Prototypische Implementierung der initialen technischen Workflows (M 2.3.3)

Version 15.03.2014
Cluster 2
Verantwortlicher Partner HKI

DARIAH-DE
Aufbau von Forschungsinfrastrukturen für die e-Humanities

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1110A bis N, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Aufbau von Forschungsinfrastrukturen für die e-Humanities
BMBF Förderkennzeichen: 01UG1110A bis N
Laufzeit: März 2011 bis Februar 2016

Dokumentstatus: Final
Verfügbarkeit: Öffentlich
Autoren: Johanna Puhl, HKI

Revisionsverlauf:

Datum	Autor	Kommentare
20.02.2015	Johanna Puhl	Initiale Version
10.03.2015	Johanna Puhl	Ausarbeitung
12.03.2015	Xi Kong	Kleine Korrekturen
15.03.2015	Johanna Puhl	Finalisierung

Inhaltsverzeichnis:

1. Einleitung

Das Arbeitspaket 2.3 (Technische Aspekte des Data LifeCycle) ist mit der Aufgabe betraut, bereits entwickelte sowie noch im Aufbau befindliche technische Komponenten der DARIAH Infrastruktur aufzugreifen und die theoretischen Überlegungen zu den Funktionen eines geisteswissenschaftlichen Forschungsdatenzyklus mithilfe dieser Komponenten umzusetzen bzw. die so zwangsläufig beobachtbaren Lücken zwischen Theorie und Praxis zu dokumentieren und deren Schließung anzuregen.

Der hier beschriebene Milestone 3.3 beschreibt die erste Implementation des DARIAH Research Data LifeCycle. Das folgende Kapitel enthält die Ergebnisse der AG Research Data LifeCycle und schildert die einzelnen Arbeitsschritte innerhalb eines Forschungsdatenzyklus. Dabei werden eine Reihe fachwissenschaftlicher und generisch technischer Funktionen aufgeführt, welche innerhalb von DARIAH-DE Anwendung finden. Kapitel 2 schildert die initialen technischen Workflows und Kapitel 3 und 4 schildern deren Implementation sowie deren Verankerung im der DARIAH-Infrastruktur.

1.1. Fachwissenschaftliche Funktionen

Unter fachwissenschaftlichen Funktionen werden diejenigen verstanden, welche speziell für die intellektuelle Beschäftigung von Geisteswissenschaftlern mit Forschungsdaten im Fokus stehen. Dadurch, dass es sich um computergestützte Funktionen handelt, können die genannten Funktionen nichtsdestotrotz (teil-) automatisierbar eingesetzt werden.

Erste – nicht repräsentative – Umfragen ergeben hier neben dem Bedarf nach Software, die spezifische Annotationsstandards unterstützt und die Arbeit mit diesen ermöglicht (TEI) sowie dem Bedarf nach geospatialen Visualisierungsfunktionen und Bibliographietools, dass Geisteswissenschaftler nach wie vor hauptsächlich mit Officeprodukten arbeiten[1].

Daneben haben Untersuchungen von quantitativen Methoden (AP 5.2) hinsichtlich ihrer Verwendbarkeit und der Adaption geeigneter Methoden im Rahmen fachwissenschaftlicher Dienste (AP 5.3) in DARIAH-DE bisher eine beschränkte Anzahl von fachlichen UseCases ergeben, die über die Verwendung von Officeprodukten weit hinaus gehen. Der erste dort untersuchte UseCases („Narrative Techniken und Untergattungen im deutschen Roman“) wurde dabei am weitesten ausgearbeitet[2].

