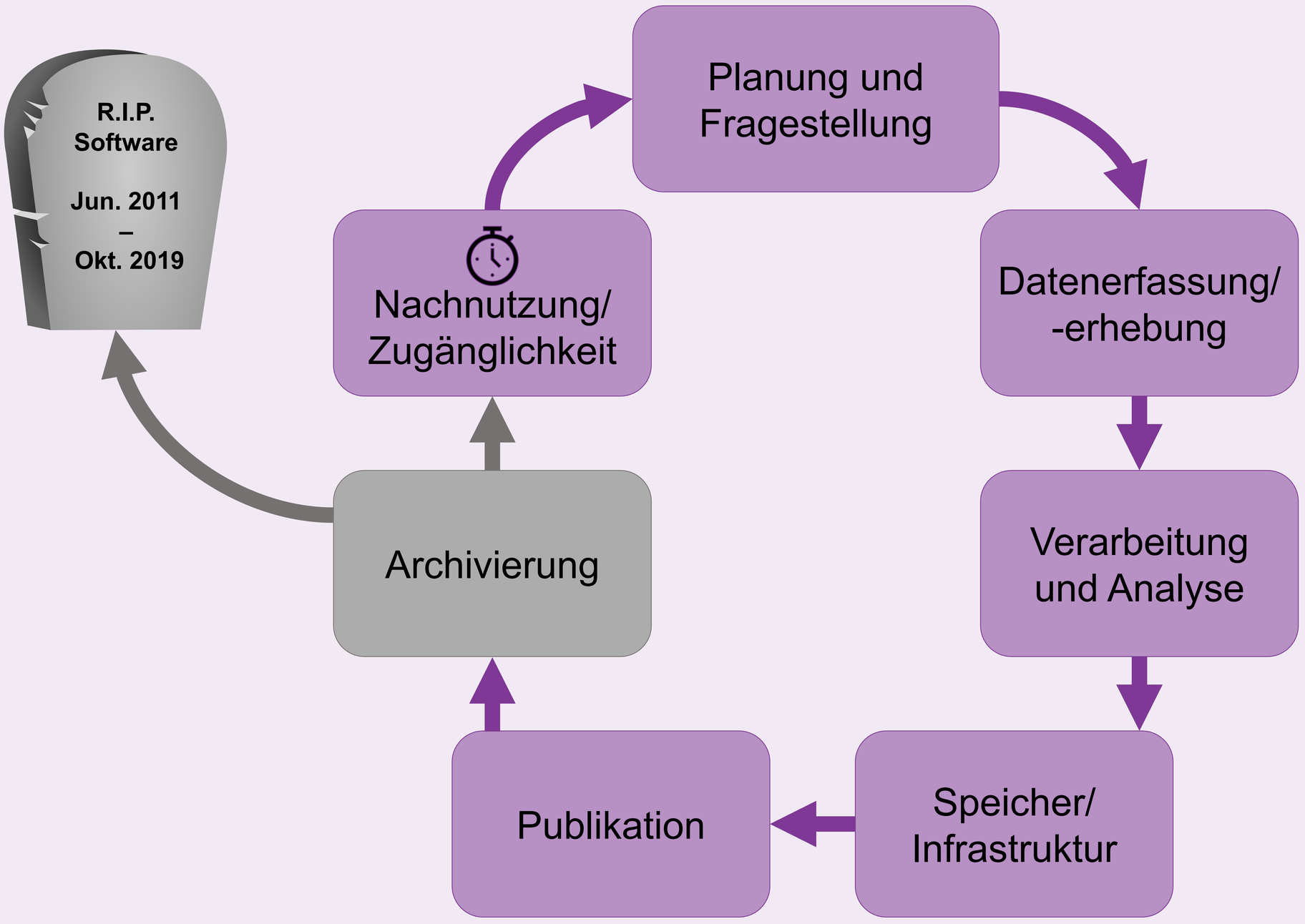
# Nachhaltigkeit von Forschungsdateninfrastrukturen

## am Beispiel von Digitalen Editionen – Was steht noch in 2019?

### Forschungsdaten und Lebende Systeme

Die nachhaltige und langfristige Sicherung, Verfügbarkeit und Nachnutzung von Forschungsdaten durch digitale Archive, Repositorien und Beschreibungsstandards befördert wissenschaftliche Arbeit, schafft Verlässlichkeit und Nachvollziehbarkeit und entspricht der guten wissenschaftlichen Praxis. Für den effektiven Umgang mit Forschungsdaten muss das Forschungsdatenmanagement (FDM) über den gesamten Forschungsprozess und den entsprechenden Forschungsdatenlebenszyklus mitgedacht werden. Insbesondere innerhalb der heterogenen geisteswissenschaftlichen Forschungslandschaft stellt dabei die hohe Diversität der entstehenden Forschungsdaten eine große Herausforderung für das Datenmanagement dar: Unterschiedliche Formate, Datentypen und Standards müssen bedient, Bedarfe und Tradition verschiedener Fachcommunities müssen berücksichtigt werden. Gleichzeitig sind häufig dynamische, Lebende Systeme wie Digitale Editionen oder interaktive Präsentationssysteme Ergebnisse geisteswissenschaftlicher Forschungsprozesse oder stellen die einzige Zugangsschicht zu Forschungsdaten dar. Auch diese Systeme müssen zur langfristigen Sicherung, Verfügbarkeit und Nachnutzung kuratiert werden. Veraltete Software, überholte Skriptsprachen, nicht mehr unterstützte Anwendungen und proprietäre Systeme können in der Regel jedoch nur mit hohem Aufwand bewahrt und betrieben werden. Notwendige Ansätze der Homogenisierung von Frameworks und Toolchains oder einer geplanten Virtualisierung von Systemumgebungen werden häufig nicht mitgedacht und im Forschungsprozess nicht berücksichtigt.




