Virtuelle Forschungsplattformen im Vergleich: MONK, Textgrid, Transcribo und Transkribus

Piotrowski, Michael
piotrowski@ieg-mainz.de
Université de Lausanne

Schomaker, Lambert
l.r.b.schomaker@rug.nl
Niedersächsische Staats- und Universitätsbibliothek Göttingen

Horstmann, Wolfram
horstmann@sub.uni-goettingen.de
Universität Trier

Burch, Thomas
burch@uni-trier.de
Staatsarchiv des Kantons Zürich

Hodel, Tobias
tobias.hodel@hist.uzh.ch
Leibniz-Institut für europäische Geschichte Mainz

Eine zentrale Forderung zur Unterstützung digitaler Editionen ist das Anbieten virtueller Umgebungen (Interfaces, Software) zur Produktion, aber auch zum Management digitaler Daten (BMBF 2013). In den letzten Jahren wurden aufgrund dieser durch FachwissenschaftlerInnen getragenen Nachfrage mehrere Plattformen und Softwareangebote/ Infrastrukturen geschaffen, die Prozesse der digitalen Datenerstellung von der Aufnahme von Informationen (Metadaten, Transkriptionen) über die Auswertung und Anreicherung bis zur Publikation unterstützen (DARIAH-DE (Hg.), 2015) und nachhaltig betrieben werden sollen. Unterschiedliche Konzepte und angebotene Abläufe sowie integrierte Hilfsmittel stehen für eine je eigene Profilierung der Plattformen. Merkmale der Angebote, insbesondere Leistungsfähigkeit, unterstützte Prozesse und Ausrichtungen unterscheiden sich zwangsläufig. Im Panel werden aus diesem Grund wichtige und häufig eingesetzte Plattformen in ihrem Leistungsumfang verglichen und einander gegenübergestellt. Im Sinne geisteswissenschaftlicher software studies (Andrews, 2016) müssen die Plattformen nicht nur aus pragmatischen Gründen gegeneinander abgewogen werden sondern auch, um in den angebotenen Prozessen angelegte Praktiken auf ihre Logik und dadurch entstehende Folgen zu untersuchen (Drucker, 2013). Anhand eines klar umrissenen Fragebogens präsentieren Monk, Textgrid, Transcribo und Transkribus Arbeitsabläufe, Services und Vernetzungsmöglichkeiten. Damit wird Interessierten in einem Panel aus erster Hand ein Vergleich wichtiger, produktiv nutzbarer Angebote geliefert.

Das Panel wird moderiert von Michael Piotrowski (IEG Mainz).

Folgende Frage- und Themenschwerpunkte werden schriftlich und in kurzen Präsentationen dargeboten:

- Idealtypischer/Schematisierter Ablauf für den Gebrauch der Plattform
- Zeitliche Anforderungen, um ein Projekt aufzusetzen/ ein Dokument zu verarbeiten; zu exportieren
- Herstellung von Transkriptionen
- Bild-Text-Verknüpfung
- Text-Markup
- Ausgabemöglichkeiten (für Edition und/oder Transkription)
- Vernetzungsmöglichkeiten (Wörterbücher, externe Ressourcen, Ontologien)
- Datei-/Bildverwaltung
- Projektverwaltung
- Auswertungs-/Abfrageoptionen
- Automatisierungen
- Crowdsourcing/Optionen zum Einbezug von Laien oder Externen
- Nachhaltigkeit der Plattform/der enthaltenen Daten
- Updates bis 2018

Monk (presented by Lambert Schomaker, Rijksuniversiteit Groningen)

The Monk system is a trainable search engine for handwritten material. For the humanities, it may serve as a method for getting keyword access to scanned pages at the earliest stages after a document digitisation. For pattern recognition research, it is an observatory for complicated visual material and its human-provided labels (e.g., word or character labels). The system act as an e-Science service that is continuously available.

An internal image and metadata format is used, which can be exported to, e.g., PAGE xml. Provisional transcriptions can be retrieved as flat text. Indices can be exported upon request.

The system makes a distinction between four different forms of annotation: page (scan) descriptors, typically page titles, page regions of interest (tags for visual objects), transcription of segmented lines, and finally, word labeling. The system could export in TEI, however, within the OCR community, there is a preference for layout-centric description languages, as opposed to editorial descriptions. In practice, both TEI and PAGE are used, as well as other formalisms that allow to provide metadata to polygonal image sections.

In order to proceed data in Monk, scans are uploaded via sftp or mailed hard disks. The collection is then judged

on the required preprocessing steps (multicolumn, contrast enhancement, line segmentations), and 'ingested'. Within one or two days users can start to label words. The system performs data mining on the collection and presents hit lists for words which can be labeled further, and so on. Static indices and provisional transcriptions are updated nightly.

