TextGrid-Produktivsystem (M4.3.3.2)

Version 08.10.2018
Cluster 4
Verantwortlicher Partner SUB Göttingen

Weiterentwickeltes Produktivsystem des TextGrid Repository

Dieses Forschungs- und Entwicklungsprojekt wird / wurde mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF), Förderkennzeichen 01UG1610A bis J, gefördert und vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR) betreut.

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Projekt: DARIAH-DE: Überführung der digitalen Forschungsinfrastrukturen für die e-Humanities in die Operational Phase (Betriebsphase)
BMBF Förderkennzeichen: 01UG1610A bis J
Laufzeit: März 2016 bis Februar 2019

Dokumentstatus: Final
Verfügbarkeit: öffentlich
Autoren: Raisa Barthauer, Stefan Funk

Revisionsverlauf:

Datum	Autor	Kommentare
08.10.2018	Raisa Barthauer	Entwurf
17.10.2018	Stefan E. Funk	Finalisierung

Inhaltsverzeichnis

1. Einleitung

Forschungsdaten und digitale Werkzeuge und Dienste zu ihrer Nutzung sind wesentliche Grundlagen der geistes- und kulturwissenschaftlichen Forschung. Durch die Verknüpfung, Archivierung und Nachnutzung dieser Daten wird nicht allein eine vorher unbekannte Forschungsdynamik möglich, sondern es können auch völlig neue Forschungsfragen bearbeitet werden. DARIAH-DE als digitale Forschungsinfrastruktur für die Geistes- und Kulturwissenschaften trägt dieser Prämisse mit seinem Angebot Rechnung, indem es zum einen Forschungsdaten verfügbar macht, zum anderen digitale Werkzeuge zum Forschungsdatenmanagement entwickelt und bereitstellt.
TextGrid[1] stellt in diesem Rahmen eine virtuelle Forschungsumgebung für Geisteswissenschaftlerinnen und Geisteswissenschaftler dar. Der Dienst bietet Open-Source-Werkzeuge und -Services an, die den gesamten Forschungsprozess unterstützen sollen. Im TextGrid Repository[2] können darüber hinaus Forschungsdaten langzeitarchiviert werden. Neben der sicheren Speicherung bleiben die Daten dort durchsuchbar und können publiziert werden.
Um das TextGrid-Produktivsystem zu fördern, besser nutzbar zu machen und weiter voranzubringen, werden von DARIAH-DE zahlreiche Anstrengungen unternommen. Für die Weiterentwicklung des Produktivsystems wurden neue Dienste in TextGrid eingebunden und nutzbar gemacht, Systeme geprüft und deren Leistungsfähigkeit verbessert sowie schneller und zentraler konfigurierbar gemacht.
Das vorliegende Dokument dient als Nachweis des zum 1.10.2018 im Rahmen von Cluster 4 „Wissenschaftliche Sammlungen“ erbrachten Meilensteins „M4.3.2.2 Weiterentwicklung Produktivsystem TextGrid-Repository“. Es fasst die einzelnen Bestandteile der erfolgten Weiterentwicklung kurz zusammen. Das TextGrid Repository ist über folgende URL erreichbar: https://textgridrep.org/

2. Funktionale Weiterentwicklung des TextGrid-Produktivsystems

Bis Oktober 2018 wurden mehrere entscheidende Ziele umgesetzt, die TextGrid nachhaltig sowohl in der Umsetzung als auch in der Anwendung verbessern werden, die im Folgenden kurz beschrieben werden. Es handelt sich dabei um:

- Die Einrichtung des Dienstes Voyant2,
- die Erweiterung und Stabilisierung des Dienstes Digilib samt Mirador-Viewer,
- die Einbindung von Metriken für Ressourcen-Nutzung,
- die „Puppetisierung“ der TextGrid Repository-Server,
- die Verbesserung des Release-Managements der TextGrid Repository-Dienste,
- den Test des ISILON-Dateisystems sowie

[1] https://textgrid.de/
[2] https://de.dariah.eu/repository

- die Einführung von Annotationen in die TextGrid Repository Webseite.

