Weitere Tagungsberichte

Vom Bücherrad zum Holodeck. Der Expertenworkshop „Suchtechnologien“ des Forschungsverbunds Marbach Weimar Wolfenbüttel und DARIAH-DE in Weimar

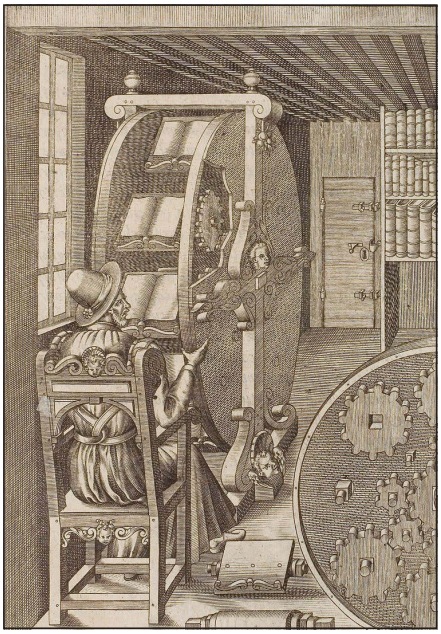

Abb. 1: Bücherrad aus Agostino Ramellis Le diverse et artificiose machine, 1588. Creative Commons Public Domain Mark 1.0 Lizenz.

Die gegenwärtigen Anforderungen an Suchtechnologien bewegen sich in einem Spannungsfeld zwischen spezifischen und generischen Zugängen zu Information. Zugleich gilt: Je umfangreicher und heterogener Informationen vorliegen, desto größer die Herausforderung, eine zielgenaue und effektive Recherche zu ermöglichen.

Wie aber können Suchverfahren den unterschiedlichen Bedarfen einer multidisziplinären wissenschaftlichen Nutzerschaft gerecht werden? Wie können heterogene Datenbestände in eine gemeinsame Suche integriert und somit gemeinsam durchsuchbar gemacht werden? Und weshalb ist die Nutzung von Normdaten ein wichtiger Entwicklungsschritt auf dem Weg zum aus Star Trek bekannten Holodeck? Diesen und weiteren Fragestellungen widmete sich der Workshop „Suchtechnologien” des Forschungsverbunds Marbach Weimar Wolfenbüttel und DARIAH-DE vom 23.-25. Mai 2018.[1] Der Workshop versammelte Expertinnen und Experten aus den Informationswissenschaften, der Informatik und den Digital Humanities sowie Wissenschaftler/innen aus Bibliotheken, Archiven und Museen, um Trends und Potenziale von Suchtechnologien gemeinsam auszuloten und zu diskutieren.

Keynote und Opener verorteten die diskutierten Themenfelder in einem übergreifenden Kontext: Michael Büchner (Deutsche Digitale Bibliothek) zeigte auf, in welche Richtung sich Suchverfahren künftig entwickeln könnten. In seiner Keynote griff er die Vision des Holodecks auf, bei welchem mittels einfacher Sprachbefehle Simulationen und virtuelle Welten gestaltet werden. Dass Lösungswege hin zu effizienten Suchverfahren zu allen Zeiten kreativ und innovativ waren, vermittelte der Beitrag von Swantje Dogunke, Corinna Mayer und Timo Steyer (Forschungsverbund MWW), der sich den historischen Dimensionen des Workshopthemas widmete: So ermöglichte etwa der historische

1 Tagungsprogramm unter <https://vfr.mww-forschung.de/web/suchtechnologien/programm>, Stand: 25.10.2018.

Bücherradkatalog Herzog August des Jüngeren von Braunschweig-Lüneburg (1579-1666), bestehend aus drehbaren Auflageflächen für die handschriftlichen Katalogbände, ein bequemeres Lesen, Annotieren, Wechseln und Vergleichen der bibliographischen Nachweise (Abb. 1).

Context matters – spartenspezifische und spartenübergreifende Suche

Eine Grundproblematik bei der Entwicklung von übergreifenden Suchportalen ist die große Vielfalt und Heterogenität der Datenmengen. Online-Plattformen wie etwa die Deutsche Digitale Bibliothek[2] oder *bavarikon*[3] stehen vor vergleichbaren Herausforderungen: (Meta-)Daten entstehen typischerweise in lokalen Kontexten und genügen sowohl in technischer als auch inhaltlicher Hinsicht spezifischen Anforderungen. Wie aber können Daten aus dem lokalen Anwendungsfall in einen übergreifenden überführt und dort interpretier- und suchbar gemacht werden?