At the moment 400 documents from different periods and handwriting styles are being processed. The Monk system is one of the first 24/7 machine learning systems. The system detects where compute resources should be directed, on the basis of observed user activities and interests.

The Monk system is part of the large multi-petabyte Target platform of the university of Groningen, in collaboration with astronomy, genomics and the IBM company.

TextGrid (präsentiert durch Wolfram Horstmann, Niedersächsische Staats- und Universitätsbibliothek Göttingen)

Hintergrund

Die Entwicklung von TextGrid, einer Virtuellen Forschungsumgebung für die Geistes- und Kulturwissenschaften, wurde durch die zunehmende Nachfrage aus den Fachwissenschaften nach digitalen Werkzeugen v.a. des philologischen Edierens und kollaborativen Arbeitens angestoßen. Das Bundesministerium für Bildung und Forschung (BMBF) hat TextGrid als Verbundprojekt mit über zehn institutionellen und universitären Partnern zwischen 2006 und 2015 gefördert.

Die Software steht mittlerweile in einer stabilen Version 3.0 zum kostenfreien Download bereit. Software, Archiv und damit das gesamte Angebot werden in Zusammenarbeit mit AnwenderInnen, FachwissenschaftlerInnen und Fachgesellschaften und in Kooperation mit DARIAH-DE - Digital Research Infrastructure for the Arts and Humanities weiter entwickelt und dauerhaft betrieben.

Zielpublikum

FachwissenschaftlerInnen, die mit TextGrid Forschungsprojekte wie z.B. digitale Editionen erarbeiten

EntwicklerInnen, die TextGrid-Tools und Services für eigene Vorhaben anpassen oder externe Services und Tools in TextGrid integrieren

Forschungsprojekte und -institutionen, die Daten in TextGrid archivieren und für Dritte zugänglich und nutzbar machen (Repository)

Form des Einsatzes

Die virtuelle Forschungsumgebung (VFU) TextGrid unterstützt digital arbeitende GeisteswissenschaftlerInnen im gesamten Forschungsprozess – insbesondere beim Erstellen digitaler Editionen.

Sie besteht aus drei Kernbereichen:

- Die Software **TextGrid Laboratory** stellt den Einstiegspunkt in die VFU dar und bietet unterschiedliche Open-Source-Werkzeuge und -Services für den gesamten Forschungsprozess zur Verfügung, z. B. einen Text-Bild-Link Editor für die Verknüpfung von Digitalisaten und Transkriptionen
- Im **TextGrid Repository**, einem Langzeitarchiv für geisteswissenschaftliche Forschungsdaten, können XML / TEI-kodierte Texte, Bilder und Datenbanken sicher gespeichert, publiziert und durchsucht werden.
- Die beständig wachsende **TextGrid Community** trifft sich bei regelmäßigen Nutzertreffen zu themen- bzw. anwendungsspezifischen Workshops, die nicht zuletzt auch den Austausch zwischen digitalen Forschungs- vorhaben aus den Geisteswissenschaften befördern.

Eine Stärke

TextGrid unterstützt den gesamten wissenschaftlichen Arbeitsprozess im Rahmen der Erstellung digitaler Editionen vom Ingest des Ausgangsmaterials (Text- und/ oder Bilddatei- en / Faksimiles) über die Anreicherung und Auszeichnung der Daten (Annotationen, Verknüpfungen) bis zur Veröffentlichung (Portal, Print) und nachhaltigen Archivierung (Repository) und wird stetig basierend auf konkreten fachwissenschaftlichen Anforderungen weiterentwickelt.

Eine Schwäche

Technisch setzt TextGrid auf dem Eclipse-Framework auf, aus heutiger Sicht, wären webbasierte Tools wünschenswerter. Zugleich verdeutlicht dies, dass Softwareentwicklungen permanente Weiterentwicklung benötigen, um sich neuen technologischen aber auch sich wandelnden User-Requirements stellen zu können.

Transcribo (präsentiert durch Thomas Burch, Universität Trier)

