



# Informationen – Digital verpackt

22. – 23. September 2020 / Zentralbibliothek Zürich

# Tagungsprogramm

## Dienstag, 22. September 2020

	Referenten	Titel
10:30 Uhr	<b>Christian Oesterheld,</b> <b>Organisatorinnen</b>	Begrüßung
10:45 Uhr	<b>Lena Oetjens, Luzia</b> <b>Roth, Katja Staub</b>	Einführung
11:15 Uhr	<b>Peter Bell</b>	Vom Bild zu tausend Worten. Erschließung historischer Bildwerke und Computer Vision
12:00 Uhr	<i>Mittag</i>	
13:15 Uhr	<b>Tobias Hodel</b>	Neue Zugriffe auf Handschriften und alte Drucke: Grosse Korpora durchsuchen
14:00 Uhr	<b>René Schneider,</b> <b>Julien Raemy</b>	Auspacken, Identifizieren, Anpacken: Überlegungen zur Konvergenz von Linked Data und persistenter Identifikation
14:45 Uhr	<i>Pause</i>	
15:15 Uhr	<b>Patrick Dinger</b>	Ein Portal, viele Partner - Herausforderungen beim Aufbau eines bestandsübergreifenden Zeitungsportals
16:00 Uhr	<b>Gerold Schneider</b>	Quantitative Ideengeschichte: Zeitungen, Parlamentsdebatten und medizinische Forschung vergangener Jahrhunderte
16:45 Uhr	<i>Pause</i>	
17:15 Uhr	<b>Torsten Reimer</b> Library Science Talk	The perfect storm - challenges for research libraries in the coming decade

## Mittwoch, 23. September 2020

9:00 Uhr	<b>Christof Schöch,</b> <b>Maria Hinzmann</b>	Mining and Modeling Text: Informationsextraktion und Linked Open Data für die Literaturgeschichtsschreibung
9:45 Uhr	<b>Regine Tobias</b>	Open Access und Open Science - lokale Services für Forschung und Lehre und die Rolle der Bibliothek am KIT Karlsruhe
10:30 Uhr	<i>Pause</i>	
11:00 Uhr	<b>Cyrus Beck</b>	Das Recht der digitalen Daten - überall und nirgends?
11:45 Uhr	<b>Noémie Ammann</b>	Metadatenverarbeitung im Discovery-System von der Stange
12:30 Uhr	<i>Mittag</i>	
14:00 Uhr	<b>Michael Piotrowski</b>	Ich weiss nicht, was soll es bedeuten ... Überlegungen zur Modellierung historiografischer Unsicherheit
14:45 Uhr	<b>Noah Bubenhofer</b>	Listen, Karten, Vektoren, Bäume, Partituren: Visualisierungen in der Linguistik
15:30 Uhr	Abschlussdiskussion	

## Abstracts

### **Peter Bell (Universität Erlangen): Vom Bild zu tausend Worten. Erschließung historischer Bildwerke und Computer Vision**

Deep Learning, Data Science und AI sind zu Schlagworten ganzer Industrien geworden und werden auch in den Digital Humanities diskutiert. Im Bildbereich hat der Einsatz dieser Werkzeuge aber oft noch zu sehr den Charakter von Fallstudien und Prototypen, die oft im universitären Bereich verbleiben. Der Vortrag soll daher zu einer stärkeren Etablierung von Computer Vision Verfahren im Bereich der Erschließung von Bilddaten anregen. Es geht dabei nicht nur um eine effizientere Annotation und Vernetzung, sondern um eine tiefere Erschließung von Bildern und Bilddetails. Maschinelle Bildverarbeitung bietet einen Zugang komplementär zum Text, kann aber auch zur Erzeugung von Tags und Beschreibungen genutzt werden, während vorhandene Metadaten die visuelle Erschließung verbessern können. Im Vortrag werden verschiedene Ansätze für die Interaktion von Mensch und Maschine sowie Bild und Text vorgestellt, um sie mit Bibliotheken und GLAM-Institutionen insgesamt zu diskutieren. Dabei wird auch die besondere Herausforderung historischer Bildkulturen (z.B. Buchmalerei oder Druckgrafik) berücksichtigt.