Die folgende Liste[3] gibt Aufschluss über die in diesem Dokument geschilderten fachlichen Funktionen, die GeisteswissenschaftlerInnen im Kontext dieser UseCases nutzen. Hier ist insbesondere die Funktionsweise des Tools DKPro[4] aufschlussreich, mithilfe dessen sich auf einem Framework basierend *Pipelines* von Tools und Funktionen orchestrieren lassen:

- Sprachen-Identifizierung
- Wort- und Satzgrenzbestimmung
- Lemmatisierung
- Wortstamm-Bestimmung

[1] Vgl. R 1.2.1 *Report Nutzungsverhalten in den Digital Humanities*
[2] Vgl. R 5.3.1 *Konzept Use Cases*
[3] Ebd. Ab S. 6
[4] Vgl. https://www.ukp.tu-darmstadt.de/research/current-projects/dkpro/

- Wortart-Bestimmung
- Morphologische Analyse
- Named Entity Recognition
- Chunking
- Constituency Parsing
- Dependency Parsing
- Koreferenz-Analyse
- Semantic Role Labeling
- Rechtschreibüberprüfung
- Anaphora Resolution (Z.B. mit Bart[5]:)
- Topic-Modelling gemäß Latent Dirichlet Allocation (LDA) Verfahren (Z.B. mit MALLET[6])
- DARIAH-DE Schema-Registry
- DARIAH-DE Crosswalk-Registry
- DARIAH-DE Generische Registry

1.2. Disziplinunabhängige Funktionen

Zur technischen Realisierung eines geisteswissenschaftlichen Forschungsdatenzyklus besteht neben den fachspezifischen Tools und Funktionen, die in einer prototypischen Implementation unterstützt werden sollen, außerdem der Bedarf nach ganz generischen Funktionen.

Bei diesen handelt es sich also um disziplinunabhängige, in der Tendenz eher administrativ, informatische Funktionen, die in einer Forschungsdateninfrastruktur essentiell notwendig sind.

Ihre Definition und Funktionsweise wurde im Rahmen der AG Research Data LifeCycle in DARIAH-DE genauer erörtert[7]. Die disziplinunabhängigen Funktionen des Forschungsdatenzyklus werden hier nicht im Detail erläutert, aber ihre Umsetzung wird in den folgenden Kapiteln neben den fachwissenschaftlichen Funktionen immer auch mitberücksichtigt.

[5] Vgl. www.bart-anaphora.org
[6] Vgl. http://mallet.cs.umass.edu/
[7] Vgl. dazu die Definition des DARIAH Resarch Dara LifeCycle. Work in Progress: https://docs.google.com/document/d/12tSyZdByWH7I0wb2xGAbh38cw78OezRdjHEGmPIiYIM/edit#

2. Der inititale technische Workflow

Die in R 2.3.1 *Auswahl und Beschreibung der initialen technischen Workflows und Policies für den Data LifeCycle* ausgearbeitete Übersicht alle Einzelfunktionen eines Forschungsdatenzyklus wurde in der AG Research Data LifeCycle weiter entwickelt.

Hier haben mittlerweile einige Änderungen statt gefunden:

- Bei der Definition des Begriffs „Forschungsdaten“ wurde entschieden, dass eine Unterscheidung zwischen den Begriffen „Primär-“ und „Sekundärdaten“ auf dem Feld der Geistes- und Kulturwissenschaften nicht zielführend ist, da „des einen Primärdaten des anderen Sekundärdaten sind“.
- Die Funktion der digitalen Langzeitarchivierung mit all ihren Unterfunktionen wurde als iterative Tätigkeit identifiziert, welche – genauso wie die Identifiervergabe – bei jedem offiziellen (Zwischen-) Publikationsschritt erfolgt.
- Die Kuration wurde als eher inhaltliche Tätigkeit von der Langzeitarchivierung abgespalten und dem fachwissenschafltichen Workflow zugeordnet.