Mit der neuen Version Voyant2 wurde unter www.textgridrep.de eine weitere wichtige Anwendung für Forschende nutzbar gemacht. Das Tool zur Textanalyse und -interpretation wurde in das TextGrid Repository eingebunden sowie der Zugriff auf Voyant (Version 1) erhalten. So können beide Tools bei allen im TextGrid Repository gespeicherten Texten direkt angewählt und angewendet werden.
Voyant2 ermöglicht das Sichtbarmachen und Analysieren von Sequenzen bis hin zu einzelnen Wörtern in den ausgewählten Texten. Sich wiederholende Sequenzen von Wörtern können aufgezeigt werden, die nach der Häufigkeit der Wiederholung oder der Anzahl der Wörter in den wiederholten Phrasen sortiert werden können. Phrasen können gefiltert werden, indem Suchanfragen in das Suchfeld eingegeben werden. Mit einem Schieberegler können beispielsweise die Länge der Phrasen, der Kontext oder auch die Granularität eingegrenzt werden.
Es kann aus einer Reihe von zugehörigen Werkzeugen ausgewählt werden, etwa einer Wordcloud, die die am häufigsten auftauchenden Wörter abbildet, einem Tool, das Verknüpfungen einzelner Begriffe darstellt oder Graphen, die die relative Häufigkeit von Begriffen im Text anzeigt. Die Wörter können dabei einzeln angewählt und dargestellt werden, die Anzahl der gezeigten Wörter kann individuell reguliert werden. Das Tool ermöglicht damit den Forschenden eine erweiterte Form der Textanalyse. Vorher unbemerkte Zusammenhänge werden sichtbar gemacht und können zu neuen Ergebnissen führen.