Für Michael Büchner (Deutsche Digitale Bibliothek) ist die Nutzung von Normdaten ein zentraler Baustein auf dem Weg zu einer verbesserten bestandsübergreifenden Suche. In seiner Keynote plädierte er daher für den Ausbau und die konsequente Weiterentwicklung der Gemeinsamen Normdatei (GND). Denn die Verknüpfung mit Normdaten schaffe die Voraussetzung, heterogene Metadaten in einem übergreifenden Kontext zu verorten und deren Qualität und Nachnutzbarkeit somit erheblich zu verbessern.

Abb.2 : Generische Suche von DARIAH-DE, Usecase des Forschungsverbundes MWW

Einen alternativen Ansatz verfolgt die von Tobias Gradl (Universität Bamberg) vorgestellte Datenföderationsarchitektur von DARIAH-DE: Diese ermöglicht übergreifende Suchen über das Mapping

2 Deutsche Digitale Bibliothek, <https://www.deutsche-digitale-bibliothek.de/>, Stand: 25.10.2018.
3 Bavarikon Online Portal, <https://www.bavarikon.de/>, Stand: 25.10.2018.

von verschiedenen Metadatenschemata. Das Verfahren ist insbesondere für die Entwicklung spartenübergreifender Portale interessant, wie der von Swantje Dogunke und Timo Steyer (Forschungsverbund MWW) präsentierte Usecase des Forschungsverbunds Marbach Weimar Wolfenbüttel zeigt: Technisch und inhaltlich heterogenes (Meta-)Datenmaterial aus Archiven, Bibliotheken und Museen kann mithilfe der sogenannten *Generic Search* unter einer Suchoberfläche vereint werden – neben der größeren Sichtbarkeit der Bestände wird dadurch auch eine Optimierung der Datenpflege möglich (Abb. 2).

Einblick in ein besonders spezialisiertes Suchverfahren gewährte der Beitrag „Wir suchen anders! Das Patentinformationszentrum der ULB Darmstadt“ von Rudolf Nickels (ULB Darmstadt): Dort recherchiert man vor allem für Anfragen aus den Bereichen zwischen Wirtschaft und Forschung. Suchen wie eine Neuheitsrecherche oder Recherchen in Bezug auf eine Konkurrenzanmeldung erfolgen auf einer großen Datenbasis – zugleich bestehen hohe Anforderungen an Vollständigkeit und Präzision der Suche (Abb. 3). Es wurde klar: Fachspezifische Suche erfordert Expertise. Darüber hinaus bestehen im Vergleich zu generischen Suchverfahren andere, aber keinesfalls einfachere Anforderungen an das Datenmanagement im Kontext der Patentrecherche.

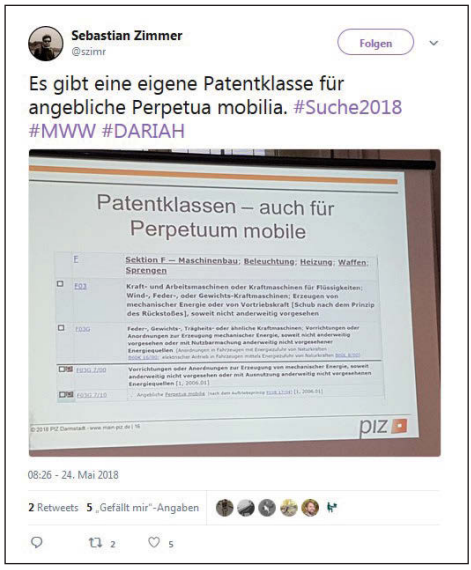

Abb. 3: Sebastian Zimmer (@szimr), 24.Mai 2018: Es gibt eine eigene Patentklasse für angebliche Perpetua mobilia. https://twitter.com/szimr/status/999672921135173632

Die Beiträge der Referentinnen und Referenten in dieser Sektion konnten zeigen, dass für eine auf digitalen Ressourcen basierende Forschung zweierlei notwendig ist, sowohl die Vereinheitlichung als auch die Spezifizierung von Suche. Eine zentrale Herausforderung für Suchverfahren besteht vor allem darin, die fachliche bzw. spartenbezogene Spezifizierung zu ermöglichen und zu bewahren sowie zugleich zu gewährleisten, dass eine Verknüpfung und Nachnutzung der Daten möglich ist, um diese in übergreifenden Kontexten neu zu verorten und auffindbar zu machen.