2.1. Der DARIAH-DE Research Data LifeCycle

Die folgende Übersicht stellt den in DARIAH entwickelten Forschungsdatenzyklus schematisch dar, wobei iterativ wiederkehrende Funktionen zur technischen Unterstützung auf der rechten Seite abgebildet sind. Die linke Seite bildet möglichst allgemeingültig dediziert geisteswissenschaftliche Tätigkeiten ab:

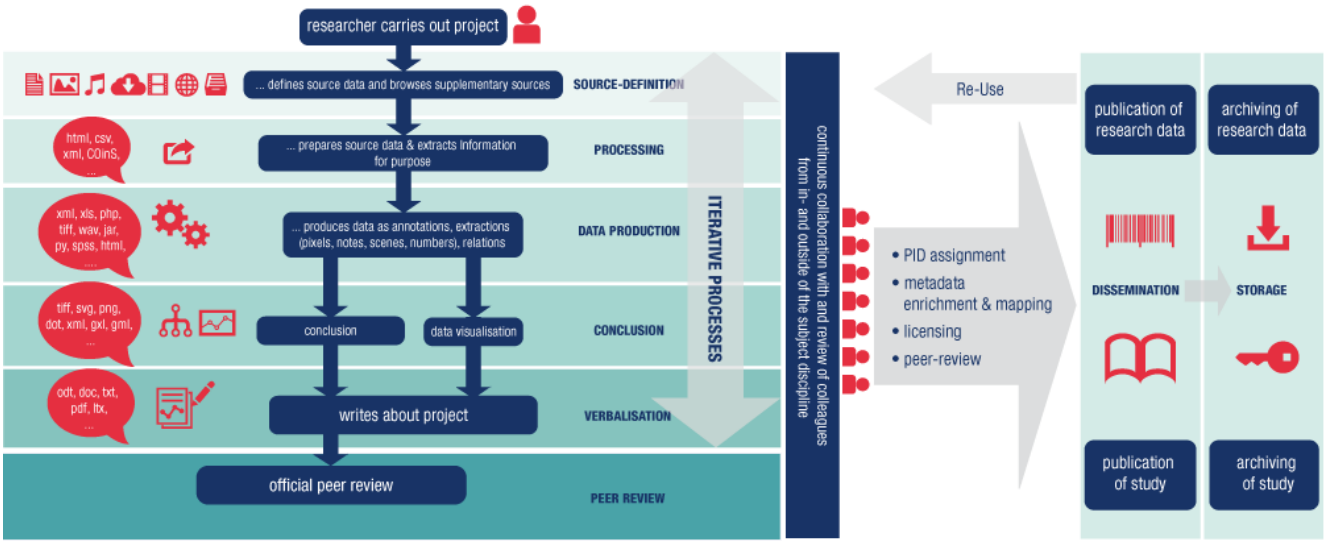

Abb. 2.1: Der DARIAH-DE Research Data LifeCycle.

2.2. Iterativität der generischen Funktionen

Die iterativ wiederkehrenden Funktionen des Forschungsdatenzyklus sind zumeist die oben beschriebenen technischen und disziplinunabhängigen Funktionen:

- Identifiervergabe
- Metadatenanreicherung /-abgleich
- Lizensierung
- Peer-Review
- Publikation

- Langzeitarchivierung

Durch die kollaborativen Arbeitsweise in den digitalen Geisteswissenschaften und den steten Bedarf – auch nicht abgeschlossene – Forschungsergebnisse der Community zur Verfügung zu stellen, können Forschungsdaten in unterschiedlichen Bearbeitungsstadien publiziert und somit auch archiviert werden, so dass eine ganze Reihe von Arbeitsschritten wiederholt aufgerufen werden.

Bei fünf der sechs „generischen“ Funktionen handelt es sich um (zumindest teil-) automatisierbare, d.h. technisch unterstützbare Funktionen. Insbesondere die Vergabe von persistenten Identifiern und die Publikation, sowie Teile der Langzeitarchivierung (Bitpreservation und automatische Metadatenextraktion) sind vollständig automatisierbar.

