Hans-Jörg Lieder, Staatsbibliothek zu Berlin, spricht zur Begrüßung vor vollem Publikum. Fotos: SBB-PK / Hagen Immel

SWIB23

Bericht über die Semantic Web in Libraries Konferenz, Berlin, 11. – 13. September

Katja Jana

Die diesjährige Semantic Web in Libraries Konferenz, SWIB23, fand vom 11. bis 13. September in der Stabi Berlin statt, zum ersten Mal seit der Corona Pandemie wieder vor Ort, nur als Präsenzveranstaltung.[1] Auch in diesem Jahr organisierten das ZBW, Leibniz-Informationszentrum Wirtschaft, und das hbz, das Hochschulbibliothekszentrum NRW, die Konferenz gemeinsam. Rund 200 Teilnehmerinnen und Teilnehmer aus über 18 Ländern fanden sich auf der seit 2012 englischsprachigen Konferenz zusammen, um sich zum Thema Linked Open Data zu informieren und auszutauschen.

Von der Stabi Berlin begrüßte Hans-Jörg Lieder[2] die Teilnehmenden mit der Frage, ob das Semantic Web überhaupt eine Zukunft habe und welche Rolle die Bibliotheken dabei spielen könnten. Dieser Frage wurde auf der Konferenz in fünf thematischen, aber untereinander eng verknüpften Panels – Authorities, Data Modelling, Utilizing Wikimedia, Collections, Aggregators – nachgegangen.[3] Zwei prominente Themen in diesem Jahr waren die Arbeit mit Wikidata und die Entwicklung von Machine Learning Verfahren.

Die Organisatorinnen und Organisatoren stellten zum Auftakt fest, dass mittlerweile immer mehr Datensammlungen im offenen Zugang zur Verfügung stehen. Ausgehend von diesem Status Quo adressierte die Keynote von Sarven Capadisli, der seit vielen Jahren aktiv an der Definition von offenen Web-Standards beteiligt ist, u.a. als Teil der W3C Solid Community Group, die Voraussetzungen für ein Web3.0 und die Bedeutung offener Standards für die weitere Etablierung von Linked Open Data und des Semantic Web.

Vor dem Hintergrund der Konzepte „Autonomy“ und „Universal Access“ behandelte er die Frage, welche neueren Entwicklungen im Bereich Offene Standards die Arbeit von Bibliotheken unterstützen können und vice versa. Dabei bezog er sich auf Self Publishing und Linked Research von Individuen, aber nicht zuletzt auch von Communities als einer Verwirklichung dieser Prinzipien.[4]

1 https://swib.org/swib23/index.html [21. November 2023]

2 Am 10.11.2023 verstarb unser Kollege Hans-Jörg Lieder, der seit 2011 Leiter der Abteilung für Überregionale Bibliographische Dienste war, plötzlich und unerwartet. Während ich diesen Bericht verfasse, sind die anderen Kolleginnen und Kollegen der Staatsbibliothek und ich immer noch fassungslos und nehmen Abschied in tiefer Trauer. https://blog.sbb.berlin/nachruf-auf-hans-joerg-lieder/

3 Die aufgezeichneten Vorträge, Slides, Abstracts und Informationen zu den Vortragenden finden sich unter https://swib.org/swib23/programme.html [21. November 2023].

4 Self Publishing ist im weitesten Sinne, die Veröffentlichung im Selbstverlag auf Onlineplattformen und Linked Research ist eine Initiative zum Self Publishing und die Umsetzung von Forschungspraktiken nach den im Vortrag vorgestellten Prinzipien (offene Standards, Open Linked Data, etc., vgl. https://linkedresearch.org/ [21. November 2023]), wie sie etwa in dem vom ihm entwickelten Editor möglich ist: dokielihttps://dokie.li/ [16. November 2023]

