden Bibliothek ihre Dissertation abgeben, können dann zusätzlich Forschungsdaten über bwDataDiss archivieren und publizieren. BwDataDiss unterstützt alle Disziplinen und Datentypen. An die Abgabe der Forschungsdaten wird lediglich die Bedingung geknüpft, dass diese einen finalen, für die Nachnutzung aufbereiteten Charakter innehaben. Der genaue Umfang des Services der Bibliothek und die endgültige Auswahl der Forschungsdaten liegt im Ermessensspielraum der zuständigen Bibliothek. Der Dienst ist eng mit den jeweiligen Repositorien vor Ort verbunden und kann auch auf weitere Publikationstypen ausgeweitet werden, die unabhängig von der Dissertationsabgabe fungieren. Im Zentrum steht, dass der Dienst vollständig in lokale Workflows der Bibliothek integriert werden kann.

Es haben sich drei Modelle herauskristallisiert, die die Nutzung des Dienstes sowohl für große Bibliotheken mit einer leistungsfähigen IT-Infrastruktur als auch für kleinere Bibliotheken attraktiv macht. Die Modelle unterscheiden sich in erster Linie anhand der Integrationstiefe der bwDataDiss-Infrastruktur in die vorhandenen Workflows und sind in Abbildung 1 dargestellt.

Abbildung 1. Integrationsmodelle von bwDataDiss

Der Projektpartner Universitätsbibliothek Freiburg folgt Modell 1, das es ermöglicht, die Workflows vollständig in FreiDok *plus*[5], dem institutionellen Repository der Universitätsbibliothek Freiburg, abzubilden. In den letzten Jahren wurde es zu einem Forschungsdateninformationssystem weiterentwickelt, um die komplette Forschungslandschaft der Universität abdecken zu können. Vor diesem Hintergrund ermöglicht FreiDok *plus* auch die Veröffentlichung und Archivierung von Forschungsdaten in einem Guss und stellt entsprechende Workflows bereit. Der Integrationsaufwand war dementsprechend hoch.

Einen anderen Ansatz verfolgte die KIT-Bibliothek mit dem zentralen Repository KITopen[6], die die Mehrwerte von bwDataDiss weniger nachprogrammieren, als direkt mit dem Repository verbinden möchte. Der Fokus der IT-Anbindung liegt darauf, zwar eine möglichst einheitliche

5 https://freidok.uni-freiburg.de/
6 https://www.bibliothek.kit.edu/cms/kitopen.php

Nutzerkommunikation anzustreben, aber bei den konkreten Anwendungen auf die Features von bwDataDiss zu verweisen. Ganz ohne Nutzerbrüche kommt der Forschungsdatenworkflow so nicht aus, aber der Integrationsaufwand wird um beträchtliche Komponenten verringert.

Das Projekt bwDataDiss konnte im Verlauf des Projekts noch ein weiteres, drittes Modell entwickeln, das dem Wunsch des Ministeriums nach Nachnutzbarkeit voll entspricht: Durch die komplexe Portalentwicklung von bwDataDiss kann eine Bibliothek auch auf relativ schnellem Wege zur Einführung dieses neuen Forschungsdatendienstes kommen, indem das Repository und bwDataDiss IT-technisch völlig getrennt voneinander betrieben werden. Dieser Ansatz kommt in dieser Form auch kleineren Bibliotheken zu Gute (Modell 3).

Im Folgenden werden die technischen Grundlagen von bwDataDiss erläutert. Die Ausführungen werden schwerpunktmäßig anhand des Modells 2 erläutert, gehen aber in den Details auch auf die Spezifikationen für Modell 1 und Modell 3 ein bzw. sind auch für diese gültig.

Konzepte und Workflows

Prinzipien

Beim Entwurf von bwDataDiss wurden folgende Prinzipien berücksichtigt:

- Möglichst einfach nutzbar für die Hauptkunden, nämlich Promovenden bzw. Forschende im Allgemeinen
- Datenintegrität jederzeit sicherstellen
- Flexible Möglichkeiten der Integration in Bibliothekssysteme

Aufgabenteilung

Zwischen der Bibliothek und bwDataDiss gibt es eine strikte Aufgabentrennung:
Bibliotheken:

- Ansprechpartner für Forschende und Promovierende
- Erfassung und Kontrolle für Metadaten

bwDataDiss:

- Archivierung von Daten
- Bereitstellung von archivierten Daten zur Nachnutzung durch die Wissenschaftsgemeinschaft
- Charakterisierung der Forschungsdaten und Bereitstellung derer Ergebnisse

Da Bibliotheken über ihre Repositorien eine zentrale Rolle im institutionellen Publikationsprozess spielen, ist es konsequent, die Abgabe von Forschungsdaten – die als Teil der Dissertation oder als Teil von weiteren Publikationen entstanden sein können – in eben diesen Prozess zu integrieren. Der Workflow besteht im Groben aus den folgenden Schritten (Reihenfolge vernachlässigbar):

1. Der Forscher überträgt seine Publikation (Bibliothekswebseite)

2. Der Forscher stellt Metadaten für die Publikation bereit (Bibliothekswebseite)
3. Die Bibliothek prüft die Publikation und die bereitgestellten Metadaten und kontaktiert ggf. den Forschenden

Bei Forschungsdaten sind mindestens zwei zusätzliche Schritte nötig:

- Übertragung der Forschungsdaten (zu bwDataDiss oder zur Bibliothek, dies hängt vom konkreten Integrationsszenario ab)
- Zu den Forschungsdaten gehörende Metadaten bereitstellen (zu bwDataDiss oder zur Bibliothek)

Möglichkeiten der Integration einzelner Komponenten

Upload

bwDataDiss unterstützt mehrere Wege der Integration in die Bibliothekssysteme, um Forschungsdaten vom Promovenden zum Archiv zu transferieren. Entweder werden die Daten direkt vom Promovenden zu bwDataDiss (und dann weiter ins Archiv) übertragen, oder die Daten werden temporär auf Servern der Bibliothek zwischengespeichert und später von dort zu bwDataDiss übertragen (Modell 1). In jedem Fall aber prüft die Bibliothek die bereitgestellten Metadaten.

Details des direkten Datentransfers: Es gibt zwei Möglichkeiten, den direkten Datentransfer vom Promovenden zu bwDataDiss zu organisieren, wobei die Bibliothek entscheidet, welche Lösung umgesetzt wird.

Für eine einheitlichere Sicht auf die Bibliotheksseite (Modell2) kann die Uploadkomponente von bwDataDiss in die Bibliothekswebseite integriert werden. Obwohl die Komponente auf der Bibliotheksseite integriert ist, überträgt diese die Daten direkt an bwDataDiss.

Die andere Möglichkeit (Modell 3) besteht darin, den Nutzer zum bwDataDiss Portal weiterzuleiten und dort die Daten hochzuladen. Allerdings stellt dies für den Nutzer natürlich einen Bruch dar.

In Abschnitt Upload Komponente finden sich weitere Details bezüglich des Uploaders.

Metadata

BwDataDiss bietet verschiedene Möglichkeiten, um Metadaten entgegenzunehmen. Die einfachste Methode besteht im Ausfüllen eines Webformulars auf dem bwDataDiss Portal. Dies setzt aber offensichtlich voraus, dass der Nutzer bwDataDiss ansteuert und damit die Bibliothekswebseiten verlässt (Modell 3).

Um bwDataDiss unsichtbar im Hintergrund zu halten, können die Metadaten auch entweder über eine Schnittstelle von der Bibliothek zu bwDatDiss übermittelt oder aber von bwDataDiss per OAI-PMH abgerufen werden.

Weitere Details zu Medataten finden sich im entsprechenden Abschnitt.

Umsetzung

bwDataDiss wurde in PHP mit Hilfe von Symfony – einem Web Application Framework – implementiert. Außerdem wird JavaScript für einige Funktionen, wie z. B. den zuverlässigen Upload sehr großer Dateien, genutzt.

Es werden drei Benutzerbasisrollen unterschieden: Bibliotheksnutzer, bwDD-Administratorrolle und reguläre Nutzer. Ein Nutzer, der über die bwDD-Administratorrolle verfügt, kann anderen Nutzern weitere Rollen zuweisen.

Web-Frontend

Im Prinzip ist bwDataDiss als Selbstbedienungsportal entworfen, allerdings kann eine Bibliothek die Nutzung auf manche Funktionen beschränken. Es ist erreichbar unter: https://bwdatadiss.kit.edu. Das gesamte Portal ist in Englischer und Deutscher Sprache verfügbar.

API und Datei Management

bwDataDiss stellt eine moderne ReST Schnittstelle (API) bereit, die sowohl xml als auch json Antworten liefern kann und über eine Schlüssel-basierte Authentifizierung verfügt. Die API stellt Funktionen bereit, um Datensätze zu erstellen, Metadaten zu bearbeiten, Dateien in Stücken hochzuladen, Datensätze nach Zustand abzurufen, etc. Außerdem können darüber Archivierungsaufgaben angezeigt und deren erfolgreiche Durchführung gemeldet werden. Ein Großteil der über das Portal bereitgestellten Funktionalität lässt sich auch über die API nutzen.

Upload- Komponente

bwDataDiss stellt ein mächtiges Upload-Werkzeug bereit. Wie im Abschnitt Prinzipien geschildert, ist Datenintegrität eines der wichtigsten Entwurfsziele von bwDataDiss und muss auch durch den Uploader sichergestellt werden. Der Uploader transferiert die Daten vom Promovenden zu bwDataDiss und ist fähig, mit Dateien beliebiger Größe umzugehen. Der Upload kann außerdem unterbrochen und zu einem späteren Zeitpunkt fortgesetzt werden. Weiterhin werden vor und nach dem Upload Prüfsummen berechnet und verglichen, um Integrität sicherzustellen. Die Funktionsweise des Uploaders ist in Abbildung 2 dargestellt.

