Automatisierung der Sacherschließung mit Semantic-Web-Technologie

Ralph Hafner, Universität Konstanz, Kommunikations-, Informations-, Medienzentrum (KIM)
Bernd Schelling, Universität Konstanz, Kommunikations-, Informations-, Medienzentrum (KIM)

Zusammenfassung:
Der vorliegende Artikel möchte einen Ansatz vorstellen, der aufzeigt, wie die Bibliothek der Universität Konstanz – und andere Bibliotheken mit einer Haussystematik – bei ihrer eigenen Systematik bleiben und trotzdem von der Sacherschließungsarbeit anderer Bibliotheken profitieren können. Vorgestellt wird ein Konzept, das zeigt, wie mithilfe von Semantic-Web-Technologie Ähnlichkeitsrelationen zwischen verbaler Sacherschließung, RVK, DDC und hauseigenen Systematiken erstellt werden können, die das Übersetzen von Sacherschließungsinformationen in andere Ordnungssysteme erlauben und damit Automatisierung in der Sacherschließung möglich machen.

Summary:
The paper presents an approach how libraries with a custom classification scheme such as the library of Konstanz University can make use of subject headings and classification data produced by other libraries. The proposed model uses semantic web technology to create relations of similarity between subject headings, RVK, DDC and custom classification schemes. These relations can be used to translate between these schemes, which makes automated classification and indexing possible.

Zitierfähiger Link (DOI): http://dx.doi.org/10.5282/o-bib/2015H4S161-175
Autorenidentifikation: Hafner, Ralph: GND 14275627X; Schelling, Bernd: GND 122761782
Schlagwörter: Sacherschließung; Inhaltserschließung; Automation; Automatisierung; Klassifikation; Semantic Web; Linked Data; Schlagwortkatalogisierung; verbale Sacherschließung; Dewey-Dezimalklassifikation; DDC; Regensburger Verbundklassifikation; RVK; hauseigene Systematik; Konstanzer Systematik

1. Ausgangslage

Die Einheitsklassifikation stand nicht rechtzeitig zur Gründung der Bibliothek der Universität Konstanz zur Verfügung. „Während der Jahre 1965/66 versuchte man, für die Bibliotheken der neugegründeten Universitäten in Bochum, Bremen, Dortmund, Konstanz und Regensburg ein einheitliches Klassifikationssystem zu erarbeiten. Dieser Vorschlag scheiterte aber an dem Zeitdruck während der Aufbauphase [...]."[1] Die bestehenden Klassifikationen wurden von den Gründerinnen und Gründern der Bibliothek der Universität Konstanz für nicht geeignet befunden. Man erlag der Verführung, dass eine eigene Systematik die Bedürfnisse besser würde erfüllen können. Diese

1 Müller-Dreier, Armin: Einheitsklassifikation. Die Geschichte einer fortwirkenden Idee, Wiesbaden: Harrassowitz, 1994 (Beiträge zum Buch- und Bibliothekswesen 35), S. 35.

Entscheidung für eine hauseigene Systematik könnte man im Rückblick als den Sündenfall der Konstanzer Sacherschließung bezeichnen. Entschuldigend kann hinzugefügt werden, dass in den 60er Jahren nicht absehbar war, dass der Fernzugriff auf Daten schon bald kein Problem mehr sein würde, die Unterschiede in der Formatierung von Daten hingegen bis heute eine sehr große Herausforderung in der Informationsverarbeitung darstellen.

Die damals in Konstanz getroffene Entscheidung für eine eigene Systematik, die bis heute im Einsatz ist, hat Vor- und Nachteile. Die Vorteile sind: Die Bibliothek der Universität Konstanz hat eine feingliedrige, eigene Systematik, die gut auf die Gegebenheiten und Bedürfnisse vor Ort abgestimmt ist. Muss die Systematik korrigiert oder erweitert werden, konnte und kann das zumeist sofort umgesetzt werden, ggf. nach Rücksprachen mit Kolleginnen und Kollegen vor Ort. Die Konstanzer Systematik ist, da die Konstanzer Bibliothekar/inn/en der ersten Generation so weise waren, sie nicht in ein Dezimalkorsett zu zwängen und mit relativen Hierarchien zu arbeiten, sehr flexibel. Die Nachteile sind: Die Notationen der Konstanzer Systematik sind lang (vgl. gri 900:p718:ka64a/t94).[2] Merken werden sich die meisten Nutzerinnen und Nutzer der Bibliothek wahrscheinlich nur die Bedeutung der Buchstabenkombination am Anfang der Signatur. Die Flexibilität, die man gewonnen hat, als man sich gegen ein dezimales System entschieden hat, bezahlt man mit der schwereren Vermittelbarkeit des Aufbaus der Systematik.[3]