Die Tätigkeit des Peer-Review ist hingegen – zumindest wenn nicht anderweitig stark institutionell formalisiert – ein stark informeller Arbeitsschritt, der aus einer Reihe nicht technisch abbildbarer Prozesse, wie Workshops, Konferenzen und Gesprächen besteht und daher auch nicht automatisierbar ist.

3. Prototypische Umsetzung

Die für den Forschungsdatenzyklus identifizierten Tätigkeiten sind größtenteils durch DARIAH-DE Services implementiert.

Einige (nicht automatisierbare Schritte, wie der oben beschriebene Peer-Review) sind naturgemäß von einer solchen Implementation ausgeschlossen, werden aber durch die Bereitstellung von DARIAH-DE Diensten erleichtert.

Daneben ermöglichen gerade die Vorarbeiten aus Cluster 5 die Abbildung fachwissenschaftlicher Forschungsdatenzyklen auf Basis des generischen Basiszyklus.

3.1. Übersicht der Workflowfunktionen und den korrespondierenden Services

Die folgende Tabelle bietet eine Übersicht abstrahierter geisteswissenschaftlicher Tätigkeiten und die hierfür verwendbaren DARIAH Services.

Schritt	Aktivität	Notiz	Korrespondierender Service
1	Formulierung Forschungsfrage, Benennung von Methoden	Die Formulierung der Forschungsfrage als auch die Benennung der dazu verwendeten Methoden sind ein kreativer Akt und können nicht automatisiert werden.	Brainstorming und Kollaboration prinzipiell möglich mithife von **Wikis und Etherpads**
2	Auswahl Daten	Ist eng an die Forschungsfrage geknüpft und damit nicht automatisierbar.	Suche & Download geeigneter Forschungsdaten in der DARIAH-Collection Registry
3	Vorbereitung / Verwendung von Tools	Ist prinzipiell (teil) automatisierbar, hängt stark von der Detailtiefe in der maschinenlesbaren Beschreibung der Metadaten ab.	Ggf. Nutzung einzelner in DARIAH-DE bereitgestellter oder weiter entwickelter Tools, wie Schema-Registry, Digivoy, MEISE, Geobrowser,...
4	Generierung von (Zwischen-) Ergebnissen/ Verwendung von Tools	Wenn alle Vorbereitungen getroffen und spezifiziert sind: Ja	Nutzung externer Tools
5	Visualisierung	(Teil-)automatisiert möglich	Möglich durch einige in DARIAH bereitsgestellte Tools, wie Geobrowser und Cosmo-Tool
6	Beschreibung der veröffentlichungs- würdigen Ergebnisse und Erkenntnisse	Ist ein kreativer Akt. Produktion begleitender Publikationen	Nutzung von Confluence-Wiki und Publikation mithilfe von DARIAH-DE Publish
7	Kuration / Vorbereitung der Archivierung	Ist ein intellektueller Akt. Ist nur teilautomatisiert möglich	Empfehlungen der AG RDLC und Sammlungszuweisung über Metadaten im DARIAH Repository

Tabelle 3.1.1: Gegenüberstellung fachwissenschaftlicher Tätigkeiten aus dem RDLC und DARIAH-DE Services

Ergänzend hierzu wurden die 6 iterative Arbeitsschritte identifiziert, welche nach jeden der obigen inhaltlichen Tätigkeiten wiederkehren:

a	Identifiervergabe	automatisiert möglich	EPIC PID-Vergabe durch DARIAH-DE
b	Metadatenanreicherung /-abgleich	automatisiert möglich, durch Parsing entsprechender XML-Daten	DublinCore im DARIAH Repository
c	Lizensierung	halb automatisch	Im Rahmen der Metadatenvergabe im DARIAH Repository
d	Publikation	Manuell, technisch unterstützt	Sammlungszusammenführung und Publikation im DARIAH-Repository
e	Peer-Review	manuell	Ggf. durch Nutzung von kollaborativen Tools, wie Wikis
f	Langzeitarchivierung	Iterativ (Metadatenextraktion, Bitstream-perservation)	DARIAH Repository

Tabelle 3.1.2: Iterativ technische Schritte im Research Data LifeCycle

3.2. Aufbauende fachwissenschaftliche Workflows

Die prototypische Implementation des Cosmo-Tools in DARIAH[8] bietet daneben einen **dediziert geisteswissenschaftlichen Workflow**, der die DARIAH-DE Services nutzt.