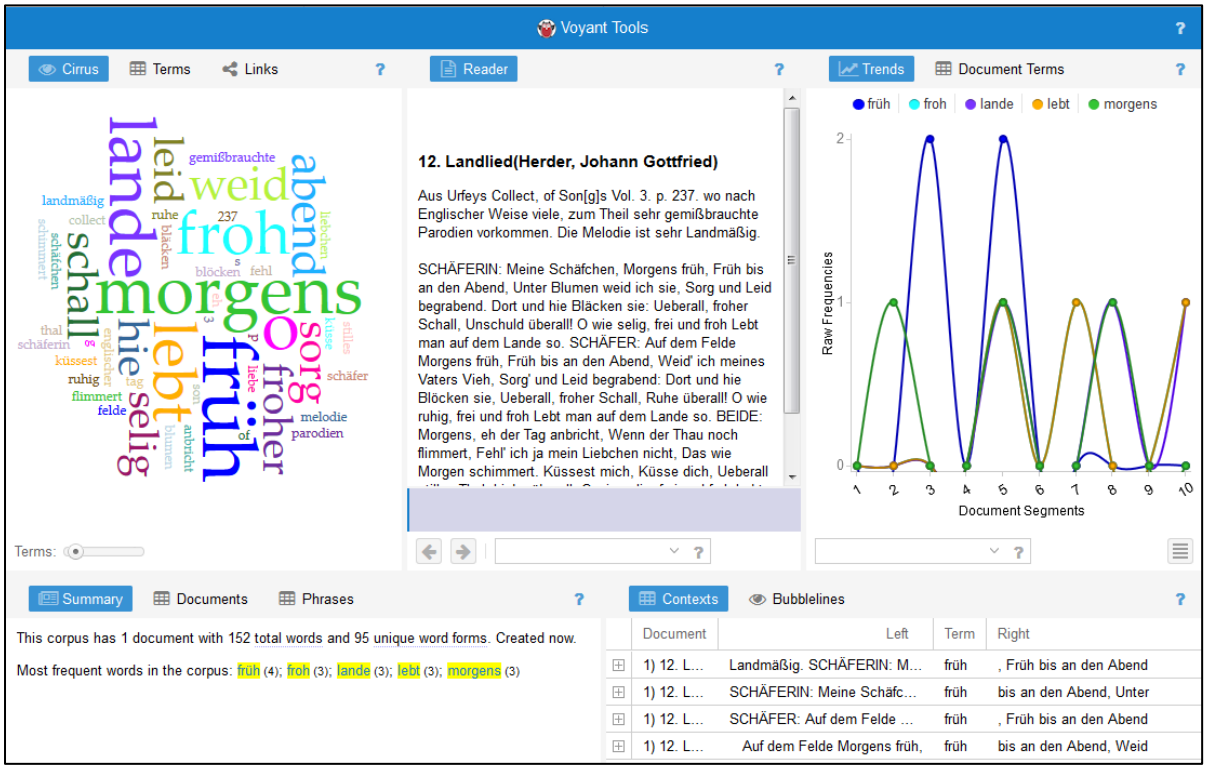

Abbildung 1: Wordcloud und Trend-Analyse mit Voyant2

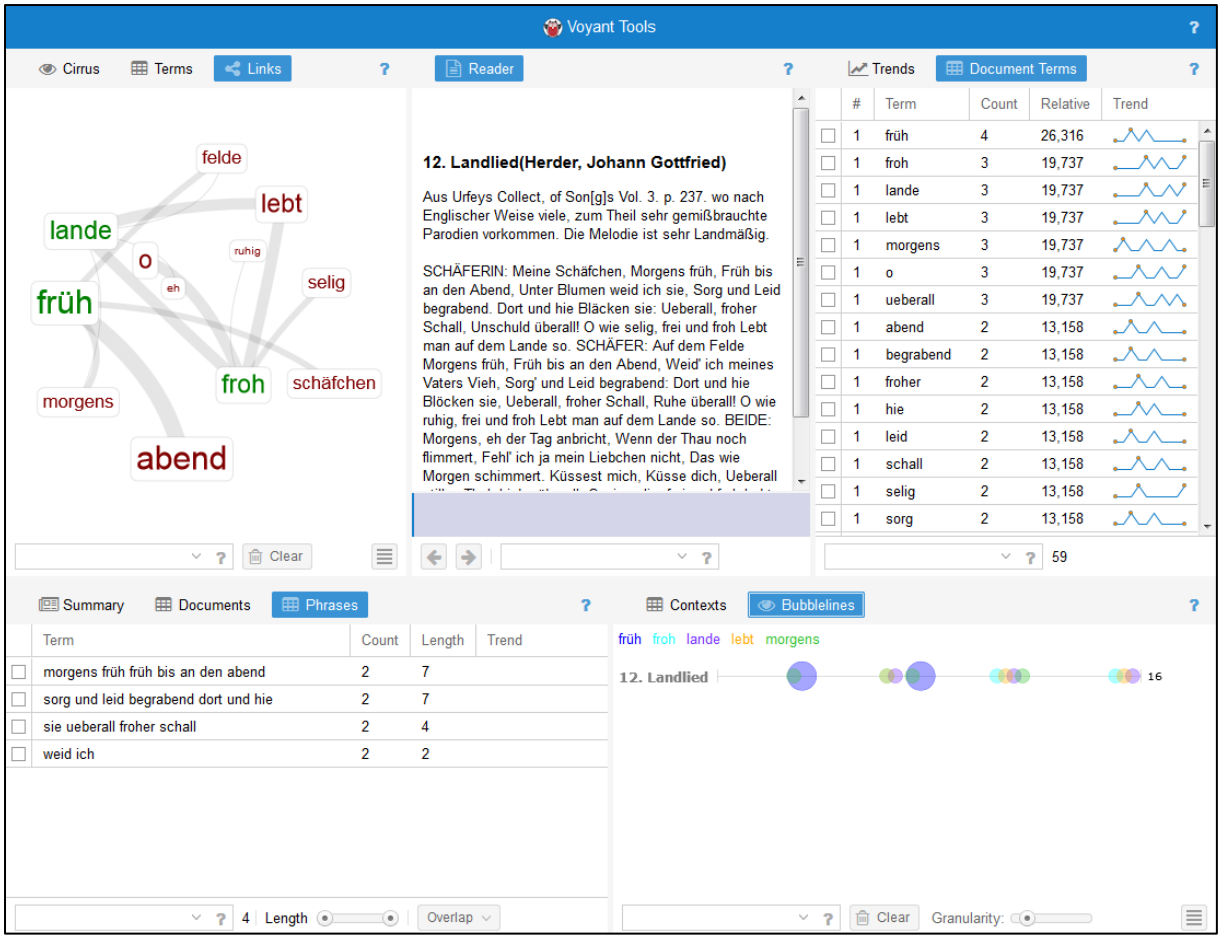

Abbildung 2: Wortanalyse und Wortverknüpfungen mit Voyant2

Ein Service, den TextGrid bereits anbietet, ist Digilib. Mit diesem Bildbetrachtungs-Tool können interaktiv Bilddateien und Ausschnitte daraus über ein Webinterface wissenschaftlich analysiert werden. Dabei enthält das Bild oder der Bildausschnitt alle Informationen, um den kompletten Darstellungskontext zu reproduzieren. Das Bild oder der Ausschnitt können als Referenz zitiert und weitergegeben werden. In diesem vorhandenen Tool wurden noch bestehende Performanz- und Stabilitätsprobleme identifiziert und behoben und somit die Anwendung sicherer, verlässlicher und effizienter gemacht. Die Bilder des TextGrid Repositorys, die mit Digilib ausgeliefert werden können, werden per IIIF-Protokoll exportiert und können im Mirador-Viewer[3] als Edition oder Kollektion visualisiert werden.