### **Tobias Hodel (Universität Bern): Neue Zugriffe auf Handschriften und alte Drucke: Grosse Korpora durchsuchen**

Seit einigen Jahren digitalisieren Bibliotheken und Archive ihre Bestände und Sammlungen professionell und im grossen Stil. Standards zur Bildaufnahme und -publikation sowie für Metadaten setzen sich allmählich durch, sodass der Nutzung nichts mehr im Wege zu stehen scheint. Um die meist textuellen Dokumente aufzubereiten und vor allem zugänglich zu machen, bedarf es jedoch noch der Texterkennung. Ein Problem das erst für Drucke, die nach 1830 produziert wurden, seriell funktioniert. Für Drucke der Frühen Neuzeit und Handschriften im Allgemeinen besteht seit dem *machine learning turn* die Möglichkeit, spezifische Erkennungsmodelle zu trainieren, die grosse Masse wird von spezialisierten Modellen jedoch ausgenommen. Eine erfolgsversprechende Form des Zugriffs ermöglicht die spezifische Indexierung und das Durchsuchen von Variantenlesungen (sogenanntes Keyword Spotting). Die beiden Ansätze führen zu einem Volltextzugriff, der rasch Einblicke erlaubt ohne vorgängige perfekte Erkennung oder Korrektur. Obwohl die Verfahren erst in Prototypen verfügbar sind, ist absehbar wie in Zukunft eine Vielzahl von Beständen und Sammlungen zugänglich gemacht werden kann.

Anhand von Beispielen aus dem 10., 16. und 19. Jahrhundert demonstriert der Beitrag, wie Keyword Spotting eingesetzt werden kann, um heterogene Korpora für unterschiedliche Abfragen zu nutzen. Dazu wird die Transkribus-Plattform (transkribus.eu) genutzt, die unterschiedliche Handwritten Text Recognition Engines verbindet. Ergänzt wird die Demonstration mit der Beschreibung, wie die Erkennung technisch und theoretisch funktioniert. Die Forschungen entstanden im Rahmen eines Infrastrukturprojekts, das durch die EU ermöglicht wurde und zur Überarbeitung der Transkribus-Infrastruktur führte. Die READ Kooperative (readcoop.eu) pflegt und erweitert aktuell diese Infrastruktur.

**René Schneider, Julien Raemy (Haute école de gestion de Genève): Auspacken, Identifizieren, Anpacken: Überlegungen zur Konvergenz von Linked Data und persistenter Identifikation**

Von Linked Data sprechen alle. Von persistenten Identifikatoren wenige. Dahinter verstecken sich unterschiedliche aber auch verwandte Konzepte, die sich alle mit der Frage nach der Permanenz des Zugangs zu digitalen Ressourcen beschäftigen.

Im Vortrag sollen die Vor- und Nachteile insbesondere aber die Möglichkeiten einer Kombination oder sogar Konvergenz der Ansätze diskutiert werden. Dazu stellen wir existierende und in Planung befindliche Umsetzungen aus den Bereichen der Informations- und Geisteswissenschaften vor.

**Patrick Dinger (Deutsche Digitale Bibliothek Frankfurt a.M.): Ein Portal, viele Partner – Herausforderungen beim Aufbau eines bestandsübergreifenden Zeitungsportals**