Ausgangsdaten sind hier Wikidata-Einträge aus denen Geodaten sowie Datumsangaben und Ergebnisse von NLP-Tools extrahiert werden.

Der 2. UseCase Biographien aus R 5.3.1 *Konzept UseCases* baut genau hierauf auf.

Dabei wird als Wissensbasis Wikidata herangezogen und mithilfe von Tools, die natürliche Sprache analysieren können, auf Personendaten und -Biographien automatisiert untersucht.

Die so entstandenen Metadaten werden mithilfe der DARIAH-DE *Schema-Registry* und der DARIAH-DE *Crosswalk-Registry* für unterschiedliche Anfragen aufbereitet und können durch die DARIAH-DE *Generische Registry* abgefragt werden.

Eine Ableitung des abstrakten Research Data LifeCycle auf diesen Workflow kann folgendermaßen dargestellt werden:

Schritt	Aktivität	Notiz	Fachwissenschaftliche Beschreibung	Korrespondierender Service
1	Formulierung Forschungsfrage, Benennung von Methoden	Die Formulierung der Forschungsfrage als auch die Benennung der dazu verwendeten Methoden sind ein kreativer Akt und können nicht automatisiert werden.	Ziel ist die Ermöglichung qualitativer, historischer Forschung abgeleitet aus dem Cosmobilities Projekt. „So soll mit Hilfe automatischer Methoden zur Analyse und Visualisierung von Daten die qualitative Forschung wie folgt unterstützt werden: Biographische Informationen aus unterschiedlichen Quellen sollen zu (poten-	Brainstorming und Kollaboration prinzipiell möglich mithife von **Wikis und Etherpads**

[8] Vgl. http://search.de.dariah.eu/cosmotool/search

			ziell) transnationalen Lebens- und Bewegungsprofilen historischer Personen zusammengeführt werden. Auf Basis dieser Profile sollen Eigenschaften und Regeln identifiziert werden können, welche als so genannte Internationalitätskriterien Rückschlüsse über die Wahrscheinlichkeit einer Mobilität korrelierter Personen erlauben.“ (R 5.3.1)	
2	Auswahl Daten	Ist eng an die Forschungsfrage geknüpft und damit nicht automatisierbar.	„Als Konsequenz wurde die Datenbank von Wikidata als erste Basis für die Analyse und Korrelation biographischer Daten identifiziert. Neben biographischen Daten, wie Geburts- und Sterbedaten, Berufsbezeichnungen oder nächsten Verwandten zeichnet sich Wikidata insbesondere auch durch beinhaltete Referenzen zu weiteren Datenbanken12 aus: typische Verweise umfassen beispielsweise die Bezeichner von Gemeinsamer Normdatei (GND), Virtual International Authority File (VIAF) oder des Library of Congress Name Authority File (LCNAF)“ (Ebd)	Wikidata
3	Vorbereitung / Verwendung von Tools	Ist prinzipiell (teil) automatisierbar, hängt stark von der Detailtiefe in der maschinenlesbaren Beschreibung der Metadaten ab.	„Parsing, Transformation und Indexierung der Wikidata Daten“ durch *Schema Registry* und *Crosswalk Registry*	http://dev3.dariah.eu/schereg/ http://search.de.dariah.eu
4	Generierung von (Zwischen-) Ergebnissen/ Verwendung von Tools	Wenn alle Vorbereitungen getroffen und spezifiziert sind: Ja	• „generische Extraktion von Text aus MediaWiki oder HTML Markup“ (R 5.3.1) • „Ansteuerung analytischer und verarbeitender Tools“ (R 5.3.1)	NLP Portfolio
5	Visualisierung	(Teil-)automatisiert möglich	Visualisierung biographischer Ereignisse	Hier mithilfe von OpenStreetMap
6	Beschreibung der veröffentlichungswürdigen Ergebnisse	Ist ein kreativer Akt. Produktion begleitender Publikationen	Publikation bspw. als DARIAH Working-Paper mit URN	Nutzung von Confluence-Wiki und Publikation mithilfe von DARIAH-DE Publish
7	Kuration als intellektueller Prozess	Ist ein intellektueller Akt.	Speziell die Wahl beschreibender Metadatenstandards und die Zuweisung zu intellektuellen Sammlungen, erfolgt aufgrund der spezifischen Forschungsfrage und best practice Methoden aus der Community.	Empfehlungen durch DARIAH Public Wiki. Sammlungszuweisung über Metadaten im DARIAH Repository.