Neue Zugänge: Browsen und Visualisierung

Durch moderne Suchtechnologien werden innovative Zugangsmethoden zu digitalen Informationen geschaffen, welche über die klassische „einfache“ und „erweiterte“ Suche hinausgehen. Verschiedenste Datenquellen einer oder mehrerer Einrichtungen unter einer Suche zu vereinen, ist mit aktuellen technischen Möglichkeiten grundsätzlich realisierbar. Verschiedene Grundvoraussetzungen mit Blick auf Metadatenschemata, Normierung und Feldbenennungen stellen aber nach wie vor große Herausforderungen dar.

Wie der Beitrag Florian Sepps (Bayerische Staatsbibliothek München) über das Kulturportal *bavarikon* zeigte, machen heterogene Datengrundlagen und unterschiedlichste Nutzungsszenarien die Suche zu einer komplexen Thematik: Die Online-Plattform, welche Kunst- und Kulturschätze aus bayerischen Kultureinrichtungen digital präsentiert, soll sowohl ein wissenschaftliches als auch nicht-wissenschaftliches Publikum ansprechen. Um Inhalte nutzergerecht auffindbar zu machen, werden mehrere Suchmöglichkeiten angeboten: Nicht nur Suchschlitz und thematische Suche sollen künftig einen Einstieg bieten, sondern insbesondere auch kuratierte Zusammenstellungen, etwa in Form virtueller Ausstellungen. Darüber hinaus soll eine stärkere Vernetzung auf Objektebene inhaltliche Zusammenhänge sichtbar machen. Durch Visualisierungen werden Verbindungen zwischen Daten und Einrichtungen sichtbar, die durch eine herkömmliche Suche nicht möglich gewesen wären.

Klassische Suchstrategien gehen jedoch davon aus, dass die Nutzerinnen und Nutzer bereits mindestens ein konkretes Informationsbedürfnis oder Basiswissen mitbringen. Genau das ist beim *VIKUS* Viewer[4] nicht notwendig, wie Katrin Glinka (Staatliche Museen zu Berlin) in ihrem Vortrag „Den Suchschlitz überwinden. Visualisierungen als Zugang zu digitalisiertem Kulturgut" vorstellte. *VIKUS* ermöglicht ein „Finden ohne Suche", sodass kein Fachvokabular von Nutzerinnen und Nutzern benötigt wird. Als Sucheinstieg wurde der Weg der Visualisierung gewählt, sodass Zusammenhänge zwischen den Daten computergestützt sichtbar gemacht werden können. Abstrakte Datenstrukturen werden so in eine visuelle Form überführt. Die Analyse und Bereinigung der Daten nimmt einen großen Stellenwert ein, bevor mit der Visualisierung begonnen werden kann, die immer bestandsgerecht erfolgen muss.[5]

Dies gilt auch für den *Catalogue of Digital Editions*[6], vorgestellt von Greta Franzini (Universität Göttingen) und Peter Andorfer (Österreichische Akademie der Wissenschaften), welcher digitale Editionen nach normierten Kriterien verzeichnet. So sind über vernetzte Daten und Visualisierungen neuartige Zugänge zu den Informationen der digitalen Editionen möglich. Diese Beispiele zeigen, dass neben dem klassischen Suchschlitz weitere innovative computergestützte Verfahren Vorteile bieten und von immer mehr Einrichtungen und Projekten favorisiert werden. Sowohl nicht-wissenschaftliche als auch wissenschaftliche Anliegen können von neuen Zugangsformen profitieren: Auf der einen Seite wird kein Fachwissen mehr für eine Suchanfrage benötigt, andererseits können so im wissenschaftlichen Kontext – durch die Verlinkung von Datenquellen und Visualisierungen – Verbindungen hergestellt werden, die auf herkömmlichem Weg nicht möglich gewesen wären.

Qualität trotz Quantität – Potenziale der Automatisierung

Im Rahmen des Workshops wurden weiterhin Möglichkeiten und Potenziale der Automatisierung im Kontext von Suchtechnologien vorgestellt: Im Zentrum der Diskussion stand dabei die Frage nach der Gewährleistung von Qualität im Vergleich zu intellektuellen Verfahren. So referierte Elisa

4 VIKUS Viewer, <https://uclab.fh-potsdam.de/vikus/>, Stand: 25.10.2018.
5 Demo VIKUS Viewer "Past Visions"- Zeichnungen Friedrich Wilhelms IV. von Preußen: <https://uclab.fh-potsdam.de/vikus/content/1-projects/3-pastvisions/pastvisions.gif>, Stand: 25.10.2018.
6 Catalogue of Digital Editions, <https://dig-ed-cat.acdh.oeaw.ac.at/>, Stand: 25.10.2018.