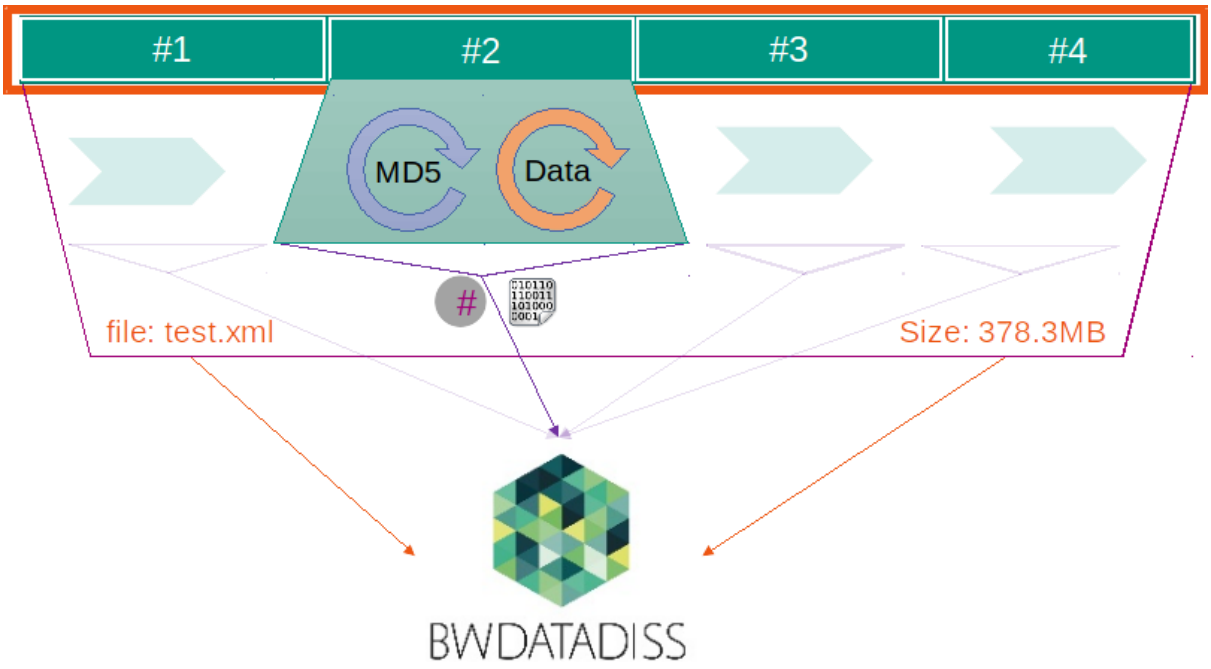

Abbildung 2. bwDataDiss Uploader: Funktionsweise

Beim Upload einer Datei wird diese in Stücke einer festen Größe aufgeteilt und dann separat verarbeitet: zunächst wird eine Prüfsumme berechnet, dann werden die Daten des Dateiausschnittes, zusammen mit der Prüfsumme zu bwDataDiss übertragen. Ein Beispiel ist in Abbildung 2 dargestellt. Der Uploader ist in JavaScript geschrieben und kann mehrere Threads nutzen, um Dateien parallel zu bearbeiten.

Benutzerauthentifizierung und bwIDM

Um Benutzer zu identifizieren setzt bwDataDiss auf einen anderen Dienst, nämlich das „Föderierte Identitätsmanagement der baden-württembergischen Hochschulen (bwIDM)“. bwIDM ermöglicht Nutzern von Universitäten und Hochschulen in Baden-Württemberg. sich per Shibboleth (SAML-basiertes Single Sign On) zu authentifizieren. Für Benutzer, die keiner Universität (mehr) angehören – und sich somit nicht per bwIDM anmelden können – wurde eine Einladungsfunktion geschaffen. Solch eine Einladung kann von der jeweiligen Bibliothek ausgelöst werden, um externe Nutzer zu bwDataDiss einzuladen.

Authentifizierung und Integration durch Bibliotheken

bwDataDiss stellt eine Reihe von Diensten bereit, die durch Bibliotheken genutzt werden können. Das einfachste Integrationsszenario sieht dabei schlicht die Archivierung von Daten und das Bereitstellen von Metadaten vor.

Wie im Abschnitt Aufgabenteilung dargestellt, ist die Bibliothek der Hauptansprechpartner von Promovenden und der Ausgangspunkt für den Veröffentlichungsprozess der Dissertation.

Nun hat eine Bibliothek mehrere Möglichkeiten, die Übertragung der Forschungsdaten zu organisieren: Die Forschungsdaten werden vom Promovenden zur Bibliothek übertragen und später dann von dieser (unter Nutzung der API) weiter zu bwDataDiss. Dies erfordert aber natürlich, dass die Bibliothek entsprechende Systeme vorhält.

Eine andere Möglichkeit besteht darin, den Uploader von bwDataDiss in die Webseite der Bibliothek zu integrieren und die Daten direkt vom Promovenden zu bwDataDiss zu übertragen.

Um authentifizierte Benutzer-Kontexte von der Bibliothekswebseite zum API-Key basierten bwDataDiss Backend delegieren zu können, authentifiziert eine Bibliothek einen Nutzer zunächst per Shibboleth WebSSO. Dann kann sich die Bibliothek selbst mit dem HMAC-SHA256 signierten Benutzer-Shibboleth Token authentifizieren und von der bwDataDiss API einen API-Key für den Benutzer erhalten. Auf diese Art können auch asynchrone API-Anfragen von der Bibliothek an bwDataDiss gestellt werden, ohne den initial hergestellten Nutzerkontext zu verlieren.