Zentrale Fragen sind in diesem Zusammenhang die Kontrolle bzw. Wahl von IDs, Daten, Services, Tools und Anwendungen. Dem liegt der Gedanke des Web als soziale Maschine zugrunde, als einem Werkzeug, dass die Gesellschaft mit all ihren Problemen reflektiert. Anzustreben ist deswegen ein dezentralisiertes Web, in dem bestimmte Standards, etwa die W3C Linked Data Notification eingehalten und weiterentwickelt werden.[5] Als wesentlich betrachtet er zudem die Kompatibilität und vor allem Interoperabilität von Anwendungen. Interoperabilität beinhaltet den Bezug auf Standards, eine Verständigung über Regeln und Grundsätze, deren Weiterentwicklung oder Verwerfung. Die Standards sollten daran gemessen werden, ob sie gesellschaftliche Probleme lösen und ob sie transparent dazu sind, wie etwas funktionieren soll und warum bestimmte Entscheidungen getroffen werden sollen. Diese weitgehend theoretischen Prinzipien könnten von Bibliothekarinnen und Bibliothekaren in den unterschiedlichsten Anwendungsfällen erprobt werden, so Capadisli, ohne konkrete Beispiele zu nennen. Bibliotheken hätten die Möglichkeit kostengünstig auf offenen Standards aufzubauen und diese Lösungsansätze und Implementierungsversuche in ihrer Rolle als Multiplikatorinnen zu teilen.

Ob das im Anschluss an die Keynote vorgestellte Projekt, das Linked Data und Discovery System der Nationalbibliothek von Singapor den von Sarven Capadisli vorgestellten Standards entspricht, könnte diskutiert werden. Richard Wallis, ausgewiesener Experte für Linked Data und Semantic Web und Leiter der W3C Community Groups, stellte die Genese des 2022 live gegangenen Systems vor. In dem Projekt arbeiteten mehrere kommerzielle Partner zusammen. Open-Source-Entwicklungen seien an passender Stelle berücksichtigt worden. Ziel des Projekts war es, einen nationalen Discovery-Service aufzubauen und darin Normdaten-Entitäten aus verschiedenen Quellen in einem Interface mit den Katalogdaten zu verknüpfen. Umfangreiche Datentransformationen nach einem auf BIBFRAME und schema.org beruhenden Linked Data Modell wurden dafür implementiert und die Daten in einen Knowledge Graphen geladen. Eine Herausforderung stellt vor allem die Zusammenführung der zahlreichen Normdaten dar.[6] Aus dem Vortrag wurde nicht klar ersichtlich, inwiefern die Daten und die nicht offenen ETL-Prozesse nachnutzbar sind. Sicherlich kann das Projekt aber Orientierung und Anregung für ähnlich gelagerte Projekte geben.

Unter etwas anderen Vorzeichen wurde das Thema der Zusammenführung von Normdaten im Panel „Authorities" fortgesetzt. Steven Folsom von der Cornell University Library berichtete von der geplanten Weiterentwicklung der Rechercheplattform für Normdaten LD4P.[7] Ziel bei LD4P ist es, bisherige Caching basierte Prozesse weitestgehend durch eine Metadata Management API zu ersetzen. Es geht vor allem darum, einen robusten Suchdienst anzubieten, der auf die Bedürfnisse der Katalogisiererinnen und Katalogisierer zugeschnitten ist, dafür Spezifikationen festzulegen und diese im Katalogisierungsformular von Sinopia zu implementieren. Schwierigkeiten bereiteten bisher die Indizierung der verschiedenen Vokabulare, Data Dumps aus verschiedenen Vokabularen und die Tatsache, dass beim Caching in den Data Dumps kürzlich vorgenommene Änderungen nicht sichtbar sind. Der Wechsel auf APIs wird zwar angestrebt, momentan bestehe aber noch das Problem, dass nicht alle Datenquellen über solche verfügen.

Auch das Dutch Digital Heritage Network (NDE) stellte die Weiterentwicklung seiner Dienste seit der letzten Präsentation auf der SWIB vor sechs Jahren vor. Zwar ist inzwischen einiges vorangegangen, jedoch sind die digitalen Sammlungen noch immer schwer auffindbar. Auch hier soll ein nutzerzentrierter Ansatz die Richtung vorgeben, um ein Netzwerk von Netzwerken zu schaffen, dass die Sichtbarkeit des digitalen kulturellen Erbes erhöht. Entwickelt wurde aus dieser Idee eines generisch angelegten „Network of Terms", dass es auf andere Anwendungsfälle übertragbar ist. Auf der Basis von Gemeinsamkeiten wurden Anwendungsfälle beschrieben und diese in eine gemeinsame an den FAIR-Prinzipien orientierten Linked-Data-Architektur integriert. Dieses Collection Managment System beinhaltet eine Dataset Registry, die mittels SPARQL abgefragt werden kann. Die Normdaten werden durch SPARQL and GraphQL mit einander verbunden. Zudem wurden API-Reconciliation-Dienste eingerichtet und eine Volltextsuche.

