KIM Workshop 2019

Wann: 2. und 3. April 2019

Wo: Universität Mannheim, Fuchs-Petrolub-Festsaal (O 138), Anfahrt

Veranstalter: DINI-AG KIM, UB Mannheim

Teilnahmegebühr: 40 Euro

TeilnehmerInnen: Öffentliche Liste (*Eintragung optional*)

Twitter: Hashtag #kimws19

Abendessen (Selbstzahler): Ellinn, E3,1, Di ab 19:00 Uhr (Wegbeschreibung)

Hoteloptionen: https://www.uni-mannheim.de/media/Universitaet/hotelliste.pdf

Kontakt: kim-info@dini.de

Programm

Dienstag, 02.04.2019

11:30 - 12:00	**Begrüßung** Sabine Gehrlein (UB Mannheim, Leitende Bibliotheksdirektorin) Stefanie Rühle (SUB Göttingen) und Jana Hentschke (DNB), DINI-AG KIM-Sprecherinnen

12:00 - 13:30	**Session: Datenvalidierung, Datenbereinigung, Workflowsteuerung** Moderation: Stefanie Rühle (SUB Göttingen) ***MAPS - Workflow-Tool zur Validierung und Transformation von XML Daten (Folien)*** Karl-Ulrich Becker (SUB Göttingen), Timo Schleier (SUB Göttingen) Das Tool MAPS wird an der SUB Göttingen (DDB Fachstelle Bibliothek) für die Steuerung und Automatisierung von Validierungs- und Transformationsprozessen im Rahmen von Datenlieferungen an die Deutsche Digitale Bibliothek entwickelt. MAPS besteht aus einem Preprocessing Tool, mit dem Daten gebündelt in eine eXist-XML Datenbank hochgeladen werden und einer eXist Webapp, über die Prozesse gesteuert werden. Die Anwendung wird über eine Web-GUI bedient und bietet neben einer Validierung mit automatischer PDF-Erzeugung ein Modul, mit welchem XSL-Transformationen aneinandergereiht und auf eine Auswahl von XML Dateien angewendet werden können. In dem Vortrag wird der aktuelle Stand der Entwicklung vorgestellt. ***Data Preparation Tool (DPT) (Folien)*** Oliver Götze (Landesarchiv BW) Das Data Preparation Tool wurde am Landesarchiv Baden-Württemberg (DDB-Fachstelle Archiv) als Werkzeug zur Aufbereitung und Validierung von archivischen Datenlieferungen an die Deutsche Digitale Bibliothek entwickelt. Das Python-basierte Tool ermöglicht die Verarbeitung von XML-Dateien in verschiedenen Formaten sowie die Ausgabe von EAD-Dateien, welche für den Ingest in die Portale verwendet werden können. Spezifische Anpassungen an den Daten können mit relativ geringem Aufwand erstellt, miteinander verknüpft und für verschiedene Datengeber nachgenutzt werden. Daneben können Vorschau-Ansichten der Datensätze im Archivportal-D-Layout erstellt werden, wobei eine repräsentative Testmenge automatisiert ausgewählt wird. Die Desktopanwendung nutzt das Qt-Framework für die GUI und setzt auf lxml für die performante Prozessierung auch großer XML-Dateien. Es sollen die Motivation für die Entwicklung eines eigenen Tools, der aktuelle Stand sowie mögliche Perspektiven präsentiert werden. ***Nightwatch - Because the night is dark and full of errors... (Folien)*** Daniel Opitz (SuUB Bremen) Viele Datenquellen, viele Formate, unsaubere Daten, unterschiedliche Lieferzeiten, komplexe Verfahren. Metadatenverarbeitung ist nicht unbedingt die einfachste und angenehmste Disziplin. SuUB Bremen entwickelt Nightwatch, um mehr Transparenz, Klarheit und Einfachheit in die Gestaltung und den Ablauf der Prozesse zu bringen. Nightwatch ermöglicht die Abbildung von Verarbeitungspipelines, die einerseits zur Dokumentationszwecken genutzt wird, aber auch die Grundlage für die Überwachung der Prozesse bildet. Die Verarbeitungsskripte, die die Anbindung an Nightwach implementieren, können den aktuellen Verlauf der Pipeline melden und zu jedem der Verarbeitungsschritte die Logs an Nightwatch senden, damit diese an einer zentralen Stelle für die Nutzer verfügbar sind. Darüber hinaus ist Nightwatch im Stande, selbst Pipelines zu steuern und die Verarbeitung der Daten über mehrere Server zu verteilen. In dem Vortrag werden die Funktionen und Einsatzzwecke von Nightwatch vorgestellt und die kommenden Entwicklungen erläutert.
	Mittagspause mit Verpflegung

