Kriterien für die Auswahl zusätzlich benötigter technischer Workflows und Policies (R 2.3.2)

Version 1 – 28.11.2014

Cluster 2.3

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
18.11.2014	Johanna Puhl	First Draft
29.11.2014	Manfred Thaller	Ergänzungen
01.12.2014	Johanna Puhl	Final Version

Inhalt

1. Einleitung

Aufbauend auf dem Report 2.3.1 *Auswahl und Beschreibung der initialen technischen Workflows und Policies für den Data LifeCycle*[1] sowie ganz generell den Vorarbeiten aus der ersten Forschungphase von DARIAH-DE, sollen in diesem Report Kriterien definiert werden, auf deren Basis weitere Workflows und Policies zur Unterstützung eines Research Data LifeCycle ausgewählt und in die technische Infrastruktur von DARIAH-DE überführt werden können.

Dabei beschäftigt sich Kapitel 2 mit der Rekapitulation bereits vorhandener Workflows und Policies aus den beiden eben genannten Vorhaben. Kapitel 3 und 4 stellen organisatorische und fachliche Kriterien zusammen, auf deren Basis weitere Workflows und Policies Eingang in die Implementation eines Research Data LifeCycle finden können.

Kapitel 5 schließlich versucht einen runden Abschluss dieser Thematik, damit trotz dynamischer technischer Entwicklung im Projekt ein erster Research Data LifeCycle bereits im März 2015 erfolgreich implementiert sein kann.

2. Vorarbeiten aus R 2.3.1 und der AG RDLC

Generell verteilt sich die Arbeit zwischen Cluster 2.3 und der AG Research Data LifeCycle auf getrennte aber miteinander verbundene Aufgabenfelder:

Cluster 2.3 definiert umsetzbare Arbeitsabläufe und beschreibt die dafür notwendigen Policies, die aber direkt auf den eher theoretischen Überlegungen der AG Research Data LifeCycle[2] aufbauen.

Im Report 2.3.1 werden vor allem bereits bestehende Komponenten und Features der DARIAH-DE Infrastruktur beschrieben und gemeinsam mit den Empfehlungen der AG Research Data LifeCycle in einen umsetzbaren Basis-Workflow[3] überführt.

Basierend auf den Überlegungen zu einem Basis-Workflow wurde weiterhin der Bedarf nach einem (komplexen) Datenmodell herausgearbeitet, für welche in der AG Research Data LifeCycle aktuell nach adäquaten bzw. ausbaubaren Metadatenstandards gefahndet wird, die als Empfehlungen für die (Weiter-) Entwicklung eines Workflows in Cluster 2 verwendet werden können.

Ein äquivalenter Prozess wurde mit den Policies beschritten:

Cluster 2.3 generiert aus den Vorarbeiten der AG Research Data LifeCycle Richtlinien, die bei der Verarbeitung von Forschungsdaten durch die DARIAH-DE Infrastruktur eingehalten werden soll(t)en.

---

[1] Vgl. https://docs.google.com/document/d/1rt_0XDNoVINjUxG7a7kJAtYo4mZTYUNw229Q4X_i_IA/edit#

[2] Vgl. dazu das Dokument der AG Research Data LifeCycle: https://docs.google.com/document/d/12tSyZdByWH7I0wb2xGAbh38cw78OezRdjHEGmPIiYIM/edit#

[3] Zum Basis-Workflow vergleiche Kapitel 6.1 in R 2.3.1

Diese können und sollen gemäß der im vorliegenden Dokument skizzierten Kriterien sukzessive erweitert werden.

Es handelt sich aktuell um Richtlinien zu:

- (Persistenten) Identifiern
- Prüfsummen
- Verarbeitbaren Dateiformaten
- empfohlenen Metadatenstandards
- ggf. Rollen und Rechtemanagement

Langfristig sind außerdem Richtlinien zu verwendeten Tools und speziell zu Kurationsroutinen (Extraktion von Metadaten, Migrationstools...) zu entwickeln, für welche auch in Cluster 2.3 Kriterien zur Erweiterung und Neuaufnahme entwickelt werden müssen.

Einige der Richtlinien haben mittlerweile einen Status erreicht, auf dessen Basis sich ein Basis-Workflow aufbauen lässt, während sich andere Richtlinien noch stark im Fluss befinden.