Daten Charakterisierung

Datenformate variieren stark zwischen verschiedenen Forschungsdisziplinen und da bwDataDiss nicht auf Forschungsdaten bestimmter Disziplinen beschränkt ist, besteht trotzdem Bedarf an qualitätssichernden und die Langzeitarchivierbarkeit vorbereitenden Maßnahmen. Insbesondere die Langzeitarchivierbarkeit ist in der heutigen Zeit nicht einfach zu erreichen – da schon im geplanten Zeitraum von 10 Jahren Formate unlesbar werden könnten. Um dieses Risiko besser einschätzen zu können, müssen die Formate der gespeicherten Daten bewertet und hinterlegt werden. Für die hinterlegten Daten können dann ggf. entsprechende Erhaltungsstrategien entwickelt werden. Für diese Analyse der Formate und deren Bewertung wurde im Rahmen von bwDataDiss ein entsprechender Dienst entwickelt und integriert.

Erhaltungsrisiken

Das Ziel der Charakterisierung ist es, eine Übersicht über mögliche Erhaltungsrisiken bezüglich Nachnutzbarkeit der Daten bereitzustellen. Dafür muss die logische und strukturelle Repräsentation der Daten – das Dateiformat – bewertet werden. Dies gilt insbesondere wenn eine Dokumentation und Software zu den Dateiformaten existiert. Anhand dieser Informationen können dann Vorhersagen bzgl. langfristiger Nutzbarkeit erstellt werden.

Die Ergebnisse des Charakterisierungsdienstes können für zwei Zwecke genutzt werden: Zum einen als Werkzeug, um Daten vor der Archivierung zu bewerten und Rückmeldung bzgl. Datenformaten zu geben. Anhand dieser Rückmeldung können Wissenschaftler Empfehlungen für Dateiformate gegeben und deren Aufmerksamkeit bzgl. geeigneter Dateiformate gesteigert werden. Zum anderen können die Charakterisierungsergebnisse genutzt werden, um eine Softwaresammlung zu pflegen, die benötigt wird, um mit entsprechenden Daten zu arbeiten bzw. eine Emulationsumgebung bereitzustellen.

Daten Charakterisierung

bwDataDiss stellt einen RESTful Charakterisierungsdienst zur Analyse von Daten bereit. Es wurde FITS als Werkzeug zur Analyse der Dateien ausgewählt, da es verschiedene Charakterisierungswerkzeuge in einem einzelnen, anpassbaren Java-Framework bündelt.

Eine Charakterisierungsanfrage kann durch Stellen einer POST-Anfrage generiert werden, die sowohl eine Referenz zu einem bwDataDiss Datensatz, als auch – optional – auf eine Policy enthält. Aus Effizienzgründen werden die Daten in ein ISO9660 Container (CD-ROM / DVD format) gebündelt. Damit ist ein Vorab-Download der Daten überflüssig, da der Container aus der Ferne eingehängt werden kann und nur Daten, die für die Charakterisierung benötigt sind, übertragen werden. Für die HTTP Anfragen werden entsprechend Range-Requests genutzt. Da die Daten nur im Speicher gehalten werden, stellen parallele Anfragen und Festplattenplatz kein Problem dar.

Beispielanfrage: http://bwdatadiss.eaas.uni-freiburg.de:8080/bwdatadiss/FileFmtCheck/init

```
{
   "objectUrl": "http://bwdatadiss/myset.iso",
   "policyUrl": "http://bwdatadiss/base-policy.txt"
}
```

Der Dienst gibt sofort eine Session ID zurück, welche für Abfragen bzgl. des Charakterisierungsstatus genutzt werden kann. Abhängig vom Umfang der Daten kann es eine Weile dauern bevor die Charakterisierung abgeschlossen ist. Mithilfe der Session ID können die Ergebnisse abgerufen werden – liegen diese noch nicht vor, muss der Aufruf wiederholt werden.

Das bwDataDiss Charakterisierungsergebnis ist eine Dateiformatverteilung bzw. die Anzahl der Dateien pro Dateityp (PRONOM ID).

Beispiel: http://132.230.3.211:8080/bwdatadiss/FileFmtCheck/getResultSummary?sessId=5

```
{
   "summary": [
   {
     "type": "x-fmt/111",
     "value": "GREEN",
     "count": "242"
   },
   {
     "type": "fmt/16",
     "value": "GREEN",
     "count": "2"
   },
   {
     "type": "x-fmt/411",
     "value": "RED",
     "count": "1"
   }
}
```

Es kann auch ein detailliertes Ergebnis angefordert werden, welches eine Liste von Dateien (inkl. relativem Pfad) zu jeder PRONOM ID enthält. Wenn zusätzlich eine Policy angegeben wurde, wird zu jedem Format eine „Bewertung“ angehängt. Im obigen Beispiel enthält die Policy Ampelfarben, wobei den PRONOM IDs „x-fmt/111“ (Plain Text) und „fmt/16“ (PDF) die Farbe Grün und „x-fmt/411(Windows Executable COFF) die Farbe Rot zugewiesen wird.

Archivintegration

Das Archiv basiert auf dem High Performance Storage System (HPSS) und stellt ein hierarchisches Dateisystem zur Verfügung, auf welches per SFTP zugegriffen werden kann. HPSS verfügt über einen integrierten, sehr großen Festplatten Cache, der (logisch) oberhalb des eigentlichen Bandarchivs angeordnet ist. Die Integration des Archivs in bwDataDiss wird realisiert durch zwei verschiedene Techniken:

1. Einen FUSE basierten SSHFS Mount Punkt in das bwDataDiss Hostsystem, welches das SFTP Protokoll versteckt und ein „normales“ lokales Verzeichnis bereitstellt, auf das mit gewöhnlichen Dateiwerkzeugen zugegriffen werden kann.
2. Einen direkten Zugriff per SFTP-Softwarebibliothek ohne Einbindung in das Dateisystem des Hostbetriebssystems.