Tabelle 3.2: Ein fachwissenschaftlicher Workflow in DARIAH-DE

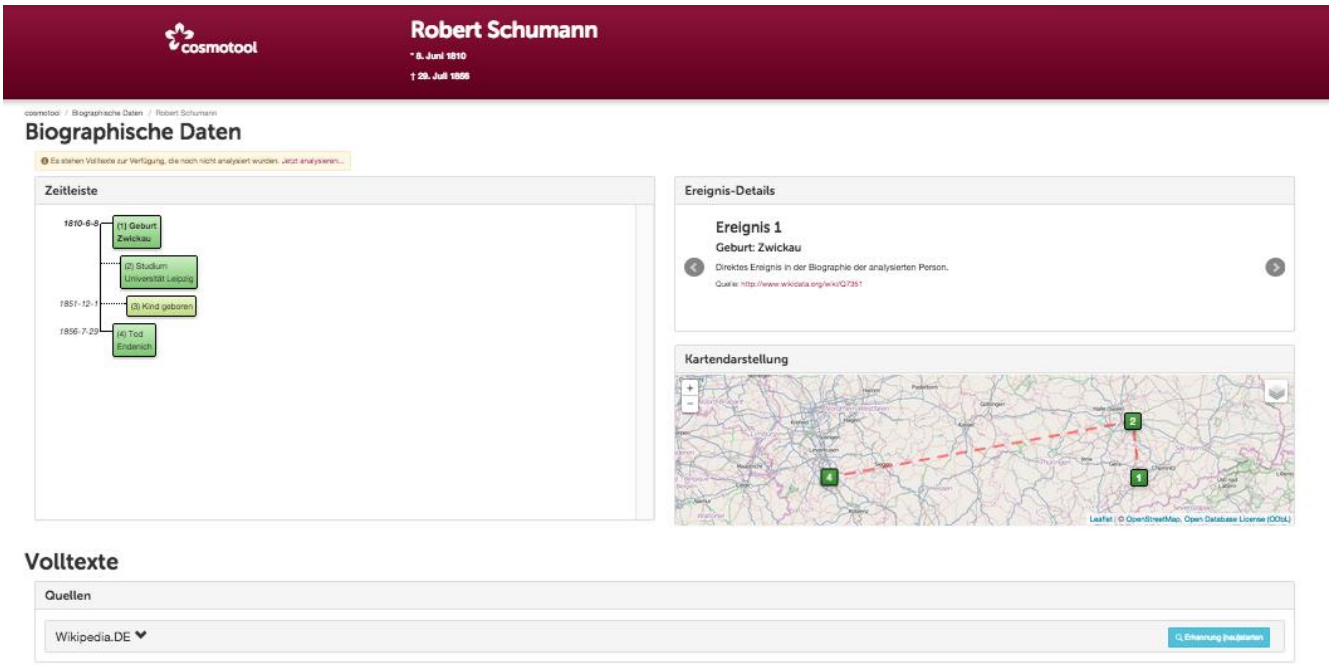

Abb. 3.2 Screenshot zu einer Cosmosearch Suche nach „Robert Schumann“

4. Verankerung der Workflows im DARIAH Repository

Die oben geschilderten Workflows sind wie folgt in der DARIAH Infrastruktur implementiert.