[3] https://textgridlab.org/1.0/iiif/mirador/?uri=textgrid:24ggz.0

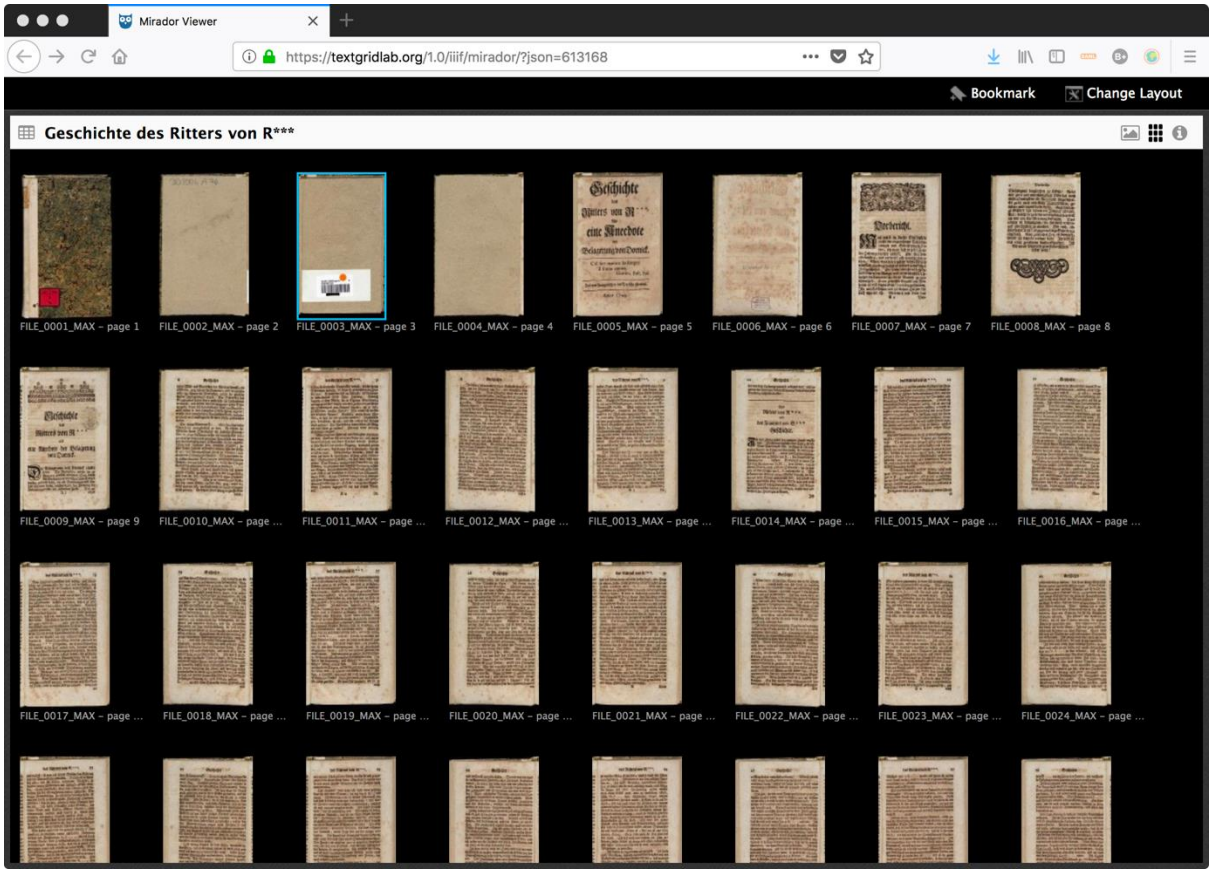

Abbildung 3: Darstellung einer Monographie im Mirador-Viewer

Das Metrics-Monitoring[4] für TextGrid Kern-Dienste wurde eingerichtet und stetig erweitert. Quantitative Aussagen über die Ressourcennutzung und die Auslastung der einzelnen Dienste können hier beobachtet und auch für einen gewissen Zeitraum analysiert werden. Das ist sehr hilfreich für die Weiterentwicklung der Dienste und zum Beispiel auch für das Feintuning der Speicherkonfiguration.

[4] https://metrics.gwdg.de (intern)