Herrmann (Herzog August Bibliothek Wolfenbüttel) über die Qualitätsfrage bei der automatischen Texterkennung: Während herkömmliche OCR-Verfahren bei Texten des 20. Jh. in der Regel eine zufriedenstellende Textgenauigkeit erzielen, erreichen sie bei älteren Vorlagen nur selten die gewünschte Qualität. Lösungsansätze werden im Rahmen des Projekts *OCR-D*[7] entwickelt und reichen von der Prüfung neuer computerlinguistischer Verfahren über die Automatisierung von Qualitätsprüfungen bis hin zur pragmatischen Neudefinition von Qualität: Denn für bestimmte wissenschaftliche Nutzungsszenarien, etwa einer groben Stichwortsuche, sind auch Texte mit einer Genauigkeit von 85% ausreichend nachzunutzen.

Wie wichtig der Zugang zu größeren Textkorpora für die digitale geisteswissenschaftliche Forschung ist, zeigte der Beitrag von Gerhard Heyer und Christian Kahmann (Universität Leipzig), der sich mit einem explorativen Suchverfahren befasste: Dieses zielt im Gegensatz zur herkömmlichen Suche, die zumeist von einem vordefinierten Informationsbedarf ausgeht, auf ein Auffinden und Identifizieren auffälliger Themen, Konzepte, Akteure und Ereignisse. Anhand einer geeigneten Datengrundlage – hier etwa das digitale Archiv der Zeitschrift Guardian – kann mittels Text-Mining beispielsweise der Bedeutungswandel bestimmter Themenfelder über einen längeren Zeitraum hinweg untersucht werden.

Mit den Möglichkeiten der automatisierten Suche in digitalisierten Bildbeständen beschäftigte sich der Beitrag Rosa Riccis (Universität Bamberg): Ausgangspunkt hierzu bildete Riccis Untersuchung ähnlicher Bildsegmente in Emblem-Picturas des 17. Jahrhunderts. Bei dieser Form der visuellen Suche wird versucht, die essentielle Struktur eines Ausgangsbildes zu erfassen, indem ausschlaggebende Bildmerkmale (sog. *features*) extrahiert und mithilfe von Deskriptoren klassifiziert werden. Auf diese Weise ist es möglich, in Bildbeständen nach ähnlichen Motiven und Inhalten zu suchen – die Ähnlichkeitssuche birgt in diesem Sinne großes Potenzial für die automatisierte Suche und Analyse von digitalem Bildmaterial.

Einen weiteren Aspekt der Automatisierung in Bezug auf Suchverfahren, den der automatisierten Erschließung, stellte Elisabeth Mödden (Deutsche Nationalbibliothek) vor: Die DNB setzt bei der Verschlagwortung von Publikationen zunehmend auf computerlinguistische Verfahren. Deren Vorteile sind: Der Aufwand einer intellektuellen Sacherschließung wird verringert, Erschließungslücken angesichts wachsender Datenmengen können geschlossen werden. Argumente aus der derzeitigen Diskussion über die Risiken eines Qualitätsverlusts durch automatische Erschließungsverfahren werden, so Mödden, aufgenommen und geben Impulse für die Weiterentwicklung dieser Techniken. Automatisierte Verfahren der Verfügbarmachung von Inhalten im Netz, wie Texterkennung oder automatisierte Erschließung digitaler Inhalte, sind zunächst zur Bewältigung wachsender Datenmengen wichtig. Die Beiträge von Ricci, Kahmann und Heyer zeigten darüber hinaus auch: Sie sind Bedingung für das Forschen mit digitalen Ressourcen, welches auf maschinell verarbeitbare Inhalte angewiesen ist.