Aber der Punkt, auf den dieser Aufsatz fokussiert, ist folgender Nachteil der hauseigenen Systematik: Konstanz konnte und kann bislang nicht von der Sacherschließungsarbeit anderer Bibliotheken profitieren, im Gegensatz zu Bibliotheken, die auf verbale Sacherschließung oder auf die RVK gesetzt haben. Die Konstanzer Fachreferent/inn/en systematisieren also seit bald fünfzig Jahren alle neuen Medien selbst und kommen dabei inzwischen auf eine Anzahl von knapp zwei Millionen Einheiten. Umgekehrt kann bislang auch keine andere Institution von der Konstanzer Sacherschließungsleistung profitieren.

Man könnte daher überlegen, die Konstanzer Sacherschließung umzustellen, z.B. auf RVK, um auf diese Weise die Sacherschließung anderer Institutionen nutzen zu können und seinerseits etwas in dieses System einzubringen. Allerdings wäre der Aufwand, knapp 2 Mio. Bände umzusystematisieren, immens.

Der vorliegende Artikel möchte einen Ansatz vorstellen, der aufzeigt, wie Konstanz – und andere Bibliotheken mit einer Haussystematik – bei ihrer eigenen Systematik bleiben und trotzdem von der Sacherschließungsarbeit anderer Bibliotheken profitieren können.

2 Näheres zur Struktur der Konstanzer Notationen: s. Bösing, Laurenz; Stoltzenburg, Joachim; Thomashoff, Barbara: Regeln für den Aufbau von Buchsignaturen. (Überarbeitete und auf den neuesten Stand gebrachte Fassung der Regeln vom April 1967 (B/Sto/Th) unter Berücksichtigung aller späteren Anhänge und Ergänzungen), Konstanz: Bibliothek der Universität Konstanz, 1969 (Bibliothek aktuell / Sonderheft 1), S. 3ff. Nota bene: gri 900:p718:ka64a ist die Systemstelle für Plato / Apologia / Kommentar.

3 Als Beispiel diene die Phonetik als Teilgebiet der Sprachwissenschaft, sie befindet sich in folgendem Bereich der spr-Systematik: spr 86 – spr 114. Eingängiger wäre ein Systematikbereich spr 80 – spr 89.

Umgekehrt versetzt der vorgestellte Ansatz auch diejenigen Institutionen, die mit Standardsacherschließungssystemen wie verbaler Sacherschließung nach der GND, mit RVK oder DDC arbeiten, in die Lage, die Sacherschließungsdaten von Einrichtungen mit lokaler Systematik zu nutzen.

2. Ziele

Folgende Ziele verfolgt das hier vorgestellte Projekt:[4]

Erreicht werden soll eine Arbeitserleichterung für Fachreferent/inn/en durch Automatisierung der Sacherschließung. In den Fällen, in denen eine vollautomatisierte Sacherschließung nicht möglich ist, soll die Sacherschließungsarbeit durch maschinell erzeugte Vorschläge unterstützt werden. Auch bisher Unerschlossenes wie E-Book-Sammlungen sollen so automatisiert nach der lokalen Systematik sacherschlossen werden können.

Ziel ist es, einen Weg aus der Isolation in der Sacherschließung zu finden. Dafür sollen die lokalen Sacherschließungsdaten in interoperable Daten umgewandelt, bestehende Konkordanzen zwischen Ordnungssystemen genutzt[5] und neue inhaltliche Ähnlichkeitsrelationen gebildet werden. Die Konstanzer Systematik sowie die erstellten Ähnlichkeitsrelationen zur Konstanzer Systematik sollen als Linked Data im Semantic Web zur Verfügung gestellt werden.