4.1. Verankerung des fachwissenschaftlichen Workflows

Der generische geisteswissenschaftliche Forschungsdatenzyklus ist mit folgenden Implementationen verankert:

Schritt	Aktivität	Korrespondierender Service
1	Formulierung Forschungsfrage, Benennung von Methoden	Etherpad Einrichtung unter https://etherpad.dariah.eu/ Einrichtung Wiki unter http://dev2.dariah.eu/wiki, Infos zu neuen Accounts unter https://dev2.dariah.eu/wiki/display/DARIAH2/Confluence+Server
2	Auswahl Daten	Referenzen zu Forschungsdaten unter http://colreg.de.dariah.eu/ Eingrenzung / Auswahl ggf. mithilfe der „Generischen Suche“: http://search.de.dariah.eu/ Suche und weitere Collections unter: http://textgridrep.de/
3	Vorbereitung / Verwendung von Tools	Schema-Registry: http://dev3.dariah.eu/schereg/mapping Textgrid mit DigiVoy: http://textgridrep.de/voyant/
4	Generierung von (Zwischen-) Ergebnissen/ Verwendung von Tools	Diverse innerhalb DARIAH angebotene Tools • Voyant Tools: http://textgridrep.de/voyant/ • Datasheet Editor des GeoBrowsers: http://geobrowser.de.dariah.eu/edit/ • Cosmo-Tool: http://search.de.dariah.eu/cosmotool/search • Weitere...
5	Visualisierung	Geobrowser: http://geobrowser.de.dariah.eu/ und Cosmo-Tool: http://search.de.dariah.eu/cosmotool/search
6	Beschreibung der veröffentlichungs-würdigen Ergebnisse und Erkenntnisse	Confluence-Wiki https://dev2.dariah.eu/wiki/display/DARIAH2/Confluence+Server
7	Kuration als intellektueller Prozess	• Empfehlungen der AG Research Data LifeCycle • perspektische Unterstützung durch ein Ampelsystem im DARIAH Repository

Tabelle 4.1

4.2. Verankerung der generischen Funktionen des LifeCycle

Die oben geschilderten generischen und wiederkehrenden Funktionen des Research Data LifeCycle werden wie folgt abgedeckt:

Schritt	Aktivität	Korrespondierender Service
a	Identifiervergabe	EPIC PID Vergabe durch das DARIAH Repository
b	Metadatenanreicherung /-abgleich	DublinCore Vergabe / Extraktion von technischen Metadaten im DARIAH Repository
c	Lizensierung	Im Kontext der Metadatenvergabe im DARIAH Repository & perspektischer Empfehlungen auf der DARIAH-DE Seite

d	Publikation	Sammlungszusammenführung und Publikation im DARIAH-Repository
e	Peer-Review	Kollaborativ (nicht maschinell lösbar, unterstützt durch diverse DA-RIAH Services)
f	Langzeitarchivierung	DARIAH BitstreamPreservation

Tabelle 4.2

Das kürzlich im Prototyp fertig gestellte DARIAH-DE Repository deckt den größten Teil der automatisierbaren iterativen Komponenten des Research Data LifeCycle ab.

Gerade der Ingest über die DARIAH Publish GUI bietet hier umfangreiche Optionen:

Abb 4.2.1: Die Startseite der Publish GUI (Beta) unter dariah.testportal.dariah.eu/publish-gui

Hier werden Epid PIDs vergeben, Metadaten angefügt oder eingelesen, Sammlungsbeschreibungen abgegeben, Lizenzen hinzugefügt und für die Langzeitarchivierung demnächst außerdem eine Dateiformatempfehlungen für angefügte Dateien gegeben.