In beiden Fällen besteht allerdings eine wichtige Einschränkung: Der Zugriff kann deutlich langsamer sein, als ein lokales Dateisystem bzw. übliche NFS/CIFS Mounts. Die gilt insbesondere dann, wenn die Daten von Archivbändern abgerufen werden müssen und nicht vom Festplattencache kommen. Daher ist das Schreiben in das Archiv üblicherweise ausreichend schnell, das Lesen hingegen – von Dateien die nicht mehr im Cache sind – kann sehr langsam sein. Insbesondere kann es lange dauern, bis ein angeforderter Datenstrom überhaupt Daten liefert. Daher musste eine asynchrone, entkoppelte Lösung für den Zugriff auf die Daten im Archiv entwickelt werden.

Die Entkopplung wurde realisiert, indem die Archivintegration vom Apache bzw. PHP basierten Web-Dienst getrennt wurde. Um einen asynchrone Archivanbindung zu realisieren, wurde ein separater Hintergrundworker implementiert, der die bwDataDiss REST API nach Archivierungsaufgaben (schreibe Daten ins Archiv, lese Daten aus dem Archiv) abfragt und das interne bwDataDiss Datenmodell auf das SFTP-Dateisystem abbildet (Abb. 3). Dabei können die beiden o. g. Varianten (SSHFS/FUSE-Dateisystem bzw. SFTP-Direktzugriff) unabhängig voneinander für verschiedene Archivoperationen (lesen/schreiben) und verschiedene Zugangspunkte genutzt werden. Dies ermöglicht die Anbindung mehrerer Archive gleichzeitig, es ist lediglich notwendig, dass sie direkt per SFTP oder einem anderen, ins lokale Dateisystem integrierbaren Protokoll ansprechbar sind.

Zusätzlich wurde in das bwDataDiss Hostsystem ein schneller CIFS-Kurzzeitspeicher eingebunden, um hochgeladene bzw. zum Download bereitgestellte Datensätze für den Webserver performant zugreifbar zu halten.

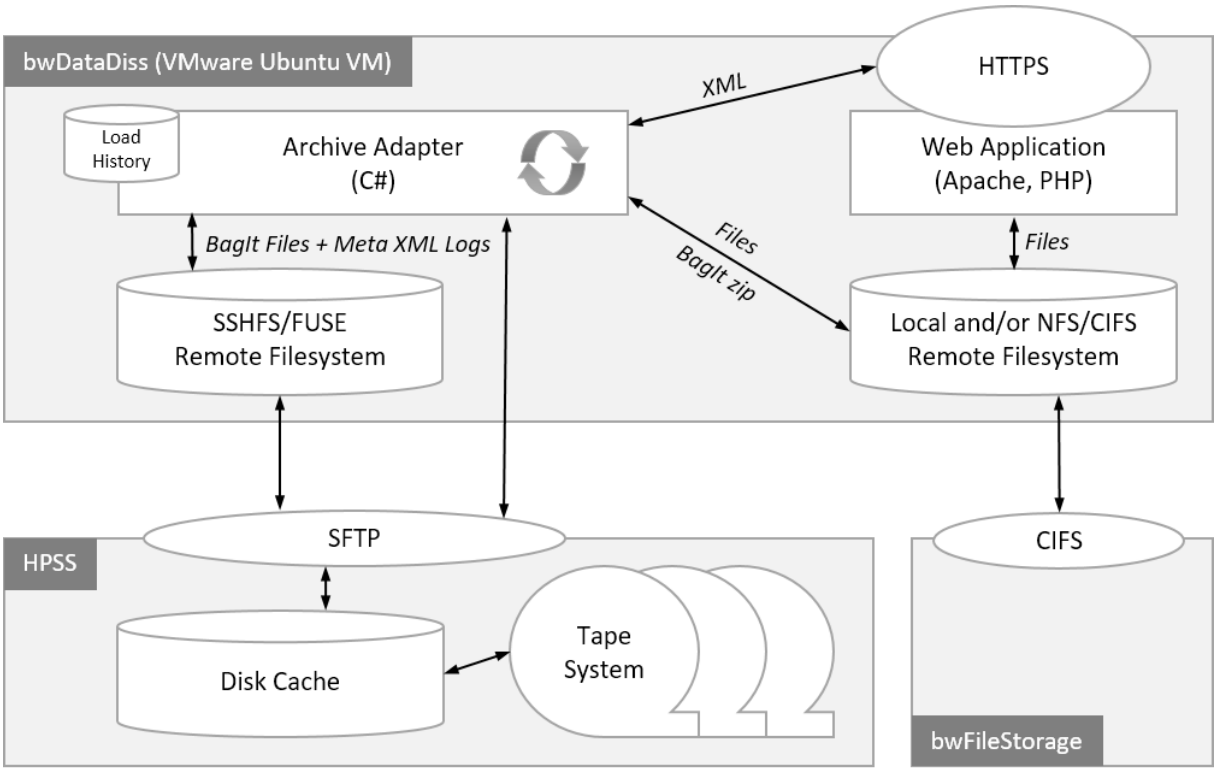

Abbildung 3. bwDataDiss und HPSS

Dort wird ein unkomprimiertes BagIt Verzeichnis angelegt, in welches die lokalen Dateien kopiert werden. Die Metadaten und das Transfer-Log werden ebenfalls in das Archiv geschrieben, was zu einer selbstbeschreibenden Archivdateisystemstruktur (gruppiert nach Bibliotheken) führt. Im Namensschema (Abbildung 4) sind sowohl der [Library-Name] und die [DataSet-ID] (in bwDataDiss) eindeutig, und die [User-EPPN] ist zumindest bei der jeweils zuständigen Bibliothek eindeutig.