Das solche Vorhaben hohe personelle und zeitlich Aufwand benötigen, wurde an einem wesentlich kleineren Projekt deutlich: Joe Cera und Michael Lindsey gaben einen Einblick in den Aufbau eines Repositoriums für die Berkeley Law Library mit der Wikibase Software. Die Wikidata QID dient als Autorinnen-ID und ersetzte bisherige institutsinterne provisorische IDs wie E-Mail-Adressen. Durch die Wikidata API können die Daten aktuell gehalten werden. Bisher fehlen allerdings die personellen Ressourcen, dass Vorhaben vollständig zu implementieren.

Im Panel „Datenmodellierung" gingen Ruth K. Tillman und Regine Heberlein auf die Komplexität archivarischer Daten ein. Die Vortragenden diskutierten verschiedene

5 Das Solid Projekt zielt darauf in diesem Bereich zu ermächtigen (Verweis auf Solid Definition: "equitable, informed and interconnected society" – "ethical web principles ").

6 Hier stellte sich der Levenshtein Algorithmus, besonders auch für Asiatische Namen, als hilfreich heraus.

7 https://wiki.lyrasis.org/pages/viewpage.action?pageId=74515029 [16. November 2023]

Linked-Data-Modelle, die die diversen Verzeichnungeinheiten der Bestände zusammenführen können. Sie betonten die tiefe Verschachtelung archivarischer Daten, die eine Vielzahl von Einzelobjekten (Items) enthalten. Drei Linked-Data-Modelle, die Forschenden dienen sollen, Materialien möglichst zielgenau zu suchen und zu identifizieren, wurden näher untersucht: BIBFRAME ARM Extension, Records in Contexts (RiC) und Linked.Art. Ausgangsfrage war, wie über ein Datenmodell die Verbindung zwischen Findbuch und physischer Lokalisation im Magazin hergestellt werden kann: "Where does a beautiful ontology meet actual stuff?"[8]

Ein weiterer Beitrag zum Thema Datenmodelle befasste sich mit der Entwicklung einer Share-VDE (Virtual Discovery Environment) Ontologie. Es handelt sich um ein Kollaborationsprojekt einer bedeutenden Anzahl von Bibliotheken in Nordamerika, Großbritannien und Skandinavien, vorgestellt von Tiziana Possemato, Jim Hahn und Oddrun Ohren. Das Projekt basiert auf OWL and will einen Share-VDE-Discovery-Service entwickeln, eine Linked-Data-Suchmaschine, die sich des BIBFRAME-Vokabulars bedient.[9] Das Ganze ist noch Work in Progress, es gibt aber bereits einen Pre-Release der Ontologie.

Die wachsende Beliebtheit und Verwendung von Wikidata kam nicht nur im Panel „Utilizing Wikimedia" zum Ausdruck, sondern zog sich durch die gesamte Konferenz. Crystal Yragui und Adam Schiff demonstrierten die Anreicherung von Archivdaten mit Wikidata in einem Kollaborationsprojekt der University of Washington Libraries (UWL) mit den Labor Archives of Washington (LAW). Sie sehen die Vorteile von Wikidata in den Datenmengen, der Zugänglichkeit durch einen SPARQL-Endpoint und der flexiblen Datenstruktur. Das Projekt lädt durch seine ausführliche Dokumentation und Offenheit zur Nachahmung ein.[10] Attraktiv an dem Ansatz ist zudem, dass zugleich die eigenen Daten als auch Wikdata aufgewertet und verbessert werden können. Die Katalogisiererinnen und Katalogisierer wurden alle in der Anwendung der Wikidata-Workflows geschult.

An die Thematik der Fortbildung in der Arbeit mit Wikidata anschließend, stellte Will Kent von Wiki Education ein Kursprogramm auf Fortgeschrittenen-Niveau vor, dass im November 2022 startete. Will Kent betont die Bedeutung der Fortgeschrittenenkurse für Wikimedia, denn strukturierte Angebote mit Live-Dozenten seien nach wie vor selten, böten aber ganz andere Möglichkeiten als Selbstlernkurse. Die Auswirkungen auf Wikidata seien enorm und der Bedarf vorhanden. Der Erfolg dieses Kurses ließe sich direkt anhand von 11.000 neuen Items messen, die in den sechs Wochen Kursdauer zu Wikidata hinzugefügt wurden.