Die Speicherung und Archivierung der Daten selbst erfolgt über die DARIAH-Bitpreservation, welche direkt an das DARIAH Repository angeschlossen ist.

Auf diese Weise lassen sich einmal abgespeicherte Forschungsdaten (-Sammlungen) langfristig nachweisen und auch erneut als Datenbasis nutzen.

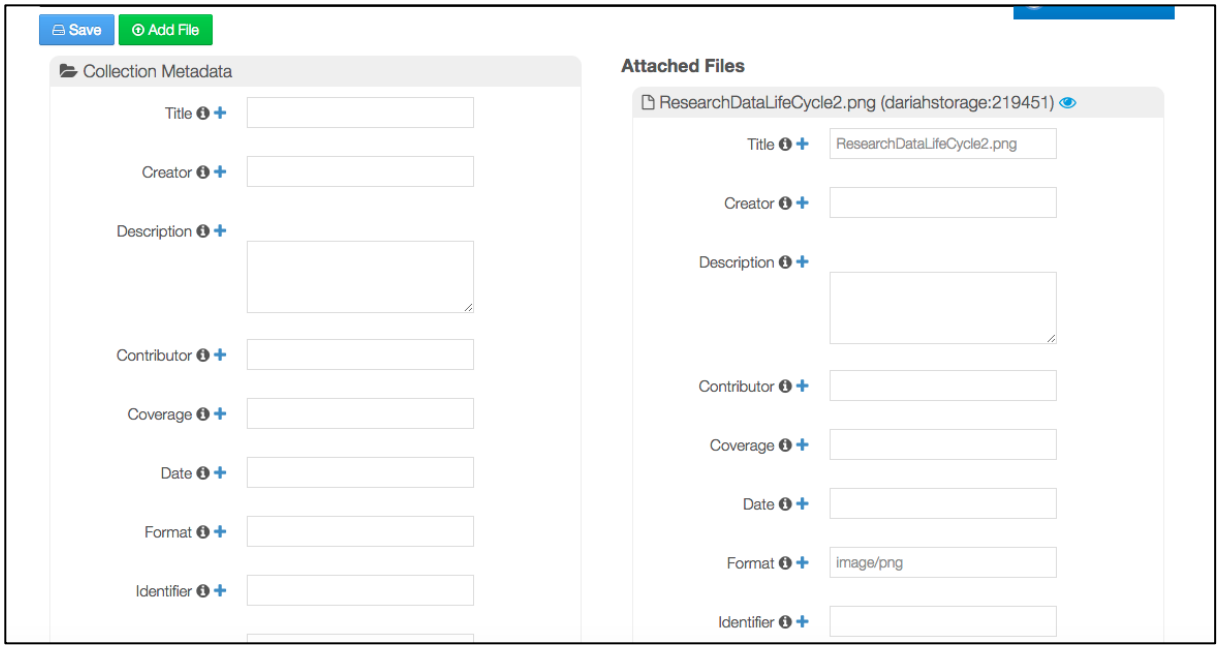

Abb. 4.2.2: Screenshot des Ingest-Screens aus der Publish-GUI Beta des DARIAH-DE Repositories

4.3. Feststellung

Die hier prototypisch umgesetzten Workflows bieten zusammen mit dem theoretischen Basisgerüst des DARIAH Research Data LifeCycle eine erste Übersicht über die Möglichkeiten, wie geisteswissenschaftliche Arbeitsprozesse mithilfe von sinnvoll eingesetzten Verfahren der Informationstechnologie unterstützt oder sogar vollautomatisch abgebildet werden können. Für das folgende und letzten Jahr der zweiten Förderphase von DARIAH-DE ist eine konsistente und nachhaltige Implementation der genannten Workflows sowie ggf. deren inhaltliche Erweiterung geplant.