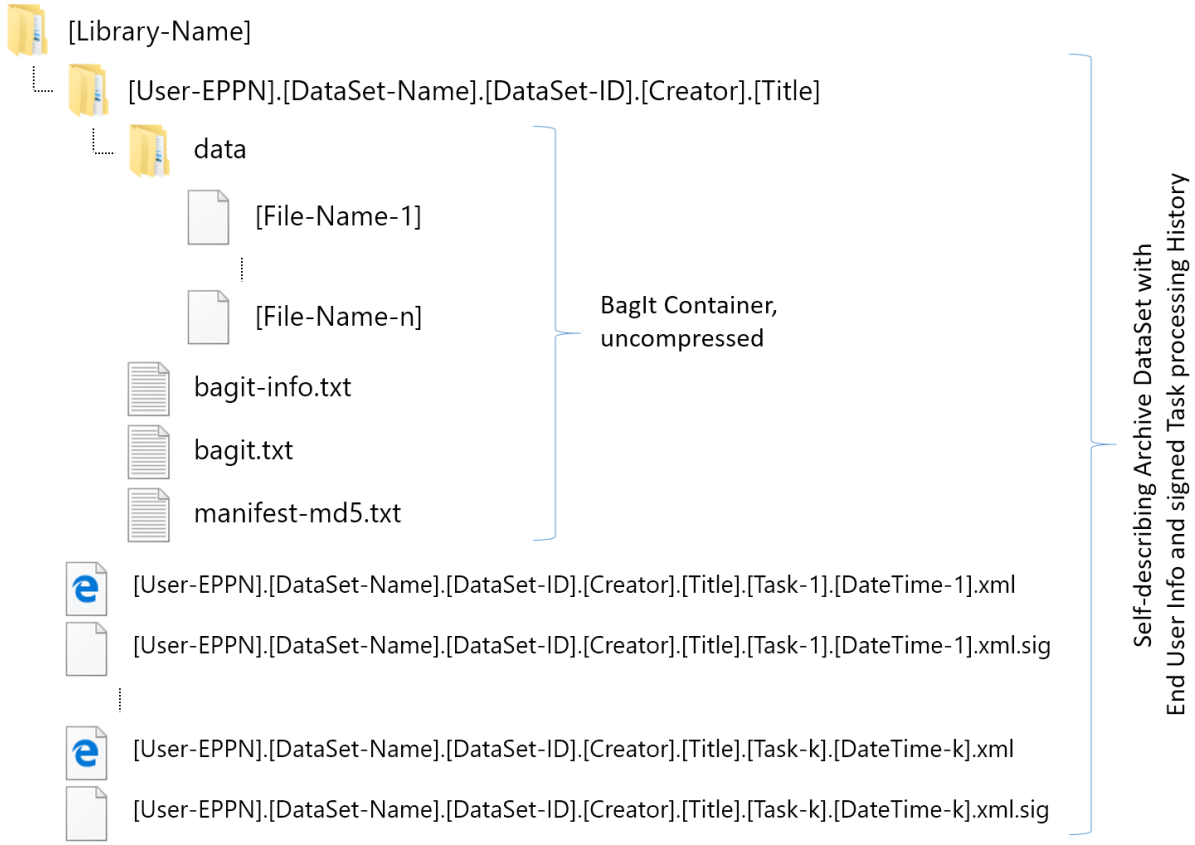

Abbildung 4. bwDataDiss Archivnamensschema

Um einen archivierten Datensatz zu lesen, wird das komplette BagIt Verzeichnis asynchron auf den lokalen Dateisystemcache kopiert, von wo aus der Webserver direkten Zugriff auf die Dateien hat (Abb. 3, CIFS-Mount von bwFileStorage). Optional kann auch eine Zip Datei vom Archivworker angefordert werden, die alle Dateien inklusive der BagIt Dateien enthält. Alle Archivierungsaufgaben werden mit Prüfsummen gegengeprüft und bei Erfolg der bwDataDiss API entsprechend mitgeteilt. Sollte ein Fehler auftreten, wird die Aufgabe wiederholt.

Metadata

bwDataDiss kann durch alle Hochschulbibliotheken im Land Baden-Württemberg genutzt werden und ist auf keine Fachdisziplin festgelegt. Das Metadatenschema von bwDataDiss trägt dem Rechnung und ist entsprechend generisch angelegt. Das Schema ist außerdem an den RADAR[7] Metadaten Kernel angelehnt.

7 Das Projekt RADAR oder Research Data Repositorium stellt Infrastruktur für das Forschungsdatenmana gement bereit und ist ein gemeinsames Projekt des FIZ Karlsruhe, des Steinbuch Centre for Computing (SCC), der Ludwig-Maximilians-Universität München (LMU) and der Technischen Informationsbibliothek (TIB).