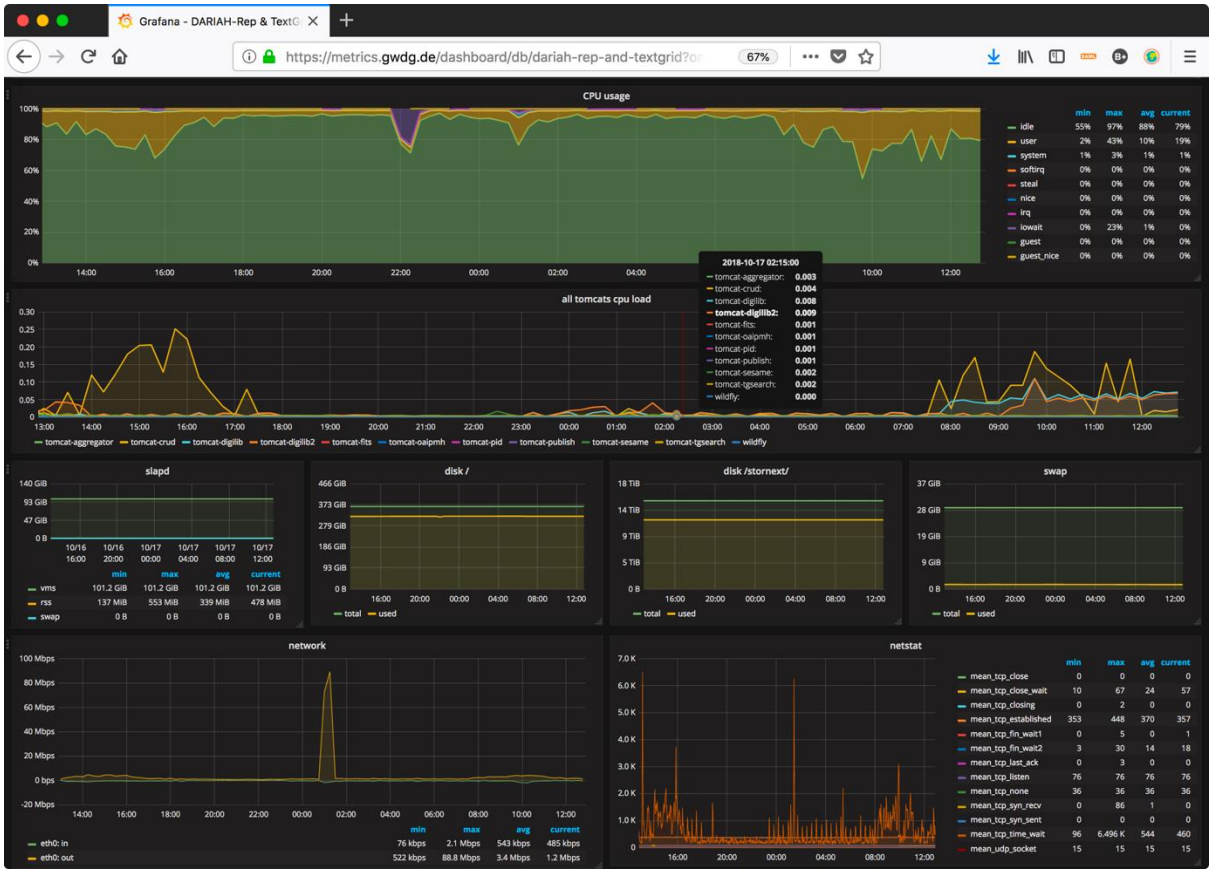

Abbildung 4: Ressourenanalyse per Grafana

Im TextGrid Repository wurde die zentrale Betreuung durch das Konfigurationstool Puppet erweitert und mit der Konfiguration des DARIAH-DE Repositorys zusammengeführt. Zu diesem Zweck wurde ein Puppet-Modul „dhrep“[5] erstellt, mit dem mittels virtueller Maschinen und der Puppet-Konfiguration jeweils einfach TextGrid- und DARIAH-DE Repository-Server erstellt und hochgefahren werden können. Gemeinsam mit dem DARIAH-DE Puppetmodul[6] ist es nun möglich, die Test-, Entwicklungs- und Produktivinstanzen des TextGrid Repositorys zentral und automatisiert zu konfigurieren und zu aktualisieren.

Ein weiterer Fortschritt auf Seiten der technischen Realisation ist die Verbesserung des Release-Managements unter Aspekten der Continuous Integration (CI). Beispielsweise wurde das Build- und Deployment-Management größtenteils auf Jenkinsfiles umgestellt, in den meisten Fällen wird automatisiert nach dem Einchecken des Quellcodes der Buildprozess angestoßen und Debian.Pakete (DEB) erstellt, die dann bei jedem Puppet-Lauf automatisch deployed werden, abhängig von der Versionsnummer jeweils auf den Produktiv- bzw. Entwicklungsservern.

Weiterhin wurden Dateisystemtests mit ISILON als Ersatz für das momentan genutzte System StorNext gemeinsam mit der GWDG durchgeführt, um den Dateizugriff auf die Daten des TextGrid Repositorys zuverlässiger und auch wartungsfreundlicher zu machen. Derzeit läuft der Entwicklungsserver stabil mit ISILON. Außerdem stellte sich heraus, dass ISILON einen schnelleren Zugriff auf die Daten ermöglicht. Sobald eine performante

[5] https://github.com/DARIAH-DE/puppetmodule-dhrep
[6] https://gitlab.gwdg.de/dariah-de-puppet (intern)

Backup-Strategie für ISILON eingerichtet wurde, wird auch das Produktivsystem zu ISILON migriert werden.

Als weiterer neuer Dienst wurde die Annotation-Sandbox[7] in die Seite www.textgridrep.org eingebunden. Mit der Annotation-Sandbox wurden die Tools Annotator.js und das TextGrid Repository miteinander verknüpft. Sie erlaubt den Nutzerinnen und Nutzern das Annotieren von Text- und Bilddaten, die vom TextGrid Repository gehostet werden, sowie von Websites, die mit dem DARIAH-DE Annotation-Proxy ausgestattet sind. Die erzeugten Annotationen werden dann im DARIAH-DE Annotationsstorage gespeichert. Die Annotation-Sandbox verbindet Forschungsprimärdaten, ihre Nachnutzung, Anreicherung und Analyse sowie die daraus entstehenden Ergebnisse, die wiederum in einem digital unterstützten Forschungszyklus zu Primärdaten werden können.

3. Zusammenfassung

In Bezug auf die technische Umsetzung wie im Hinblick auf angebotene Tools wurden wichtige Fortschritte gemacht. Neue relevante Dienste stehen nun unter www.textgridrep.de zur Verfügung, bereits bestehende Tools wurden um aktuelle Versionen ergänzt. Forschende werden somit in Analyse und Annotation u. a. von Texten in einem noch höheren Maß unterstützt, neue Forschungsdaten können generiert werden. Die Betreuung der Dienste wurde auf technischer Seite erleichtert, die Funktion der Dienste kann somit zuverlässiger sichergestellt werden.

[7] https://annotation.de.dariah.eu/