14:30 - 16:00	**Hands-On-Tutorial**
	
	Option 1:
	
	Automatisierte Datenverarbeitung mit OpenRefine
	
	Leitung: Felix Lohmeier (Open Culture Consulting)
	
	Die Software OpenRefine ist bekannt für ihre grafische Oberfläche, die einem Tabellenverarbeitungsprogramm ähnelt. Sie wird oft als Desktop-Software installiert und zur Analyse und Bereinigung von heterogenen Daten eingesetzt (siehe auch Präsentation von Maike Kittelmann auf KIM WS 2016). Weniger bekannt sind die Automatisierungsmöglichkeiten von OpenRefine. Durch die Client-Server-Architektur lässt sich OpenRefine auch auf einem Webserver installieren und über die Kommandozeile steuern. Das hat den Charme, dass Transformationsregeln in der Oberfläche spielerisch erprobt werden können und dann beispielsweise täglich automatisiert auf neue Datenlieferungen angewendet werden können.
	
	Im ersten Teil des Hands-On-Tutorials steuern wir OpenRefine über die Kommandozeile. Dazu verwenden wir einen Client, der als Programm für Windows, MacOS und Linux bereitsteht und ohne weitere Installation ausgeführt werden kann.
	
	Im zweiten Teil des Hands-On-Tutorials schreiben wir auf dem bereitgestellten Webserver beispielhaft kleine Shell-Skripte, um eine vollautomatische Datenverarbeitung zu erproben. Dabei orientieren wir uns an Erfahrungswerten aus dem Projekt Hamburg Open Science Schaufenster.
	
	Materialien (Folien, Installationsanleitung, Handouts mit Aufgaben, Beispieldateien)
	
	Zielgruppe: Personen, die gelegentlich mit OpenRefine arbeiten und Automatisierungsmöglichkeiten kennenlernen möchten.
	
	Vorkenntnisse: Vorerfahrungen mit OpenRefine sind hilfreich, aber nicht erforderlich.
	
	Voraussetzungen: Ein Notebook mit Windows, MacOS oder Linux. Es muss keine Software vorab installiert werden, da ein Webserver mit OpenRefine für alle TeilnehmerInnen bereitgestellt wird. Es sind keine Programmierkenntnisse erforderlich, die nötigen Schritte werden einzeln demonstriert.
	
	Option 2:
	
	Indexierung von Fedora Datensätzen in Apache Solr mit Apache Camel
	
	Leitung: Jaime Penagos (UB LMU), Ralf Claussnitzer (SLUB Dresden), Rainer Gnan (UB LMU)
	
	Im Rahmen dieses Hands-On-Tutorials wird zunächst ein allgemeiner konzeptioneller Überblick bezüglich der im Titel genannten Anwendungen geboten. Im speziellen soll anschließend deren Zusammenspiel vor dem Hintergrund eines einfachen aber dennoch praxisorientierten Anwendungsfalls sowohl theoretisch erläutert, als auch praktisch auf technischer Ebene demonstriert werden. Dieser Workshop möchte Interessierten damit primär die Gelegenheit bieten, sich einen ersten Eindruck über die hier vorgestellten Technologien zu verschaffen und zudem den Einstieg in die Arbeit mit diesem Setup zu erleichtern.
	
	Folgende Agenda erwartet die Teilnehmer*innen:
	
	• Einführung in Fedora (Jaime Penagos)
	  • Fedora GUI
	  • Fedora HTTP REST API
	  • Fedora Events und Messaging
	• Indexierung mit Apache Solr (Rainer Gnan)
	  • Indexieren
	  • Index-Schema und Solr-Konfiguration
	  • Sucheinstieg
	• Integration von Fedora und Solr mit Apache Camel (Ralf Claussnitzer)
	  • Konsumieren von Fedora Events
	  • Ansteuern der Camel Routen für Ingest, Update, Delete
	  • Extrahierung von Informationen aus Dublin Core XML Dokumenten
	  • Beschicken von Solr
	
	Materialien:
	
	• Anleitung VM installieren
	• Folien der Präsentation
	• Fedora Update: Current Developments and Future Plans
	
	Zielgruppe: Personen, die mit digitalen Objekten und Archiven in Repositorien arbeiten.
	