Welchen Beitrag Wikidata im Bereich des machine learning leisten kann, zeigten Kai Labusch und Clemens Neudecker vom Projekt Mensch.Maschine.Kultur der Stabi Berlin. Die Daten aus Wikidata und Wikipedia werden hier zur Linked Entity Recognition (LER) zusammengeführt. Die Schwierigkeiten, in historischen Dokumenten Personen eindeutig zu identifizieren, um sie mit Normdatenquellen zu verlinken, sollen mit der kombinierten Verwendung überwunden werden. Dafür wurde ein auf BERT basierender vortrainierter NER-tagger an den Daten der digitalisierten Sammlung der Stabi Berlin optimiert. Zu den Entitäten aus Wikidata werden Sätze aus der Wikipedia herausgefiltert, die über diese Entitäten Aussagen treffen. Sie werden dann über Satzvergleiche mit dem Originaldokument identifiziert. In der SBB sind bisher fünf Millionen Seiten der digitalen Kollektion mit dem Verfahren angereichert worden. Ein nächster Schritt ist die Veröffentlichung der Ergebnisse im digitalen Sammlungsportal und die Erprobung an anderen Anwendungsfällen.

Die Machine-Learning-Thematik führten Florian A. Grässle und Tina Trillitzsch im nachfolgenden Panel „Collections" fort und stellten die automatische Indexierung mit Annif im PSYNDEX des Leibniz Institute for Psychology (ZPID) vor.[11] Das Vokabular für die Indexierung besteht aus psychologiespezifischen kontrollierten Vokabularen. Seit Februar 2023 setzt das ZPID ANNIF ein.[12] Die Performanz soll in Zukunft verbessert werden, denn eine Reihe von Begriffen schlägt ANNIF entweder zu oft vor oder zu selten. Eine Lösung hierfür könnte ein SKOS-Vokabular sein oder häufigere Updates des jetzigen Vokabulars und die Einführung von Blocklisten.

Die Erschließung der Albrecht-Haupt-Sammlung der TIB Hannover und der Aufbau eines Portals mit der Software Vitro war Thema der Präsentation von Georgy Litvinov, Birte Rubach und Tatiana Walther von der TIB Hannover. Das von der DFG finanzierte GESAH(+) GESAH - Graphic Arts Ontology Projekt umfasst die Einzelblätter (davon bisher 6.200 digitalisiert) der Sammlung. Der erste Schritt war, die für Vitro benötigte OWL-Ontologie zu schaffen. Die Ontologie orientiert sich an LIDO (XML Schema and exchange format) und CIDOC CRM als Referenzmodell. Das

8 https://youtu.be/oyrVaKJYt7w [21. November 2023]

9 https://svde.org [16. November 2023]

10 Zur Dokumentation eines Workflows mit Python und OpenRefine siehe https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/University_of_Washington/Workflow,_Trainings,_and_Resources/EAD_to_Wikidata_Workflow [16. November 2023]

11 PSYNDEX ist eine bibliographische Datenbank für die englisch- und deutschsprachige Öffentlichkeit, mit monatlich circa tausend zusätzlichen Veröffentlichungen.

12 Bisher arbeitete die Einrichtung mit AUTINDEX, eine veraltete Software, die instabil lief.

Verfahren habe sich als sehr aufwendig herausgestellt. Als nächste Schritte sind die Entwicklung einer Verlinkung zu NFDI4Culture und einer dynamischen API geplant.

Mit den Grenzen künstlicher Intelligenz beschäftigte sich ein Vortrag zu einem Projekt des FIZ Karlsruhe zur Datenanreicherung in der DDB, bzw. wie der geänderte Titel besagt, den Herausforderungen geerbter Metadaten in der Verarbeitung mit künstlicher Intelligenz. Mary Ann Tan ging der Frage nach, wie eine Ontologie zwischen einem Druck auf Marzipan und einem Buch unterscheiden könne. Anhand dieses Beispiels erläuterte sie verschiedene Schwächen des DDB-EDM-Datenmodells und der Schwierigkeit, anhand dessen Daten mit der GND abzugleichen. Beim Reconciling der Properties Title, Agent und Event konnten bei 5.658.431 Objekten nur 1.32 Prozent mit GND-Einträgen verlinkt werden. Wegen der zu geringen Informationen in den Datensätzen, insgesamt uneinheitlicher Metadatenqualität und der geringen Anzahl von GND-Datensätzen sieht Tan das Projekt vor großen Herausforderungen. Vor allem Werktitel sind in der GND im Vergleich zur Datenmenge in eher kleiner Anzahl vorhanden.