Die beiden folgenden Kapitel diskutieren daher Kriterien, die einerseits den Bedarf nach weiteren Workflows realistisch einschätzbar machen und andererseits bestimmen, auf welche Weise neue Workflows und Policies Eingang in die DARIAH-DE Infrastruktur finden.

3. Kriterien für die Auswahl von Workflows

Das bisherige Hauptkriterium bei der Auswahl eines Research Data Workflows für die DARIAH-DE Infrastruktur waren in erster Linie die unbedingte Notwendigkeit eines Basis-Workflows zum Abgleich mit der bereits bestehenden Infrastruktur in DARIAH-DE sowie das Auffinden von noch nicht implementierten Funktionen.

Dabei wurde versucht, Kollegen aus allen Clustern in DARIAH-DE zu integrieren und auf diese Art fachspezifische und gleichzeitig allgemeingültige Arbeitsprozesse von digitalen Geisteswissenschaftlern zu untersuchen und daraus einen möglichst generischen Workflow für alle digitalen Geisteswissenschaften[4] zu destillieren.

Grundsätzlich erscheint es auch weiterhin sinnvoll, neben der generisch technischen Perspektive auf Arbeitsprozesse die fachliche Expertise spezialisierter Kollegen einzuholen.

Aus diesen Überlegungen ergeben sich zwei komplementäre Vorgehensweisen:

Zum einen sollen Kriterien beschrieben werden, wie weitere Workflows speziell aus den Fachwissenschaften in die Erweiterung der DARIAH-DE Infrastruktur einfließen können.

Zum anderen soll eine Vorgehensweise beschrieben werden, wie technische Entwicklungen überwacht und beobachtet werden können, so dass sie in die (Weiter-)

[4] Die Geisteswissenschaften haben hier definitiven Nachholbedarf: In der allgemeinen Sammlung *Workflows for e-Science: Scientific Workflows for Grids,* ed. by I.J.Taylor et al.*,* Springer 2007 (Nachdruck 2014) tauchen sie in keiner Weise auf.

Entwicklung von Workflows miteinbezogen werden können.

3.1 Kriterien von den Nutzern / Fachwissenschaftlern

An dieser Stelle sollen alle in DARIAH-DE Clustern und affillierten Projekten stattfindende Arbeiten daraufhin untersucht werden, ob sich aus ihnen weitere Workflows oder zu ergänzende Workflowbestandteile ableiten lassen.

Bisher sind hier zu nennen:

- Cluster 1: *Wissenschaftliche Begleitforschung*, hier insbesondere Report 1.2.1 und 1.2.3[5], Status: noch nicht veröffentlicht.
- Cluster 4: Wissenschaftliche Sammlungen und Forschungsdaten. R 4.2.3: *Dokumentation theorie- und verfahrensgeleiteter Sammlungskonzepte* und R 4.2.4.: *Aufbau und Nutzung von wissenschaftlichen Sammlungen*
- Cluster 5: *R 5.2.1 – Beschreibung der Use Cases*.
- TextGrid-Projekte

Die Analyse all dieser Arbeit kann zur Generierung weiterer Workflows beitragen. Grundsätzlich wird die Arbeit jedoch dadurch erschwert, dass in manchen Projekten zwar nachdrücklich auf die Existenz von Workflows hingewiesen wird, diese aber überwiegend nur als Graphiken dokumentiert sind[6].

Daneben ist nach wie vor das Hauptkriterium zur Aufnahme eines Workflows dessen (fachlicher) Abgleich mit der AG Research Data LifeCycle.

Hier wird momentan eine Umfrage geplant:

3.1.1 Umfrage zu Workflows und Spezifikationen in Projekten der (digitalen) Geisteswissenschaften

Im Kontext des Research Data Life Cycles in DARIAH-DE wird derzeit[7] eine Umfrage ausgearbeitet, in der sowohl wirklich genutzte Richtlinien als auch Gesamtworkflows in einzelnen Projekten aus den digitalen Geisteswissenschaften möglichst in kleinen Schritten abgefragt werden.

Neben administrativen Angaben, die zur Schätzung des Stellenwerts eines Projekts innerhalb der Community beitragen können, werden hier sowohl erwünschte Funktionen eines Workflows als auch Gesamtschilderungen eines Datenflusses abgefragt.