	Vorkenntnisse: Unix/Linux, Java
	
	Voraussetzungen:
	
	• Ein eigenes Notebook mit vorinstalliertem Vagrant oder Oracle VirtualBox
	• SSH Client
	• Editor mit Syntax Unterstützung für XML und Java
	• Java 8 Entwicklungsumgebung, Maven oder idealerweise IDE

	Pause
16:30 - 18:00	**Hands-On-Tutorial (*Fortsetzung*)** **Option 1:** **Automatisierte Datenverarbeitung mit OpenRefine *(Fortsetzung)*** **Option 2:** **Indexierung von Fedora Datensätzen in Apache Solr mit Apache Camel *(Fortsetzung)***
	Für Interessierte: Schloss- und/oder Eis-Spaziergang zum Restaurant Ellinn (E3,1, Di, ab 19:00 Uhr, Wegbeschreibung)

Mittwoch, 03.04.2019

09:00 - 09:45	**Datensetbeschreibungen in DCAT - Modell und Implementierung (Folien)** Lars G. Svensson (DNB), Panagiotis Kitmeridis (DNB) Damit offen zugängliche Datensets auch nachgenutzt werden können, müssen sie auffindbar sein: Dies ist die Aufgabe von Suchmaschinen und Datenportalen. Um die dafür notwendigen Metadaten zu strukturieren, wurde 2012-2014 das RDF-Vokabular Data Catalog Vocabulary (DCAT) entwickelt, das mittlerweile weite Verwendung findet. Wir wollen vorstellen, was DCAT ist, wer es pflegt, in welchen Bereichen es derzeit überarbeitet wird, und wie wir es in der DNB verwenden, um unsere Datensets zu beschreiben.
09:45 - 10:15	**Datenstrukturen für strukturierte Daten: Aus der Arbeit der W3C Data Exchange Working Group (Folien)** Lars G. Svensson (DNB) In seiner nächsten Version wird das Data Catalog Vocabulary (DCAT) neben der Beschreibung von Datensets auch die Möglichkeit bieten, APIs für den Datenzugriff zu beschreiben. Diese neuen Funktionalitäten zu spezifizieren ist Aufgabe der W3C Data Exchange Working Group (DXWG). Der Auftrag der Arbeitsgruppe umfasst aber auch Spezifikationen für die Beschreibung von Anwendungsprofilen und Best Practices dafür, wie man beschreiben kann, welchem Anwendungsprofil Daten entsprechen und wie man dies in der Suche und für HTTP Content Negotiation einsetzen kann. Auch Verbesserungen in der Beschreibung von komprimierten Daten und Provenienzinformation stehen auf der Agenda. Diese und weitere Themen der Arbeit sollen hier angerissen und erläutert werden.
	Pause
10:45 - 11:15	**Lightning Talks** *Michael Büchner*: EF2SO – Schema.org für GND-Entitäten *Mathias Wyser*: Schulungskonzept – Go for Data Analysis (Folien) *Ralf Claußnitzer*: urnlib – Eine Java-Bibliothek für standardkonforme URNs *Alessandro Aprile*: “Hat jemand Erfahrung mit” Metadatenaustausch zwischen Katalogisierungsclient WinIBW und Repository (DSpace)? *Anna Kasprzik*: Metadaten für die Automatisierung (Folien) *Christiane Klaes*: LexBib – Modellierung von Provenienzmetadaten einer Fachbibliographie (Folien)
11:15 - 12:45	**Session: Anforderungen an moderne Metadatenhaltung (Folien)** Moderation: Jana Hentschke (DNB) In dieser partizipativen Session wollen wir mit allen Workshop-TeilnehmerInnen zusammen Anforderungen an moderne Metadatenhaltung erarbeiten und einem Realitätscheck unterziehen. Wenn wir mal unabhängig von bestehender Infrastruktur denken: was wäre wünschenswert für eine zukunftsfähige, transparente Metadatenhaltung (Administrative Metadaten, Datenprovenienz, Versionierung, ...)? Welche Anwendungsfälle kennen wir bereits aus unserer täglichen Arbeit? Als Ergebnis soll eine nach Alltagsrelevanz priorisierte Anforderungsliste stehen. Im nächsten Schritt wollen wir existierende Systeme, Anwendungen, Datenveröffentlichungen, Datenschnittstellen, Metadatenstandards mit dieser Liste abgleichen und eine Übersicht erstellen. Nach dieser Inventur können wir gemeinsam analysieren und diskutieren. Zusammenfassung, Status-Quo-Übersicht-Matrix (online, weiter editierbar), Status-Quo-Übersicht-Matrix (Stand Session-Ende)
	Mittagspause mit Verpflegung

13:45 - 15:30

Session: Zusammenbringen, was zusammengehört

Moderation: Philipp Zumstein (UB Mannheim)

Bei der Arbeit mit Daten aus verschiedenen Datentöpfen ist es eine häufig auftretende Herausforderung, automatisch Gleiches auf Gleiches abzubilden. Je nach Anwendungskontext und lokalen sprachlichen Gepflogenheiten ist die Rede von "Matching", "Clustering", "Zusammenführen", "Deduplizierung", "Bündelung", "Merging", ...