Historische Zeitungen wurden in den letzten Jahren verstärkt digitalisiert und zugänglich gemacht. Um perspektivisch alle in Deutschland vorhandenen digitalisierten historischen Zeitungen in einem Portal zusammenzuführen, wird seit Anfang 2019 ein nationales Zeitungsportale auf Basis der technischen und organisatorischen Infrastruktur der Deutschen Digitalen Bibliothek (DDB) aufgebaut. Damit dies gelingen kann, haben sich die Deutsche Nationalbibliothek, die Staatsbibliothek zu Berlin, die SLUB Dresden und FIZ Karlsruhe zusammengetan und den zugrundeliegenden DFG-Förderantrag gemeinsam entwickelt. Neben der Kooperation im Aufbau des Portals ist die DDB aber auch auf ein Netzwerk an Partnerorganisationen angewiesen, die Zeitungsbestände bereitstellen und – entsprechend der Datenformate der DDB – aufbereiten. Durch die Konsolidierung der Daten verschiedener Kultureinrichtungen wird so eine gemeinsame Basis geschaffen, um das Portal für Methoden der Digital Humanities und moderne Präsentationsformate wie IIIF zu öffnen.

Im Rahmen des Vortrags werden am Beispiel des Zeitungsportals die unterschiedlichen Ebenen der Kooperation – von der Konzeption über rechtliche Fragen bis hin zur Verwendung von standardisierten Metadatenprofilen – beleuchtet, die es ermöglichen, dezentral erschlossene Bestände über ein Portal zu präsentieren und für die Analyse bereitzustellen. Es wird dabei auch der Frage nachgegangen, welche Funktion wissenschaftliche Bibliotheken heute und in Zukunft in diesem Prozess übernehmen könnten.

**Gerold Schneider (Universität Zürich): Quantitative Ideengeschichte: Zeitungen, Parlamentsdebatten und medizinische Forschung vergangener Jahrhunderte**

Wir stellen drei Studien vor, die datengetriebene Methoden aus Computerlinguistik und Digital Humanities verwenden, um Textsammlungen automatisch zu analysieren und Entwicklungen der Geschichte, Kultur und des menschlichen Denkens aufzuzeigen. Wir verwenden Methoden wie Kollokationen, Dokumentenklassifikation, Topic Modelling, Sentiment Detection und automatisch erstellte konzeptuelle Karten.

Erstens verfolgen wir die Ideengeschichte der medizinischen Forschung von 1500 bis jetzt, und zeigen den Wandel von Scholastik zu empirischer Forschung und klinischen Studien auf. Wir illustrieren, wie der Wandel zur Empirie stufenweise und über einen langen Zeitraum erfolgte.

Zweitens untersuchen wir die Debatten im britischen Parlament: wie hat sich die Sprache verändert, gibt es Unterschiede zwischen dem House of Lords und House of Commons? Welche Zusammenhänge zwischen Sprachgebrauch und Demokratisierung bestehen?

Drittens stellen wir das Zeitungskorpus und Projekt Impresso vor, in dem wir u.a. schweizerische Zeitungen über mehrere Jahrhunderte sammeln und verarbeiten. Wir visualisieren die Entwicklung des 2. Weltkrieges, und zeigen, wie Europa und Krise miteinander verbunden sind.

### **Torsten Reimer (British Library London): The perfect storm – challenges for research libraries in the coming decade**

Seit Jahren wird immer wieder die Frage diskutiert, ob Bibliotheken in einer digitalen Welt noch eine Zukunft haben. Inzwischen sollte allerdings auch Skeptikern klar geworden sein, dass weder das gedruckte Buch noch die Bibliothek auf absehbare Zeit verschwinden werden. Interessanter und relevanter ist dagegen die Frage, wie die Zukunft von Bibliotheken aussehen könnte. In meinem Vortrag möchte ich eine Reihe der Herausforderungen diskutieren, denen wir uns in dieser Dekade stellen müssen. Ein zentrales Thema wird dabei die Frage sein, ob Forschungsbibliotheken tatsächlich in der Lage sind, in der Open Science Welt eine zentrale Rolle zu spielen, und ob und wie wir in einem Umfeld digitaler Marktkonzentration noch die Kontrolle über unsere Daten behalten können. Durch die Open Science-Bewegung verlagern sich das Finden und der Zugriff auf Informationen immer mehr aus dem Bibliotheksraum in externe digitale Dienste – Dienste, bei denen das Risiko besteht, dass sie von Bibliotheken zunehmend weder verstanden noch kontrolliert werden können. Das wirft wichtige Fragen hinsichtlich von Transparenz und Vertrauen auf, zwei Themen, die zentral für das Selbst- und Nutzerverständnis von Bibliotheken sind. In meinem Vortrag werde ich diese und andere Fragen zur digitalen Transformation diskutieren, verbunden mit einem Bericht über meine aktuelle Arbeit an der Umsetzung einer Zehnjahresstrategie für den digitalen Wandel im Verbund der britischen Forschungsbibliotheken.