Dieses System soll keine Einbahnstraße sein. Die in Konstanz bereits geleistete intellektuelle Sacherschließungsarbeit kann so ebenfalls in andere Ordnungssysteme übersetzt werden. Von der Interoperabilität der Daten würden alle Seiten profitieren.

Insgesamt geht es bei diesem Projekt darum, wie die bereits intellektuell geleistete Sacherschließungsarbeit maschinell ausgewertet und für andere Systeme fruchtbar gemacht und nachgenutzt werden kann.

3. Konzept

Damit die Maschine eine gute Unterstützung leisten kann, benötigt sie einerseits Daten und andererseits Methoden, mit diesen Daten zu arbeiten. Die Methoden müssen exakt den Umgang mit den Daten beschreiben, sodass neue, idealerweise relevantere Daten daraus entstehen können. Je intelligenter die Daten verarbeitet werden sollen, desto spezifischere, aber auch mengenmäßig

4 Hafner, Ralph; Schelling, Bernd: Automatisierung der Sacherschließung mit Semantic Web Technologie. http://www.kim.uni-konstanz.de/das-kim/projekte-und-mitgliedschaften/aktuelle-projekte/automatisierte-sacherschliessung/ (08.11.2015).

5 Darunter: Das DFG-Projekt CrissCross (Fachhochschule Köln; Deutsche Nationalbibliothek: CrissCross. http://ixtrieve.fh-koeln.de/crisscross/index.html (15.10.2015)). Das Projekt Cocoda (Colibri Concordance Database for library classification systems) als Teilprojekt von „coli-conc" (Balakrishnan, Uma; Krausz, Andreas: Cocoda - ein Konkordanztool für bibliothekarische Klassifikationssysteme. https://opus4.kobv.de/opus4-bib-info/frontdoor/index/index/year/2015/docId/1676 (27.08.2015)). Einen guten Überblick über Projekte zur automatisierten Sacherschließung gibt: Kasprzik, Anna: Automatisierte und semiautomatisierte Klassifizierung - eine Analyse aktueller Projekte. In: Perspektive Bibliothek 3 (2014), S. 85–110. http://journals.ub.uni-heidelberg.de/index.php/bibliothek/article/view/14022 (12.12.2015).

mehr Methoden sind erforderlich. Im Falle von klassifikatorischen Sacherschließungsdaten wie im hier vorgestellten Projekt finden sich üblicherweise hierarchisierte Begriffssysteme in Form von kontextualisierten Wörtern oder Wortfolgen.[6]

Die Beschaffenheit einer bibliothekarischen Systematik exakt über (maschinenlesbare) formale Sprachen zu beschreiben ist sehr aufwändig, da Systematiken über einen langen Zeitraum gepflegt werden und dadurch gewissen praktischen aber auch kulturellen Zwängen unterliegen. Im Lauf der Zeit entstehen durch die Anpassungen Unzulänglichkeiten und teilweise auch Widersprüche. Lorenz kritisiert, dass „zur Wissensvermittlung verwendete Klassifikationen diese Entwicklung nur in verzögerter und abgeschwächter Weise darstellen können."[7] Für einen genauen Vergleich von mehreren Systematiken miteinander müsste jede auf eine vergleichbare Art maschinenlesbar und widerspruchsfrei beschrieben werden, damit die Maschine sie auswerten könnte. Besonders die Widerspruchsfreiheit zwischen mehreren Systemen würde eine Koordination zwischen den Erstellern der unterschiedlichen Systematiken erforderlich machen, die kaum umzusetzen sein dürfte. Sie würde zudem wesentliche Stärken einzelner Systematiken – nämlich die Optimierung für einen bestimmten Zweck – konterkarieren. Um die Komplexität und Größe des Problems zu reduzieren, werden in diesem Ansatz daher bewusst und kontrolliert Unschärfen eingebracht, die Widersprüche und Definitionslücken zulassen. Der Vergleich zwischen unterschiedlichen Systematiken kann so mit heuristischen Verfahren durchgeführt und mit Methoden aus dem Umfeld des Semantic Web optimiert werden.