7 Projekt OCR-D, <http://ocr-d.de/>, Stand: 25.10.2018.

Analyse der Suchergebnisse

Den zweiten Workshoptag leitete Manuel Burghardt (Universität Leipzig) mit seinem Beitrag „Music Information Retrieval meets Digital Humanities“ ein. Der Begriff des Distant Reading ist in den Digital Humanities mittlerweile etabliert, doch gibt es auch Distant Hearing? Burghardt näherte sich dieser Fragestellung über eine verwandte Domäne, dem Music Information Retrieval, an. Untersucht wurden Volksliedblätter, dabei speziell die Symbole (Noten). Im Ergebnis ist es mit den erhobenen Daten möglich, eine Suche und weiterführende Untersuchung von Melodien, Intervallen und Parson-Code durchzuführen, sodass beispielsweise ein Vergleich zwischen Melodic Similarity und Notengleichheit der einzelnen Liedblätter oder die Erstellung eines Melodic-Similarity-Netzwerks möglich sind. Doch um zu diesem Punkt zu gelangen, waren umfangreiche Vorarbeiten notwendig: Eine Digitalisierung von 140.000 Liedblättern musste erfolgen, Metadaten waren über Zettelkästen der Regensburger Universitätsbibliothek vorhanden. Für die Erschließung entschied man sich nach einer Betrachtung von Optical Music Recognition und einer unzureichenden Erkennungsrate von 4-36% für einen Crowdsourcing-Ansatz hinsichtlich der Notentranskription. Dafür wurde das Tool *Allegro*[8] entwickelt, welches es auch Musik-Laien möglich macht, intuitiv die Transkription der Noten vorzunehmen. Zunächst erfolgt im Tool die Segmentierung in Takte durch die User, dann werden Metadaten (Signatur, Tonart, Rhythmus) ergänzt und eine taktweise Transkription erfolgt (Abb. 4). Zur Qualitätssicherung wird ein semi-automatischer Double-Keying-Check durchgeführt.

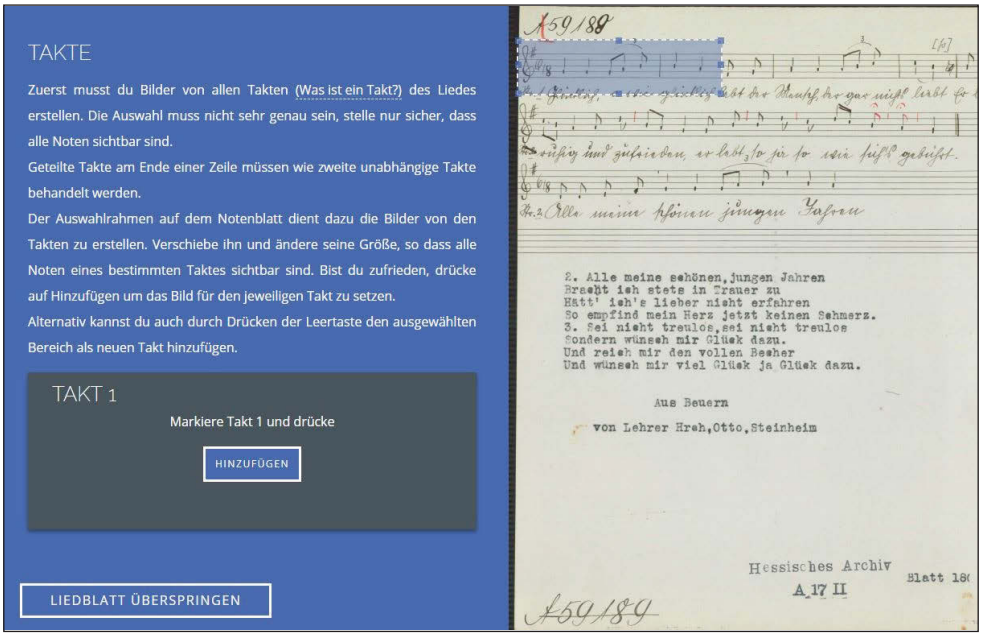

Abb. 4: Screenshot des Transkriptionsmodus für Takte und Noten im Tool Allegro

8 Allegro, <http://138.68.106.29>, Stand: 25.10.2018.

Auch im Projekt „Niklas Luhmann – Theorie als Passion. Wissenschaftliche Erschließung und Edition des Nachlasses“, vorgestellt von Johannes Schmidt (Universität Bielefeld) und Sebastian Zimmer (Cologne Center for eHumanities CCeH, Universität Köln), sind aufwändige Vorarbeiten notwendig, bis die angedachte Volltextsuche und Visualisierungen mit Graphen, Diagrammen und Branch-Views umgesetzt werden können. Luhmanns (1927 - 1998) umfangreicher Zettelkasten wurde in einem ersten Schritt digitalisiert und erschlossen, weiterhin wurde ein Schlagwort- und Personenregister erstellt. Erst dann kann die Fülle der Daten in Suche und Analyse weiterverwendet werden.

