

Forschungsdatenmanagement: Abschluss Prototypentwicklung

erstellt am: 21.11.2018	von: Beate Rajski (TUHH) Kai Wörner (UHH)
geändert am: 22.11.2018	von: Beate Rajski (TUHH)
Version: 1.0	
Status: Zur Abnahme	

Übersicht:

Programm:	Hamburg Open Science
Projekt:	Forschungsdatenmanagement
Projektlaufzeit:	2018
M6:	Konzeption und Umsetzung:3. Iteration
Z7:	Abschluss Prototypentwicklung
Auftraggeber:	Prof. Dr.-Ing. Timm-Giel, Vizepräsident Forschung TUHH
Projektleitung:	Beate Rajski, TUHH



Abschluss Prototypentwicklung

Inhalt

Operatives Ziel der PEV	2
Zielerreichung.....	2
Prototypen.....	2
Prototyp V 0.9 mit den wesentlichen Funktionen	3
Prototyp 1.0 mit Anbindung an den Archivspeicher der Hamburger Hochschulen	3
Software-Entscheidung aller beteiligten Institutionen.....	4

Operatives Ziel der PEV

Zur Herbeiführung einer Entscheidung für ein FDM-System pro Institution werden bis Mitte November (M6) zwei Prototypen entwickelt und exemplarisch mit Testdaten befüllt:

1. TUHH: auf Basis von DSpace CRIS (Open Source)
2. UHH: auf Basis von Invenio/Zenodo (Open Source)

Zielerreichung

Für die Umsetzung sollen folgende Bedingungen erfüllt sein:

- Es liegen zwei Prototypen mit Testdaten aus jeweils zwei Einrichtungen vor.
- Der UHH Prototyp ist an den Archivspeicher der Hamburger Hochschulen angeschlossen.
- Die beiden Umsetzungsalternativen werden der Steuerungsgruppe vorgestellt und abgenommen.
- Eine Entscheidung aller beteiligten Institutionen aufgrund der Prototypen und der Modellrechnung für die Kosten für die Einführung eines FDM-Systems liegt vor.

Alle Ziele konnten umgesetzt werden:

Prototypen

Im Projektjahr 2018 wurden zwei Prototypen für ein institutionsspezifisches Forschungsdatenrepositorium im Rahmen von Hamburg Open Science erstellt.

1. TUHH auf Basis von DSpace-CRIS: <https://tor.tuhh.de>
2. UHH auf Basis von Invenio/Zenodo: <https://staging.fdr.uni-hamburg.de/>

Prototyp V 0.9 mit den wesentlichen Funktionen

In der Sitzung der Steuerungsgruppe am 26.9.18 wurden von UHH und TUHH die Prototypen in der Version 0.9 mit den wesentlichen Funktionen für Upload und Bereitstellen von Forschungsdaten vorgestellt und abgenommen.

Funktionalität:

1. Anmeldung / Login
 - Gesicherte institutionelle Anmeldung über Shibboleth
 - TUHH: Anlage und Login mit lokalem Account für Nicht-TU-Angehörige
 - UHH: Für die Evaluierung werden für nicht UHH-Mitglieder vom Administrator lokale Accounts erstellt/genutzt.
2. Upload von Forschungsdaten
 - Upload von Forschungsdaten in den verschiedensten Formaten
 - Setzen einer Embargofrist für Dateien
 - Unterstützung bei der präzisen Beschreibung der Inhalte mit Metadaten
 - Unterstützung bei der Wahl der richtigen Lizenz
 - Verlinkungsmöglichkeit zu Datensätzen und Publikationen, die mit den hochgeladenen Daten in Beziehung stehen.
3. Veröffentlichungsworkflow
 - TUHH: Freischaltung erst nach formaler Kontrolle durch Fachpersonal (konfigurierbar)
 - UHH: Freischaltung durch die Hochladenden
4. ORCID-Verknüpfung
 - UHH: ORCID iD kann manuell eingetragen werden
 - Authentifizierte ORCID iD (Sandbox) und Push-Möglichkeit zu ORCID (Sandbox)
5. Finden und Verlinken
 - Metadaten und Volltexte sind für die Suche verfügbar
 - Für die Datensätze werden persistente Identifier vergeben
6. OAI-PMH-Schnittstelle für das Harvesten ist verfügbar

Beide Prototypen wurden von den beteiligten Einrichtungen getestet und dabei auch mit Daten befüllt und einem Usability Test unterzogen.