### **Christoph Schöch, Maria Hinzmann (Universität Trier): Mining and Modeling Text: Informationsextraktion und Linked Open Data für die Literaturgeschichtsschreibung**

Das Vorhaben "Mining and Modeling Text" (<https://mimotext.uni-trier.de>) setzt quantitative Verfahren der Informationsextraktion aus Texten und die Repräsentation von Wissensbeständen in Form von Linked Open Data für eine digitale Literaturgeschichtsschreibung ein. Das Ziel des Vorhabens ist, ein mit anderen Ressourcen des Semantic Web verknüpfbares Wissensnetzwerk aufzubauen, das sich aus drei wesentlichen Informationsquellen speist, die jeweils für die Informationsextraktion genutzt werden: Metadaten aus bibliographischen Nachweissystemen, inhaltliche und formale Eigenschaften der Primärliteratur sowie Aussagen der literaturgeschichtlichen Fachliteratur. Erste Anwendungsdomäne des Vorhabens ist dabei der französische Romane der zweiten Hälfte des 18. Jahrhunderts.

Die Modellierung des Netzwerks als Linked Open Data ermöglicht die Verknüpfung der verschiedenartig extrahierten Informationen und setzt voraus, dass (beschreibende und wertende) Aussagetypen über literaturwissenschaftlich relevante Entitäten eingegrenzt werden. In einem ersten Schritt hat sich für die Informationsextraktion aus der Fachliteratur die Erkennung von Personen, Orten und Werken als zentral erweisen. Der Beitrag geht exemplarisch auf die Modellierung der Verknüpfungen zwischen den aus der Fachliteratur extrahierten Informationen mit den bibliographischen Metadaten ein. Dabei wird deutlich, wie wichtig die Verfügbarkeit reichhaltiger Metadaten ist. Zudem zeigen sich die Herausforderungen, die sich aus der Verschränkung von quantitativen Verfahren der Informationsextraktion mit einer semantisch orientierten Datenmodellierung für die Domäne der Literaturgeschichtsschreibung ergeben.

## **Regine Tobias (Karlsruher Institut für Technologie): Open Access und Open Science – lokale Services für Forschung und Lehre und die Rolle der Bibliothek am KIT Karlsruhe**

Der Vortrag gibt Einblick in die strategischen Konzepte und operativen Schritte des Ausbaus des institutionellen Repositorys KITopen zu einem umfassenden Forschungs- und Lehrservice für Wissenschaftlerinnen und Wissenschaftler des KIT.

Open Access und Open Science bilden in der KIT-Bibliothek seit vielen Jahren das strategische Leitmotiv beim Ausbau und der Weiterentwicklung ihrer grundlegenden Services für Forschung und Lehre. Im Zentrum dabei steht das institutionelle Repository KITopen, das neben dem erfolgreichen Open-Access-Verlag KIT Scientific Publishing eine der zentralen Säulen der Publikations-Infrastruktur bildet. Im Vortrag werden die operativen Schritte des Ausbaus erläutert, die mit der Abstimmung KIT-weiter Policies im Bereich von Forschung und Lehre einhergehen. Aus einem anfänglich reinen Open-Access-Dokumentenserver entwickelte sich in einem ersten Schritt ein professionelles Tool für Publikationsmanagement. Damit war der Grundstein gelegt, um als zentrales Basissystem eines modularen Forschungsinformationssystems am KIT entsprechende Berichtsfunktionalitäten und forschungsnahe Mehrwertservices bereit zu stellen. Dann folgte der Ausbau in Richtung von Open Science und ermöglichte die Publikation von Forschungsdaten nach den FAIR-Prinzipien. Als jüngste Entwicklung erweitert KITopen seinen Fokus von einem Publikations-Repository für Volltexte und Forschungsdaten hin zur temporären und lernraumbezogenen Unterstützung von Dozierenden bei der Bereitstellung digitaler Lerninhalte und wird dadurch zu einem Baustein der digitalen Lehre am KIT.