In dieser Session berichten verschiedene Akteure von ihren Strategien und Erfahrungen in diesen Disziplinen. Es geht schwerpunktmäßig um den Gegenstand "bibliographische Daten". Thematisiert werden Algorithmen, Datenformate von Ergebnisse, Workflows, die Datennutzersicht sowie alles Weitere, zu dem Austauschbedarf aufkommt.

Clustern von Daten auf der swissbib Plattform (Folien)

Silvia Witzig (swissbib), Günter Hipler (swissbib)

swissbib wurde im Verlaufe der letzten 10 Jahren aufgebaut und ist in dieser Dekade zu einem wesentlichen Element des Zugriffs auf bibliographische Informationen in der Schweiz geworden. 45 Millionen Ressourcen werden ausschliesslich automatisiert aggregiert, dedupliziert, geclustert und mit externen Informationen verlinkt.

NutzerInnen kennen vor allem die Discoveryfunktionalität oder die Möglichkeiten der angebotenen maschinellen Schnittstellen. Der Nukleus oder das Herz des Service, auf dem alle bisherigen und zukünftigen Dienste aufbauen, sind jedoch die Möglichkeiten des data-processing und der gezielten Bereitstellung von Informationen für sehr unterschiedliche dedizierte Services.

Wir fokussieren in unserem Beitrag die Datenaufbereitung und die Mechanismen, die im swissbib-Datenhub zur Dedublierung verwendet werden. Feedback von den Workshop-TeilnehmerInnen ist willkommen, ob wir tendenziell eher auf die Regeln des Clusterns oder die Abläufe jetzt und evtl. in der Zukunft eingehen sollen.

De-Duplikationsverfahren und Einsatzszenarien im Gemeinsamen Verbündeindex (GVI) (Folien)

Uwe Reh (HeBIS), Stefan Winkler (BSZ), Thomas Kirchhoff (BSZ), Stefan Lohrum (KOBV)

Im Vortrag werden unterschiedliche Strategien (cluster- vs. schlüsselbasierte Verfahren) vorgestellt und ihre Anwendung in verschiedenen Kontexten (z.B. Präsentation von Rechercheergebnissen Frontend, Nachrecherche in der Fernleihe) diskutiert. Abschließend wird über spezifische Herausforderungen und Erfahrungen bei der Anwendung der Verfahren in einem großen Datenbestand (der GVI enthält rund 170 Mio Datensätze) berichtet.

Hervorholen, was in unseren Daten steckt - Mehrwerte durch Analysen großer Bibliotheksdatenbestände (Culturegraph) (Folien)

Angela Vorndran (DNB), Stefan Grund (DNB)

Das Team Datenabgleich und Datenanalyse der DNB führt in Culturegraph die Datenbestände der Bibliotheksverbünde aus Deutschland und Österreich zusammen (zurzeit mehr als 171 Millionen Titeldatensätze) für Vernetzungen innerhalb des Bestandes und mit Daten aus externen Quellen. Beispielsweise werden Publikationen, die einem Werk zuzurechnen sind, gebündelt, so dass inhaltserschließende Informationen unter den Bündelmitgliedern ausgetauscht und somit intellektuell erstellte Metadaten in größerem Umfang ausgetauscht werden können. Des Weiteren wurden öffentliche Daten der Open Researcher and Contributor ID (ORCID) mit den Daten in Culturegraph und der GND abgeglichen, um die ORCID in größerer Zahl in GND-Personendatensätze eintragen zu können. Auch statistische Auswertungen auf dem Datenbestand werden vorgenommen.

Mit CultureGraph zu einer virtuellen Systematik (Folien)

Hans-Georg Becker (UB Dortmund)

Umfangreiche Anpassungen des Bestandsmanagement aufgrund der nahenden Sanierung der Bibliothek, aber auch wegen der "e-preferred" Strategie beim Medienerwerb, machen ein Festhalten an der Aufstellungssystematik der UB Dortmund unmöglich. Die UB Dortmund arbeitet daher daran, eine "virtuelle Systematik" der Bestände auf Basis der RVK anzubieten.

Da die UB Dortmund die inhaltliche Erschließung bisher eher stiefmütterlich behandelt und die RVK im hbz-Verbund quasi keine Bedeutung hat, liegen kaum inhaltserschließende Daten vor. Mithilfe der Daten aus CultureGraph wird daher versucht, die "virtuelle Systematik" aus den Erschließungsdaten der anderen Verbünde aufzubauen.

Der Vortrag stellt das Verfahren und die Ergebnisse vor und zeigt weitere Clustering-Ansätze zum Schließen der verbleibenden Lücken vor.

15:30 - 15:45	**Closing** Stefanie Rühle, Jana Hentschke
15:45 - 16:45	**Öffentliche Sitzung der DINI-AG KIM** Agenda