In der klassischen künstlichen Intelligenz ist zur Beantwortung von natürlichsprachigen Fragen neben einem Verständnis der Grammatik auch ein Korpus für die Definition der Bedeutung von Wörtern erforderlich. Diese Definitionsmenge wird technische Ontologie genannt und üblicherweise als eine Menge von Faktensätzen in der Form Subjekt-Prädikat-Objekt in einer Datenbank gespeichert. Ihre exakte Erstellung ist aufgrund der zahlreichen Regeln, Ausnahmen und Unschärfen natürlicher Umgebungen eine große Herausforderung – insbesondere das Übersetzen in sich nicht widersprechende Regelsätze. Solche Ontologien müssen frei von Widersprüchen bleiben und dürfen keine Zirkelschlüsse[8] zulassen, wenn eine Baumstruktur gebildet werden soll, wie das in den meisten bibliothekarischen Systematiken der Fall ist. Formale Widerspruchs- und Zirkelfreiheit sind schwierige Probleme bei der Modellierung technischer Ontologien für natürlichsprachige Sachverhalte. Janich stellt die Herausforderungen so dar: „Wer sich aus diesem Widerspruch durch den Trick zu retten versucht, die Inhalte der Ontologie, also das Seiende und seine Eigenschaften, durch bloße Definition als das zu bestimmen, was wir wissen, also (etwa durch Wissenschaft) erkannt haben, verfällt dem Dilemma zwischen zwei weiteren Widersprüchen: er muß dafür entweder schon

6 Solche bibliothekarischen Klassifikationen werden von Menschen erstellt und gepflegt. Es gibt davon zahlreiche Vertreter – die Konstanzer Systematik ist ein Beispiel einer Klassifikation, die für die Gegebenheiten einer lokalen Institution optimiert wurde. Andere Vertreter, wie beispielsweise die DDC, werden weltweit zum Klassifizieren von Medien und zum Präsentieren von Beständen in Bibliotheken eingesetzt. Mit Kenntnissen der Konstanzer Systematik können sich Nutzerinnen und Nutzer nur in der Konstanzer Bibliothek orientieren, mit Kenntnissen der DDC finden sie sich in vielen Bibliotheken weltweit zurecht.

7 Lorenz, Bernd: Systematische Aufstellung in Vergangenheit und Gegenwart, Wiesbaden: Harrassowitz, 2003 (Beiträge zum Buch- und Bibliothekswesen 45), S. 294.

8 Ein Begriff verweist über Zwischenschritte wieder auf sich selbst.

Erkenntnis und Irrtum unterscheiden können (also das Seiende und seine Eigenschaften kennen), oder er muß annehmen, dass auch Irrtümer Abbilder von Eigenschaften tatsächlich existierender Gegenstände seien."[9] Es existiert bis heute kein Korpus, der eine natürliche Sprache vollständig erfassen könnte. Stattdessen wurden Ontologien für einzelne, klar umrissene Anwendungsgebiete geschaffen. Heute können Expertensysteme damit qualifiziert und zuverlässig bei der Beantwortung von fachspezifischen Fragen helfen. Der hier vorgestellte Ansatz macht sich diese Ergebnisse (z.B. in Form der Begriffszusammenhänge in der GND) zunutze, hat aber eine andere Zielsetzung: Da im hier vorliegenden Fall eine Systemstelle bereits bekannt ist (entweder im Quell- oder im Zielsystem), ist die Aufgabe des Verfahrens, möglichst gute Entsprechungen im anderen System zu finden. Dafür ist kein Verständnis der gesamten Systematiken notwendig, sondern man kann – analog zu einer der möglichen Vorgehensweisen bei der intellektuellen Tätigkeit des Systematisierens – mit einem Begriff starten und im zweiten Schritt den Kontext einbeziehen. Dadurch lassen sich die für die Maschine besonders aufwändigen vollständigen Betrachtungen aller Sonderfälle nahezu ausschließen und die Qualität von Suchergebnissen durch einen Vergleich des Kontexts beider involvierter Systeme ermitteln. Da ein anderes Ergebnis entsteht, wenn bereits eines der Systeme sich ändert, sprechen wir nicht von statischen Beziehungen wie Konkordanzen, sondern von *Verschränkungen* von Daten.