Neben dem permanenten Ausbau ihrer Repository-Infrastruktur liegt der Fokus auf Beratungsleistungen rund um das wissenschaftliche Publizieren, das Forschungsdatenmanagement und nun verstärkt auch auf den Bereich der digitalen Lehrmaterialien, bei deren Erstellung Dozierende in der Regel mit den gleichen Fragestellungen konfrontiert werden wie bei ihrer Forschungstätigkeit. Die wissenschaftspolitische Zielsetzung der Bibliotheksservices dreht sich dabei im Kern immer um die Unterstützung bei der bestmöglichen Wahrung der Autorenrechte, sowohl im Umgang mit Urheberrecht und Datenschutz, als auch bei Fragen rund um Lizenzierungen für die Nachnutzung eigener und fremder Lehrmaterialien.

## **Cyrus Beck (Zentralbibliothek Zürich): Das Recht der digitalen Daten – überall und nirgends?**

Der Vortrag beleuchtet den Blick auf digitale Daten in unterschiedlichen Rechtsbereichen. Die Spannweite dieses Streifzugs reicht dabei mit einigen Zwischentritten von der alten Rechtsmaterie des Sachenrechts über das Straf-, Urheber- und Lauterkeitsrecht bis hin zum jungen Datenschutz- und jüngsten Blockchainrecht. Jeweils wird dabei der technische in Verbindung mit dem rechtlichen Blickwinkel auf die Daten untersucht, so etwa die Semantik der oder den Umgang mit digitalen Daten. Dabei zeigt sich, dass die immense Bedeutung der digitalen Daten in der alltäglichen und wirtschaftlichen Lebenswelt in der rechtlichen Welt zwar vielseitig rezipiert wird («überall»), aber letztlich eine Randerscheinung ist und es auch keine (einheitliche) rechtliche Definition von digitalen Daten gibt («nirgends»). Als Schlussfolgerung werden Überlegungen *de lege ferenda* angestellt: Wäre ein neues Rechtsinstitut des «digitalen Datums» überhaupt zweckdienlich? Der Vortrag bezieht sich insbesondere auf das schweizerische Recht, weitet den Blick aber auch über den Tellerrand hinaus auf weitere Rechtsordnungen.

### **Noémie Ammann (ETH-Bibliothek Zürich): Metadatenverarbeitung im Discovery-System von der Stange**

In wissenschaftlichen Bibliotheken sind die Aufgaben und Themen im Zusammenhang mit der Verarbeitung von Metadaten durch die technische Entwicklung gegenüber früheren Zeiten wesentlich komplexer geworden. Immer mehr Datenformate, Schnittstellen und Systeme, mit denen Metadaten verwaltet werden, werden gleichzeitig benutzt und eingesetzt. Die in wissenschaftlichen Bibliotheken als Recherchertools für die Benutzenden breit eingesetzten Discovery-Systeme sind davon direkt betroffen. Als One-Stop-Shops konzipiert, sollen sie den Benutzenden einen möglichst einheitlichen und nahtlosen Zugang zu allen Bibliotheksressourcen gewähren. Doch während wissenschaftliche Bibliotheken in anderen Belangen auf innovative Tools oder sogar Eigenentwicklungen setzen, greifen sie bei den Discovery-Systemen in der Regel zu den gängigen Produkten von der Stange, wobei dem jüngsten Trend zufolge kein eigenständiges System mehr eingekauft wird, sondern das Discovery-System bevorzugt als Komponente in der Produktpalette der Library Service Platform gewählt wird.