Den Abschluss bildete das Panel „Aggregators". Das Portal FrenchArchives, vorgestellt von Élodie Thieblin, Fabien Amarger, Saurfelt Katia, Mathilde Daugas ist eine Linked-Data-Plattform für zahlreiche Archive, momentan 137, in ganz Frankreich. Es basiert auf der Open-Source-Software CubicWeb. Mittels der Ontologie Records In Contexts (RiC-O) werden die Daten mit verschiedenen Normdatenquellen – neben Wikidata sind dies Data.bnf und Persée – abgeglichen. Stand Mai 2023 waren 274 Millionen von 385 Millionen Tripel mit Normdaten verlinkt, das ist allerdings immer noch eine geringe Anzahl von 1,4% der Entitäten. Wie beim Projekt der University of Washington Libraries ist auch hier der Export von Normdaten in die Quellen Data.bnf, Wikidata und Persée vorgesehen.

Nach welchen Kriterien kann eine Validierung der Datenanreicherung erfolgen? Gegen welche Qualitätsansprüche müssen die Anreicherungen validiert werden? Wie kann Transparenz darüber geschaffen werden, auf welche Weise die Daten angereichert wurden? Zu guter Letzt widmete sich diesen Fragen eine Gruppe des Netzwerks Jewish Heritage der Europeana und diskutierte in ihrer Präsentation die Evaluationskriterien und Prozesse für die Anreicherung von Geo-Entitäten im Rahmen der Jewish History Tours. Allerdings stellt diese händische Evaluation von Samples nur einen ersten Ansatz dar. Offene Fragen sind zuverlässige Confidence Scores, um Übertragbarkeit auf andere Daten zu gewährleisten und was repräsentative Stichproben sind. Die notwendigen personellen Ressourcen zur umfangreichen manuellen Überprüfung sind

Best Practices for sharing and discovering ETL workflows, Foto: SBB-PK / Hagen Immel

enorm und momentan nicht vorhanden. Die Evaluation durch Ground Truth wäre eine gangbare, ressourcenschonende Möglichkeit.

Am zweiten Konferenztag gab es am Nachmittag eine Reihe von Lightening Talks, die kurze Einblicke zum Beispiel zum Thema Metadaten für graue PDF Literatur, den neuesten SkoHub-Entwicklungen und zusätzliche Informationen zum „Network of Terms" boten und die Ankündigung einer Mailingliste für Wikimedians in Bibliotheken.[13]

Die Workshops im Vorlauf der Konferenz am Montag und die Breakout-Sessions am Dienstag schufen Gelegenheiten, neben den informationsgeladenen Vorträgen konkret Tools auszuprobieren und über gemeinsame Fragestellungen in den Austausch zu kommen. In einer der Breakout-Sessions beschäftigten die Teilnehmenden sich zum Beispiel mit der Frage, wie Anwendungsfälle für ETL-Prozesse auffindbar und wiederverwendet werden können. Bisher findet in dem Bereich noch wenig Vernetzung statt, wie die Organisatoren und die Teilnehmenden feststellten. Die Frage dabei ist, wie man sehr spezialisierte Anwendungsfälle im ETL-Bereich wieder verwertbar machen kann, etwa durch Modularisierung. Auch eine Best-Practice-Guideline wäre wünschenswert.

Nicht zuletzt durch die unterschiedlichen Austauschformate war die SWIB23 eine gelungene und inspirierende Mischung aus fachlichem Austausch, Lernangeboten und über die Tagung hinausgehenden Anknüpfungs- und Vernetzungsmöglichkeiten.

Dr. Katja Jana
seit 3/2022 in der Abteilung Überregionale Bibliographische Dienste der Stabi Berlin als Bibliothekarin tätig; Historikerin, promoviert in Göttingen 2018, seit 2015 in der Redaktion von WerkstattGeschichte, Studium der Geschichte (FU Berlin) und GenderStudies (HU Berlin) in Berlin und Istanbul, MA LIS IBI HU Berlin 10/2022.
katja.jana@sbb.spk-berlin.de

13 openglam-de@lists.wikimedia.org