Die Ergebnisse dieser Umfrage können direkt in Hinblick auf die Frage, welche zusätzlichen Workflows benötigt werden und wie hoch der Bedarf ist, ausgewertet werden.

---

[5] Zitat: “Ziel ist eine Form von Kartierung der bisher genutzten Tools und Methoden, mit dem Ziel Defizite und Stärken der bisherigen Infrastruktur klarer fassen zu können, die im Rahmen des Reports 1.2.3 abschließend veröffentlicht werden soll. Zwischenergebnisse sollen jedoch frühzeitig – mit Erscheinen des Reports 1.2.1 – im Netz sichtbar sein”

[6] Z.B. http://www.deutschestextarchiv.de/dtae

[7] derzeit= Ende November 2014. Erste Draft zur Umfrage befindet sich auf http://cluster1.pad.dariah.eu/19.
Die Umfrage soll Ende 2014 / Anfang 2015 veröffentlicht werden.

3.1.2 Konkreter UseCase eines abstrakten Workflows – die Schema- / Crosswalk-Registry

Als konkreter Workflow der DARIAH-DE Infrastruktur kann die Beschreibung der Nutzung der DARIAH Schema- & Crosswalk-Registry gelten.

Um die Weiterverarbeitung auch von Metadaten inhaltlicher Natur in DARIAH-DE zu gewährleisten wurde in DARIAH-DE die Schema- und Crosswalk-Registry implementiert[8].

In der Schema-Registry werden dabei sukzessive alle in DARIAH-DE gebräuchlichen und aus den Fachwissenschaften empfohlenen Metadatenschemata eingegliedert.

Die Crosswalk-Registry hingegen wird für das Überführen von einzelnen Metadatenfeldern aus einem Schema in ein anderes verwendet, so dass geisteswissenschaftliche Forschungsdaten und ihre Metadaten auch interdisziplinär zwischen den einzelnen Disziplinen weitergereicht und verwendet werden können, ohne dass Ihre inhaltliche Beschreibung nutzlos wird und so verloren ist.

Die folgende Grafik beschreibt exemplarisch ein Workflow auf Basis der DARIAH Infrastruktur:

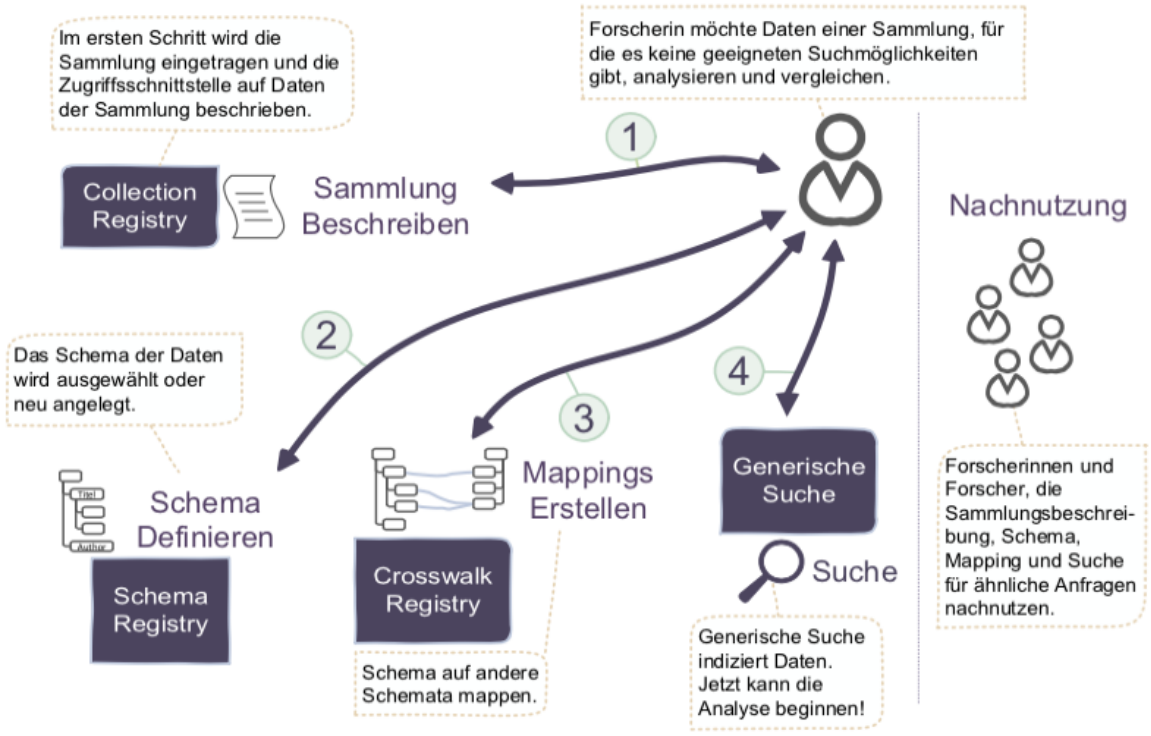

Abb. Schritte der Registrierung von Kollektionen und Schemata[9].

[8] Vgl. http://dev3.dariah.eu/schereg/

[9] Vgl. Plutte, Gradl, Henrich: Die DARIAH-DE Architektur zur forschungsorientierten Föderation von

Hier handelt es sich also um einen Workflow, der schon in der DARIAH-Infrastruktur vorhanden ist aber ggf. noch auf Basis fachwissenschaftlichen Feedbacks ausgeweitet werden kann.

3.2 Generische Kriterien

Weiter Kriterien zur Bestimmung zusätzlicher oder zur Veränderung bestehender Workflows sind klassischerweise technologische Entwicklungen, die sich aus internationalen Entwicklungen rund um Standards und Technologien für Repositories allgemein oder aber speziell für bestimmte Funktionen ergeben.

Ist hier mit echten Technologiebrüchen zu rechnen, sollten DARIAH-DE Workflows rechtzeitig daraufhin angepasst und ggf. auf eine neue Version eines Workflows ausgebaut werden.

Dies erfordert eine möglichst regelmäßige Beobachtung und Überwachung von technologischen Trends.

Kriterien zur Übernahme eines externen technischen Workflows oder zur Veränderung bestehender Workflows aufgrund von **technischen Entwicklungen** lassen sich wie folgt beschreiben:

- Besitzt die beobachtete technische Entwicklung den Charakter eines weit verbreiteten Standards / Workflows?
- Handelt es sich hierbei um einen im akademischen Umfeld gebräuchlichen Standard / Workflow?
- Wird der beobachtete Standard / der beobachtete Prozess auf Basis einer Open-Source Lizenzen bereitgestellt oder auf eine solche ausbaubar?
- ...

Solcherlei Kriterien lassen sich auch als Kriterien für eine “Technology Watch” zusammenfassen.

Da es sich hierbei aber noch nicht um eine hinreichend automatisierbare Funktionalität handelt, müssen in erster Linie Kollegen aus der DARIAH-DE Infrastruktur in regelmäßigen Abständen externe Entwicklungen bspw. durch Mitgliedschaft in diversen Gremien verfolgen und entsprechend an zuständige Stellen eskalieren.

In welchem Rhythmus dies zu erfolgen hat und auf welche Weise solch ein Monitoring formalisiert werden muss, bleibt zu diskutieren.

4. Kriterien für die Auswahl von Richtlinien (Policies)

4.1 Kriterien von den Nutzern / Fachwissenschaftlern

Analog zum vorherigen Kapitel können nicht nur für gesamte Workflows oder Datenmodelle Kriterien spezifiziert werden, sondern auch für einzelne Richtlinien.

Generell gestaltet sich hier die Suche und Auswahlkriterien einfacher, da Richtlinien im Unterschied zu Workflows bisher relativ konkret ausgeprägt sind und man daher auch konkret definieren kann, was beachtet werden muss.

Nichtsdestotrotz müssen Kriterien definiert werden, die dafür geeignet sind, eine Richtlinie für so relevant zu erklären, dass sie in den bisherigen Research Data LifeCycle überführt werden kann.