Die Metadatenverarbeitung verläuft in solchen Discovery-Systemen von der Stange naturgemäss in vorgegebenen Bahnen. Gibt es dennoch Spielraum für wissenschaftliche Bibliotheken, mit den Metadaten innovative und massgeschneiderte Lösungen zu entwickeln? Und was sind die Folgen, wenn wissenschaftliche Bibliotheken nicht nur den Systembetrieb von Discovery-Systemen an den Hersteller out-sourcen, sondern auch die Metadatenverarbeitung weitgehend auslagern, indem die Daten via einen kommerziellen Megaindex zur Verfügung gestellt werden?

### **Michael Piotrowski (Universität de Lausanne): Ich weiss nicht, was soll es bedeuten... Überlegungen zur Modellierung historiografischer Unsicherheit**

Historische Informationen sind offensichtlich stets mit Unsicherheit behaftet. Selbst wenn es noch Zeitzeugen gibt, ist ihre Erinnerung nicht unbedingt verlässlich und ausserdem subjektiv gefärbt. In schriftlichen Quellen stellt sich – neben Fragen der Echtheit, Datierung, Lesbarkeit, Vollständigkeit usw. – die Frage, worauf sich etwa Bezeichnungen von Personen, Orten oder auch Titeln beziehen: beziehen sich zwei Vorkommen eines Namens auf ein und dieselbe Person oder auf zwei verschiedene Personen? Welchen Ort bezeichnet ein Orts- oder Flurname? Welches Amt war mit einem bestimmten Titel verbunden?

Wenn man diese Informationen «digital verpackt», muss man die Unsicherheit mit einpacken – ansonsten erscheinen Informationen als «genau» und «gesichert», die es in Tat und Wahrheit nicht sind. Aber wie? Dies ist eine der zentralen Forschungsfragen der Digital Humanities.

In einigen Fällen, etwa bei Orts- und Datumsangaben, scheint die Lösung einfach: statt Punktkoordinaten gibt man beispielsweise ein Rechteck an und unsichere Daten versteht man als eine Spanne zwischen dem frühesten und dem spätesten möglichen Datum.

In der Geschichtswissenschaft geht es allerdings nicht primär darum, exakte Karten oder Zeitstrahlen zu zeichnen, auch wenn diese zweifellos nützlich sein können. In der Historiografie geht es letztlich um Kausalität.

In diesem Vortrag möchte ich über Unsicherheit auf dieser Ebene nachdenken, also über historiografische statt historische Unsicherheit.

## **Noah Bubenhofer (Universität Zürich): Listen, Karten, Vektoren, Bäume, Partituren: Visualisierungen in der Linguistik**

Diagramme spielen eine wichtige Rolle in der linguistischen Forschung: Einerseits visualisieren sie Ergebnisse von quantitativen Analysen auf effiziente Art und Weise mit Balken- und Liniendiagrammen, Plots, Netzen etc. Andererseits dienen sie der Exploration von grossen Datenmengen, um diese überhaupt analysierbar zu machen. Insofern ähneln sie eher einem Mikroskop als einem Kommunikationsmittel.

Aus einer diagrammatischen Sicht sind die Basis solcher Visualisierungen diagrammatische Grundfiguren, die Daten transformieren und formen. Im Falle der Linguistik sind es sprachliche Daten, die durch diese Grundfiguren transformiert werden. So werden nicht nur neue Einsichten in die Daten möglich, sondern es entstehen überhaupt erst Untersuchungsgegenstände.

In meinem Vortrag möchte ich zeigen, dass so verstandene Visualisierungen mehr sind als hübsche Bilder am Ende eines Analyseprozesses und die Grenze zwischen Visualisierung und Verdichtung fließend ist.