Das Metadatenschema besteht aus nachfolgenden Punkten:

- Titel*
- Zusatztitel
- Ersteller*
- Beitragende
- Abstrakt*
- Schlagwörter
- Liesmich*
- Erstellungsjahr (Ende)*
- Creation Year (Beginn)
- Herausgeber*
- Jahr der Veröffentlichung*
- Klassifizierungen*
- Ressourcentyp
- Lizenz*
- Rechteinhaber
- Embargodatum
- Zusätzliche Metadaten

Punkte, die mit * markiert sind, sind Pflichtfelder und müssen bei bwDataDiss angegeben werde. Ein paar der Punkte bedürfen einer Erklärung:

Liesmich: Im Speziellen für Forschungsdaten ist es wichtig zu wissen, wie die bereitgestellten Forschungsdaten organisiert sind und ggf. genutzt werden können. Auch technische Informationen oder Hinweise zur genutzten bzw. der zu nutzenden Software können hier ihren Platz finden.

Klassifizierungen: bwDataDiss können beliebige Klassifizierungen übergeben werden – mindestens jedoch eine. Über das Portal werden Auswahlhilfen für die DDC und für eine Klassifizierung nach DGF Fachgruppen angeboten.

Resourcentyp: Vom RADAR Metadaten Kernel entliehen, kann es einen der nachfolgenden Werte, der den Typ der Forschungsdaten beschreibt, annehmen: audiovisual, collection, dataset, image, model, software, sound, text, workflow, other. Diese Beschreibung der Art der Forschungsdaten ist relevant für die Aufnahme in den Data Citation Index von Thomson Reuters /Clarivate Analytics.

Lizenz: bwDataDiss steht hinter Open Access und stellt daher alle möglichen CC Lizenzen zur Auswahl bereit (z.B.: CC-BY und CC-BY-SA). Allerdings können Situationen eintreten, wenn diese Lizenzen nicht ausreichend sind und erlauben es Bibliotheken daher, eigene Lizenzen zu hinterlegen.

Embargodatum: Unter Umständen ist es nötig, den Zugriff aus Forschungsdaten zu unterbinden. Dies kann durch die Einrichtung eines Embargos erreicht werden. Während des Embargos ist es möglich, einzelnen Nutzern trotzdem Zugriff auf die Daten zu gewähren.

Mit bwDataDiss in wenigen Schritten zum neuen Service der Forschungsdaten-veröffentlichung – am Beispiel KITopen

Wahl des Modells der Integration

bwDataDiss ist an der KIT-Bibliothek seit Anfang März im Produktivbetrieb. Damit ist es nun am Campus des KIT erstmals möglich, Forschungsdaten aller Disziplinen und Formate gemeinsam mit der Promotionsschrift kostenlos zu veröffentlichen. Neben der technischen Integration des Dienstes bwDataDiss mit dem Repository der KIT-Bibliothek waren dazu auch noch weitere organisatorische Schritte erforderlich. Das Repository KITopen ist Teil eines am KIT entstehenden Forschungsinformationssystems und befindet sich in einem großen technischen Umbruch. Schwerpunkte sind neben der Veröffentlichung von Volltexten und bibliographischen Daten von KIT-Wissenschaftlern die Bedienung der unterschiedlichen Berichtspflichten das KIT und insbesondere der Helmholtz-Gemeinschaft. Aus praktischen Gründen war daher die Abgabe der Promotionen über die KIT-Bibliothek nur in Form eines gesonderten elektronischen Formulars möglich. Die eigentliche Datenerfassung erfolgte über die Mitarbeiter in das Repository. Es galt daher für die Integration von bwDataDiss, diesen Sonderworkflow zunächst in das sonstige Repository-Umfeld einzupflegen. Darauf aufbauend erfolgten dann die weiteren Prozessschritte für die Veröffentlichung der Forschungsdaten.

Ziel war es, den Spagat zu schaffen und den Nutzern keine unnötigen Systembrüche zu verursachen und dennoch das Repository um Komponenten von bwDataDiss zu erweitern, um unnötige Doppelimplementierungsarbeiten zu vermeiden. Daher fiel die Wahl auf Modell 2 das beinhaltet, die Erfassung der Metadaten in die Workflows des Repository der Bibliothek vollständig zu integrieren und auch die Nutzerkommunikation über KITopen zu veranstalten. Für die technischen Spezialanforderungen wie den Upload großer Mengen an Forschungsdaten, die unterschiedliche Uploadzeiten und Speichernutzungskapazitäten mit sich bringen, sollte auf die technische Infrastruktur von bwDataDiss zurück gegriffen werden. In der Praxis des KIT nutzen die Promovierenden daher nun den Upload im Repository , welcher direkt die Schnittstelle von bwDataDiss anspricht. Siehe dazu auch: Upload im Abschnitt: Möglichkeiten der Integration einzelner Komponenten.

bwDataDiss stellt neben Kern- auch Hilfsfunktionen bereit. So sind am KIT ca. 40% aller Promovenden zum Zeitpunkt der Abgabe der Forschungsdaten nicht mehr am KIT beschäftigt bzw. verfügen über keinen Account mehr, mit welchem eine Anmeldung bei bwIDM möglich wäre. Für diese Fälle wird eine Einladungsfunktion von bwDataDiss bereitgestellt, die es ermöglicht, Accounts auf bwDataDiss zu erstellen. Des Weiteren ist es über die bwDataDiss API möglich, bwDataDiss Benutzerkonten und Passwörter zu prüfen und damit diese Accounts zur Autorisierung in anderen Systemen – wie KITopen zu nutzen. Nutzer können sich also auch mit bwDataDiss Accounts bei KITopen anmelden auch wenn eine Anmeldung per bwIDM nicht möglich ist.