### Das Problem der Nachhaltigkeit

- Es existieren etablierte Strategien zur Sicherung, Verfügbarhaltung und Nachnutzung von Forschungsdaten
- Häufig sind Lebende Systeme geisteswissenschaftliche Forschungsergebnisse, oder einzige Zugangsschicht zu Forschungsdaten
- Es gibt im FDM einen blinden Fleck bei der Kuratierung, Sicherung, Verfügbarhaltung und Nachnutzung von Lebenden Systemen
- Software und Systeme verschwinden, Forschungs-daten und -erkenntnisse gehen verloren

### Kuratierungslücke – Der Tod Digitaler Editionen

Ausgehend von einer umfangreichen Sammlung (Patrick Sahle: http://digitale-edition.de/vlet_a-z.html) haben wir eine Analyse der Langlebigkeit am Beispiel von Digitalen Editionen als Lebende Systeme vorgenommen. Wir haben die Sammlung mittels eines speziell dafür angefertigten Programms (SeaCliff, https://github.com/schlusslicht/seacliff) in eine Datenstruktur überführt, die über folgende Indikatoren verfügt: (1) Name der Edition, (2) Kurzbeschreibung, (3) Editoren, (4) Weblink, (5) Verfügbarkeit, (6) Popularität, (7) erster beziehungsweise (8) letzter Kontakt laut Editor und (9) erster beziehungsweise (10) letzter Kontakt laut der Wayback Machine. Alle Informationen, bis auf die Popularität und die Erst- bzw. Letztkontakte mit der Wayback Machine, stammen aus der kuratierten Liste. Die Popularität entstammt der offenen Datenbank OpenPageRank.com, die zu Vergleichszwecken von Websites Ranking-Kennzahlen bereithält. Die Erst- sowie Letztkontakte mit der Wayback Machine wurden deren entsprechenden Datensätzen entnommen. Da unsere Analyse darauf abzielt die Langlebigkeit der gelisteten Digitalen Editionen zu ermitteln, sind in der Auswertung der Daten insbesondere der erste beziehungsweise der letzte erfasste Kontakt die entscheidenden Indikatoren.

Von den 466 aufgeführten digitalen Editionen in der Sammlung ließen sich 451 maschinell auswerten, wobei hiervon 402 berücksichtigt wurden. Aus der Analyse herausgefallen sind solche Einträge, die nicht maschinell verarbeitet werden konnten und solche, denen weder manuelle noch von der Wayback Machine spezifizierte Erst- und Letztkontakte nachgewiesen werden konnten.

**Die Grundgesamtheit Digitaler Editionen**

- Sammlung kuratiert seit **1997**
- **466** Digitale Editionen erfasst
- Erstpublikationen von **1990-2018**

**Die Analyse der Daten**

- **451** maschinelle Auswertungen
- **402** Digitale Editionen betrachtet
- **86,3%** der Grundgesamtheit

**Das Aussterben Digitaler Editionen**

- ⌀ Halbwertszeit: **6 Jahre**
- ⌀ Lebenszeit: **8,5 Jahre**
- **Keine** langfristige Nutzbarkeit

### Lebende Systeme im Forschungsprozess

Bei Digitalen Editionen handelt es sich, neben interaktiven Präsentationssystemen oder kollaborativen Arbeitsumgebungen, um Lebende Systeme, die die Besonderheit bergen, dass ihre Laufzeitumgebung insofern forschungsrelevante Daten enthält, als dass der Verlust von Aggregations- oder sonstigen Aufbereitungsvorgängen der zu Grunde liegenden Primärdaten zum Verlust impliziter Erkenntnisse führt. Dieser Umstand ist vergleichbar mit dem Betrachten eines digitalen Bildes mittels eines dafür vorgesehenen Werkzeugs, beispielsweise eines Bildeditors, oder in seiner Rohform, also als Binärdaten: Im ersten Fall ist der betrachtenden Person das Motiv des Bildes direkt ersichtlich, im letzten Fall erlauben die Binärinformationen kaum ein Erkennen. Im Gegensatz zu Bildern, die alle notwendigen Informationen für ihre grafische Darstellung enthalten, verfügen Lebende Systeme über sogenannte fluide Daten, also solche, die nicht verstetigt werden können und daher nur im Kontext der entsprechenden Forschungsapplikation verfügbar sind.

Da die Überführung solcher Lebenden Systeme in statische Daten immer mit Erkenntnisverlusten behaftet ist, erforscht das Data Center for the Humanities (DCH) der Universität zu Köln gemeinsam mit dem Institut für Architektur von Anwendungssystemen (IAAS) der Universität Stuttgart derzeit die Möglichkeit, den OASIS Standard „Topology and Orchestration Specification for Cloud Applications“ (TOSCA) auf in den digitalen Geisteswissenschaften erzeugte Forschungsapplikationen anzuwenden. Neben der Möglichkeit Softwarekomponenten standardkonform abzubilden, soll so die Langlebigkeit und Nachnutzbarkeit solcher Forschungsergebnisse deutlich verbessert werden.

Patrick Helling, Philip Schildkamp, Brigitte Mathiak
Data Center for the Humanities (DCH), Universität zu Köln
Tel.: +49 221 / 470 1307
E-Mail: info-dch@uni-koeln.de

SustainLife
DCH Data Center for the Humanities Kölner Datenzentrum für die Geisteswissenschaften
Universität zu Köln