Transcribo wird in enger Zusammenarbeit von Philologen und Informatikern der Kooperationspartner entwickelt. Die grafische Nutzeroberfläche ist um das digitale Faksimile, also in der Regel den gescannten Überlieferungsträger, zentriert. Beliebig große Einheiten (z.B. Wörter, Zeilen oder Absätze) können mittels eines Rechteck- oder Polygonwerkzeugs markiert, transkribiert und annotiert werden. Dabei wird jede Bilddatei doppelt dargeboten: links liegt das Original zur Ansicht, die rechte Version dient als Arbeitsunterlage, hier wird der transkribierte Text topografisch exakt über das leicht ausgegraute Faksimile gelegt. Wo die räumliche Anordnung nicht der textuellen Wortreihenfolge entspricht, können Wörter in der grafischen Oberfläche zu Sequenzen zusammengefasst und so die semantischen Zusammenhänge im Transkript protokolliert werden. Ein zentrales Merkmal des Programms liegt außerdem in der Möglichkeit, in jeder erfassten Einheit textgenetische und editionsphilologisch relevante Phänomene zu kennzeichnen und mit Annotationen zu versehen. Dabei kommt ein Kontextmenü mit einer projektspezifischen Auswahl zum Einsatz. Diese umfasst bisher unterschiedliche Varianten von Korrekturen (wie etwa Sofortkorrekturen, Spätkorrekturen mit ein-, zwei- oder mehrfacher Durchstreichung und Überschreibung), die Kennzeichnung von Hervorhebungen sowie von unsicheren Lesungen oder nicht identifizierten Graphen. Diese Auswahl ist jedoch beliebig erweiterbar und wird über den gesamten Projektverlauf hinweg an die Erfordernisse der Textgrundlage angepasst.

Transkribus (präsentiert durch Tobias Hodel, Staatsarchiv Zürich)

Hintergrund

Transkribus ist eine Plattform, die zur automatisierten Erkennung und Annotierung von Texten dient. Sie leistet einerseits eine Verlinkung zwischen Text und Bild (auf Block, Zeilen und Wortebene), produziert andererseits standardisierte Exportformate (XML nach TEI-Standard, PDF, aber auch METS für die Integration in Repositorien). Damit steht eine vollausgerüstete Softwaresuite zur Verfügung, die von der Segmentierung über die Erkennung, Transkription und Edition bis zur Ausgabe alle Schritte in der Herstellung hochwertiger Daten unterstützt.

Die im Projekt READ weiterentwickelte Software vereint somit praxisnah die Bedürfnisse von GeisteswissenschaftlerInnen und Aufbewahrungsinstitutionen mit den technischen Möglichkeiten und Angeboten, die momentan im Bereich der Informatik und Computerlinguistik ermöglicht werden.

Die Software steht in einer stabilen Version zum kostenfreien Download bereit. Das Projekt READ wird unterstützt durch das Horizon 2020 Forschungs- und Innovationsprogramm der Europäischen Union.

Zielpublikum

Aufbewahrungsinstitutionen, die eigene Bestände und Dokumente aufbereiten und zur Verfügung stellen wollen

Geisteswissenschaftlerinnen, die eigene Transkriptionen und Editionen in Transkribus erstellen wollen oder mit darin aufbereiteten Daten arbeiten

Interessierte Laien, die sich an Crowdsourcing-Initiativen beteiligen wollen

ComputerwissenschaftlerInnen, die mit den gewonnenen Daten arbeiten und eigene Algorithmen entwickeln oder verbessern wollen

Form des Einsatzes

Auf Transkribus werden Bilddateien hochgeladen, mit Layoutverlinkungen und Transkriptionen sowie Annotationen versehen. Unterstützt werden die Vorgänge durch Automatisierungsvorgänge im Bereich der Layouterkennung und der Transkription. Der Export der gewonnen Daten ist in unterschiedlichen Formaten möglich. Zusätzlich werden Module zum Crowdsourcing und zukünftig für e-Learning und Analyse mit Smartphone bereitgestellt.

Eine Stärke

Transkribus nutzt neueste Automatisierungsprozesse (u.a. mit rekursiven neuronalen Netzen) somit werden bestmögliche Resultate in Aussicht gestellt.

Eine Schwäche

Transkribus ist eine Expertensoftware und benötigt entsprechende Einarbeitungszeit, um die Dokumente effizient und zielgerichtet zu bearbeiten.

Bibliographie

Andrews, Tara (2015): „Software and Scholarship – Editorial“, in: *Interdisciplinary Science Reviews* 40: 342–348 10.1080/03080188.2016.1165456.

BMBF (Bundesministerium für Bildung und Forschung) (eds.) (2013): *Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften* https://www.bmbf.de/pub/forschungsinfrastrukturen_geistes_und_sozialwissenschaften.pdf .

DARIAH-DE (ed.) (2015): *Handbuch Digital Humanities: Anwendungen, Forschungsdaten und Projekte* http://handbuch.io/w/DH-Handbuch .

Drucker, Johanna (2013): „Performative Materiality and Theoretical Approaches to Interface“, in: *DHQ: Digital Humanities Quarterly* 7 (1) http://digitalhumanities.org:8081/dhq/vol/7/1/000143/000143.html .

Schomaker, Lambert (2016): „Design considerations for a large-scale image-based text search engine in historical manuscript collections“, in: *Information Technology* 58 (2): 80–88 10.1515/itit-2015-0049.