Formulierung der Policy für die Nachnutzung von Forschungsdaten

Ein wichtiger Aspekt von bwDataDiss ist die Bereitstellung der Forschungsdaten zur Nachnutzung durch andere Forschende und Interessierte. Die Forschungsdaten sind dazu mit den entsprechenden Metadaten als auch mit persistenten Links und Identifier für die Zitation versehen. Laut Policy von KITopen sind alle Forschungsdaten grundsätzlich unter eine Open-Access-Lizenzen gestellt. Beim Upload können auf Wunsch des Forschenden Embargos eingerichtet werden, die die Nutzung der Forschungsdaten für einen gewissen Zeitraum verhindern. Auf eine Limitierung des Zeitraums wird vorerst verzichtet. Die Embargofrist kann von der Bibliothek gesteuert werden. Innerhalb dieser Frist können ausgewählte Nutzer Zugriff auf die Forschungsdaten erhalten. In der ersten Stufe der Einführung des Dienstes erfolgt das durch die Mitarbeiter des Teams KITopen.

Aufbau eines Services zur Qualitätssicherung und Beratung zur Langzeitarchivierung

Für die inhaltliche Vollständigkeit und Konsistenz der Forschungsdaten sind die Datengeber selbst verantwortlich. Hier muss sich die KIT-Bibliothek erst langsam an den neuen Service herantasten und schrittweise vorgehen. Zunächst liegt daher der Fokus auf der bewährten formalen Prüfung der Metadaten. Dabei wird darauf geachtet, dass die erläuternden Felder wie „Liesmich“ und „Abstract“ bzw. „Schlagwörter“ ihre beschreibenden Funktionen entsprechend erfüllen. Wichtig ist hier eine frühzeitige Rückmeldung an die Forschenden und der Einstieg in die Kommunikation. Darauf aufbauend ist der Aufbau weiterer Beratungsservices für die Langzeitarchivierung angesichts der heterogenen Landschaft der Forschungsdaten ein komplexes und drängendes Feld. BwDataDiss unterstützt diese neuen Qualitätsprüfungsprozesse durch Bibliotheken in Form von Dateitypcharakterisierungen und gibt entsprechende Rückmeldungen an die Datengeber bzw. Bibliotheken. Die Basisinstallation von bwDataDiss verweist dabei auf Empfehlung zu Dateiformaten von der Library of Congress und der Cornell University.[8] Im Fall von KITopen wird zunächst auf die automatisierte Rückmeldung der Charakterisierungsergebnisse an die Nutzer verzichtet, da man zunächst aufgrund der Rückmeldungen Erfahrungen im Umgang mit den Dateiformaten aufbauen möchte. Das dafür nötige Expertenwissen wird in den nächsten Jahren am KIT in verteilten Rollen kooperativ aufgebaut werden, so dass KITopen sich schrittweise erweitern wird. Die technische Implementierung von bwDataDiss erlaubt sogar, daran gemeinschaftlich und im Land verteilt zu arbeiten und Policies gemeinsam zu nutzen.

Rechtliches und Formales

Aufgrund der Projektbeteiligung erübrigt sich eine weitere vertragliche Regelung zwischen der KIT-Bibliothek und bwDataDiss. Für weitere teilnehmenden Bibliotheken ist aber in jedem Fall der Abschluss eines Kooperationsvertrags erforderlich.

Die durchzuführenden Maßnahmen um Datenschutzansprüchen zu genügen, unterscheiden sich je nach Integrationsmodell, können aber auch von Umsetzungsdetails abhängen. Nach unse-

8 https://ecommons.cornell.edu/page/support#format, http://www.digitalpreservation.gov/formats/fdd/browse_list.shtml, zuletzt geprüft am 24.2.2017.

rem Informationsstand ist im Falle von Modell 1 ein Vertrag zur Auftragsdatenvereinbarung zu vereinbaren, im Fall von Modell 2 kann dies auch der Fall sein und im dritten Modell ist ein solcher normalerweise nicht nötig. Allerdings möchten wir darauf hinweisen, dass Sie sich auf jeden Fall an ihren Datenschutzbeauftragten wenden sollten. Die erforderlichen Unterlagen liegen für bwDataDiss vor.

Schlusswort

bwDataDiss ist eine wichtige Initiative auf Landesebene die sich gut in die Dienste einer Bibliothek bzw. der jeweiligen Hochschule einpasst und flexibel integriert werden kann.

Mit den drei Modellen werden sowohl kleinere Einrichtungen, als auch größere, die ggf. einen höheren Integrationsaufwand betreiben können, adressiert.

Durch die Charakterisierung der Forschungsdaten wird es Bibliotheken erleichtert, die Qualität der Forschungsdaten zu sichern bzw. abzuschätzen. Außerdem wird es dadurch möglich, entsprechende Erhaltungsstrategien für Forschungsdaten im Archiv zu erstellen und eine langfristige Nachnutzbarkeit der Daten zu gewährleisten. Dies ist derzeit in der Landschaft von Serviceanbietern für Forschungsdatendienste weitgehend singulär.