Hier eignen sich zum einen Kriterien, die sich aus dem **Nutzerfeedback** auf die ein oder andere Weise ergeben, so z.B.:

- Nutzernachfrage nach Richtlinien über offizielle Kanäle, wie z.B.[10]:
    - Helpdesk
    - Mail
    - Auf Veranstaltungen
- Automatisches Nutzer-Monitoring

4.2 Technologische Kriterien

Daneben gibt es **technologische Entwicklungen**, die einen Austausch oder eine Übernahme einer Richtlinie bewirken sollten, insbesondere wenn es sich um eine gemäß aktuellen Anforderungen unterstützenswerte Komponente handelt, einfach weil es sich hierbei mittlerweile um eine state-of-the-art Technologie handelt, die man berücksichtigen sollte.

Welche diese aktuellen Anforderungen, lässt sich in verschiedener Form erheben:

Insbesondere ist hier der Begriff der “Technology Watch” zu nennen. Dabei handelt es sich – unserem Verständnis nach – um ein möglichst automatisiertes Verfahren zur Beobachtung neuer technologischer Trends und Verfahren, über welche bei entsprechender Innovationskraft ein System automatisiert benachrichtigt werden kann.

Bisher ist es allerdings noch nicht erfolgreich geglückt ein solches automatisiertes Verfahren – zumal für ein so komplexes Konstrukt, wie den Research Data LifeCycle – erfolgreich zu implementieren. Es ist also erforderlich eine Beobachtung technologischer Trends, insbesondere im Kontext von Repositorien, Publikationsverfahren, Langzeitarchivierung und Dateiformaten in den Geisteswissenschaften analog, d.h. durch eine damit beauftragte Person oder ein aus Personen bestehendes Gremium durchzuführen.

Eingeschränkt kann auch auf externe Aktivitäten dieser Art verwiesen werden, hier insbesondere auf

- TechnologyWatch Reports der internationalen Initiative DigitalPreservationCoalition rund um Langzeitarchivierungstechnologien und -trends.[11]
- Vierteljährlich erscheinende ShowCases über neue Technologien im

[10] Inspiriert durch eine vergleichbare Aufstellung der Firma Atlassian: https://confluence.atlassian.com/display/DEV/Implementation+of+New+Features+Policy

[11] http://www.dpconline.org/advice/technology-watch-reports

Museumsbereich des Canadian Heritage Information Network[12]

- Vergangene Aktivitäten im Bereich IT in der Bildung waren außerdem die von der britischen Forschungsförderungsgesellschaft JISC publizierten TechWatchReports[13].
- Auch das DAISY Consortium, welches vor allem im digitalen Publikationsbereich tätig ist, veröffentlicht regelmäßig Reports zu technologischen Entwicklungen[14]

Neben diesen externen Aktivitäten kann auch ein DARIAH-DE internes Verfahren entwickelt werden, welches neue Technologien auf Ihre Potentiale zur sinnvollen Integration in den LifeCycle überprüft.

Solche Kriterien können sein:

- Technische Empfehlungen durch internationale Konsortien (W3C, LOC, DPC…)
- rechtlich verbindliche, oder der Verbindlichkeit nahekommende nationale Vorgaben[15],
- state of the art Reife durch entsprechende Verbreitung

Und weiterhin allgemeingültige Kriterien, wie

- lizenzrechtliche Unbedenklichkeit (entsprechend frei und nachnutzbar lizensierter Standard)
- ausreichende Dokumentationstiefe
- Adaptierbarkeit durch entsprechende Schnittstellen (Restful, SOA...)
- Hohe Standardisierung

5. Fazit

Die hier aufgelisteten Kriterien und Quellen bieten hinreichend Potential zur stetigen Ergänzung und Erweiterung der bestehenden Policies und Workflows in DARIAH-DE und helfen gleichzeitig bei der Auswahl von zur Verfügung stehenden Richtlinien und Arbeistflüssen.

Dabei wurde insbesondere auf die Einteilung in geisteswissenschaftliche, eher nutzer-zentrierte, Kriterien und eher technologische grundlegende Kriterien hingewiesen.

[12] http://www.rcip-chin.gc.ca/sgc-cms/nouvelles-news/anglais-english/?p=7898

[13] http://www.webarchive.org.uk/wayback/archive/20121224235815/http://www.jisc.ac.uk/whatwedo/services/techwatch/reports.aspx

[14] http://www.daisy.org/techwatch

[15] Z.B.BSI TR 03125 – https://www.bsi.bund.de/EN/Publications/TechnicalGuidelines/TR03125/BSITR03125.html