Die beiden Projekte zeigten: Ohne maschinell verarbeitbare Daten können keine computergestützten Methoden der Digital Humanities angewandt werden und die Erschließung und Bereitstellung von Daten nimmt im digitalen Umfeld einen signifikanteren Stellenwert ein als in der analogen Erschließung, da oft mehr statt weniger Schritte erfolgen müssen, um zu einem auswertbaren Ergebnis zu gelangen.

Doch nicht nur die Aufbereitung von analogen Materialien für digitale Zwecke ist eine große Herausforderung, denn das Novum der Born-Digital-Nachlässe, die mittlerweile Kultureinrichtungen erreichen, fügt noch eine weitere Komplexitätsebene hinzu: Zwar liegen, wie Heinz-Werner Kramski (Deutsches Literaturarchiv Marbach) in seinem Vortrag „Suche in born-digital-Nachlässen: Das Beispiel Kittler“ näher erläuterte, bereits digitale Daten vor, diese stellen die archivalische Erschließung jedoch vor gänzlich neue Herausforderungen: Zum einem die große Datenmenge (über 1,7 Millionen Dateien im Nachlass Kittler), zum anderen die Angewohnheit Kittlers, mit der sog. root-Berechtigung zu speichern, so dass Daten und Anwendungen über das gesamte Dateisystem hinweg abgelegt sind und die Identifizierung archivwürdiger Daten somit enorm erschwert wird. Zur Erschließung des umfangreichen digitalen Bestandes wurde die Eigenentwicklung *Indexer* verwendet, die mittels Mechanismen wie Dateiformaterkennung und Prüfsummenmatching die Dateien mit Metadaten versieht. Eine Suche ist mittels *Apache SOLR* realisiert und ermöglicht in der Theorie eine Volltextrecherche über alle Dateien – jedoch bedeutet das, dass der Bestand für Nutzer/innen aktuell nicht zugänglich ist, da noch nicht klar ist, welche Daten privater Natur und damit für die Benutzung gesperrt sind.

Die vorgestellten Beiträge zeigen auf, dass computergestützte Verfahren kein „Wundermittel“ sind, mit dem sich das angestrebte Ziel leichter erreichen lässt: Will man analog vorliegendes Material mit innovativen Suchmöglichkeiten ausstatten, sind aufwändige Vorarbeiten notwendig, um die Datengrundlagen anzureichern. Und liegen Daten bereits digital vor, gibt es neue Herausforderungen, um der Masse an Daten Herr werden zu können. Ist dies jedoch umfangreich geschehen, können moderne Technologien ein neues Erleben und Gestalten von Such- und Analyseprozessen ermöglichen und so zu einem erweiterten Erkenntnisgewinn beitragen.

Fazit

Der Workshop zeigte die Potenziale moderner Suchtechnologien im Hinblick auf integrative Suchen über heterogene (Daten)Sammlungen auf. Angesichts der stetig wachsenden Zahl von Digitalisaten, von Forschungsportalen sowie der Publikation von Forschungsdaten kommt den Metasuchen eine

immer wichtigere Bedeutung im Forschungszyklus zu – denn nur was gefunden werden kann, ist für die Wissenschaft auch nutzbar. Das Spannungsfeld von domänenspezifischen und globalen Suchen kann dabei nicht nur technisch gelöst werden, sondern ist ebenso eine Frage der Visualisierung und der Zugangsgestaltung. Gerade in der Kombination von domänenspezifischem Wissen und aktuellen Forschungsfragen liegt die Stärke von Suchtechnologien in den Digital Humanities und ein Alleinstellungsmerkmal gegenüber kommerziellen Anbietern.

Die Folien zu den einzelnen Vorträgen können im Virtuellen Forschungsraum des MWW Forschungsverbundes eingesehen werden.[9]

Elena Luz, Klassik Stiftung Weimar
Corinna Mayer, Deutsches Literaturarchiv Marbach
Timo Steyer, Herzog August Bibliothek Wolfenbüttel

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/2018H4S287-294

9 Beiträge des Expertenworkshops „Suchtechnologien" im LAB im VFR des MWW Forschungsverbundes, <https://vfr.mww-forschung.de/web/suchtechnologien/vortraege>, Stand: 25.10.2018.