Prototyp 1.0 mit Anbindung an den Archivspeicher der Hamburger Hochschulen

Für die Prototypen der Version 1.0 wurde für die UHH die exemplarische Anbindung an den Archivspeicher des Regionalen Rechenzentrums hergestellt. Auch für die TUHH konnte zum 9.11.2018 der Zugang vom Regionalen Rechenzentrum bereitgestellt und TU-seitig angebunden werden und erste Benchmarktests durchgeführt werden.

Test der Anbindung des S3-Langzeitspeichers am RRZ durch die UHH:

Am 19.11. wurden generische Anbindungs- und Lasttests auf dem für die Forschungsdatenrepositorien der UHH und TUHH vorgesehenen S3-Objektspeicher am RRZ (Cloudian HyperStore) durchgeführt. Dazu wurden zufällig generierte Dateien unterschiedlicher Größe sequentiell erzeugt, auf den Objektspeicher gespeichert, von dort wieder heruntergeladen und anschließend gelöscht.

Die Tests wurden einmal vollständig von einem Arbeitsplatzrechner im Universitätsnetz sowie einmal von einem Rechner innerhalb der Servicekette, die das Produktivsystem des UHH-Forschungsdatenrepositoriums beherbergt, durchgeführt. Die detaillierten Ergebnisse der Tests wurden dem RRZ und den teilnehmenden Projekteinrichtungen zur Verfügung gestellt.

Beobachtungen:

Die Ergebnisse lassen sich nicht zu 100% auf den zukünftigen Produktivbetrieb der FD-Repositorien übertragen, da

- die Anfragen nicht aus den tatsächlichen Produktivsystemen erfolgt sind,
- bisher (außer Testanwendungen) keine weiteren Anwendungen auf den Langzeitspeicher zugreifen,
- noch keine aus dem Betrieb hervorgegangenen Leistungsoptimierungen vorgenommen wurden.

Es lässt sich trotzdem bereits jetzt feststellen:

- Die Leistungsfähigkeit des Archivspeichers ist im Augenblick vor allem von der Bandbreite der Anbindung abhängig. Die Tests erfolgten jeweils mit 10 parallelen Anfragen und lasteten die verfügbare Bandbreite bei größeren Dateien beinahe vollständig aus.
- Die Antwortzeiten bleiben auch bei vielen Anfragen kurz. Für das geplante Anwendungsszenario ist von dieser Seite kein Engpass zu erwarten.
- Die Leseoperationen profitieren stark von Caching. Für die meiste Zeit wird das für die Repositorien unerheblich sein, könnte aber bei stark nachgefragten, großen Dateien Vorteile bringen

Fazit:

Die ermittelten Ergebnisse lassen kein Hindernis erkennen, das der Inbetriebnahme der Repositorien mit dem LZS des RRZ als Speicherbackend im Wege steht. Sobald die Produktivsysteme im Einsatz sind, sollen weitere Tests für beide Repositorien durchgeführt werden. Das FD-Repositorium der UHH hat hierfür integrierte Benchmark-Werkzeuge.

Software-Entscheidung aller beteiligten Institutionen

Aufgrund der Prototypen und der zu diesem Zeitpunkt noch vorläufigen Modellrechnung für die Kosten für die Einführung eines FDM-System konnten alle beteiligten Einrichtungen die Entscheidung für den Einsatz einer der beiden Softwarelösungen in 2019 treffen.

DSpace-CRIS:

- Hochschule für Angewandte Wissenschaften Hamburg
- HafenCity Universität Hamburg
- Technische Universität Hamburg

Zenodo/Invenio

- Universität Hamburg
- Universitätsklinikum Hamburg-Eppendorf