Notwendige Voraussetzungen für die Verschränkung sind pro Systematik mindestens zwei Dinge: ein kontrolliertes Vokabular und das Verständnis, wie die Begriffe hierarchisiert werden. Beim Vergleich mehrerer solcher Systematiken kommt noch eine weitere Bedingung hinzu: Unterscheiden sich die Vokabulare, müssen sie übersetzt oder übersetzbar gemacht werden. Die GND enthält zur Lösung bereits an vielen Stellen nutzbare Übereinstimmungen aus den Normdateien anderer Länder (z.B. RAMEAU und LCSH), über die beispielsweise synonyme Ähnlichkeiten gefunden werden können. Wo diese nicht vorhanden sind oder nicht ausreichen, können auch weitere Datenbanken helfen: DBpedia[10], Wikidata[11], WordNet[12] und OpenCyc[13]. Mit diesen Daten können semantische Ähnlichkeiten zwischen zwei lexikalisch unterschiedlichen Begriffen ermittelt werden. Gute Kandidaten für eine Übereinstimmung weisen gleichzeitig Ähnlichkeiten von mehreren Quellen sowie Ähnlichkeiten bei Hyponymen bzw. Hyperonymen auf. Mit ihnen lassen sich darüber hinaus andere semantische Zusammenhänge wie Teilmengen leicht darstellen – beispielsweise ob Konstanz in Deutschland liegt oder ob Datenbankdesign zu den Programmierwerkzeugen gehört.

9 Janich, Peter: Wozu Ontologie für Informatiker? Objektbezug durch Sprachkritik. In: Kurt Bauknecht; Wilfried Brauer; Thomas A. Mück (Hg.): Informatik 2001. Wirtschaft und Wissenschaft in der Network Economy - Visionen und Wirklichkeit. Tagungsband der GI/OCG Jahrestagung 2001, 25. - 28. September 2001 Universität Wien, Bd. 2. Wien: Österreichische Computer Gesellschaft, 2001, S. 765–769, hier: S. 766.

10 DBpedia ist eine Datenbank, die das Faktenwissen in unterschiedlichen Sprachen der Wikipedia in einem semantischen Netz darstellen möchte. DBpedia. http://wiki.dbpedia.org/ (13.10.2015).

11 Wikidata ist eine Faktendatenbank, aus der Daten wie beispielsweise Einwohnerzahlen direkt in Wikipedia-Artikeln einheitlich über verschiedene Sprachfassungen hinweg zitiert werden können. Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page (13.10.2015).

12 WordNet ist ein semantisch-lexikalisches Netz, das Bedeutungszusammenhänge der englischen Sprache modelliert. Princeton University: WordNet. A lexical database for English. https://wordnet.princeton.edu/ (13.10.2015).

13 Im OpenCyc-Projekt wird versucht, Alltagswissen in einem semantischen Netz so darzustellen, dass eine Maschine logische Fragen dazu beantworten kann. OpenCyc.org: OpenCyc for the Semantic Web. http://sw.opencyc.org/ (13.10.2015).

Bibliothekarische Systematiken teilen Themen in Unterthemen auf. Diese Eigenschaft wird im hier vorgestellten Ansatz zur semantischen Verortung genutzt. Die bislang in der lokalen Systematik als Zeichenketten vorliegenden Schlagwörter und Schlagwortfolgen werden, wo das möglich ist,[14] in semantisch verwertbare GND-Begriffe transformiert. Der Klassifikationsbaum wird so angepasst, dass er mittels eines Algorithmus von der Maschine so „verstanden" wird, dass die darin abgebildeten semantischen Zusammenhänge ersichtlich werden: Bezeichner für Ober- und Unterthemen (vertikale Navigation) werden genauso erkannt wie ähnliche (horizontale Navigation) oder verwandte Themen (linkartige Verknüpfung). Durch das Prinzip der Erstreckungen[15] sind Hierarchisierungen auch nachträglich leicht anzupassen und die Zirkelfreiheit ist gewährleistet. Mit dem in der Maschine gespeicherten Wissen über die hauseigene Systematik können fremde Systematiken nun spezifisch nach ähnlichen Begriffen oder Begriffskonzepten durchsucht werden. Wenn statt der eigentlichen Zeichenkette das dahinterliegende Konzept für einen Vergleich herangezogen wird, spricht man auch von einer semantischen Suche. Über die Auswertung des Kontextes eines Begriffs im fremden System werden diese Ergebnisse qualifiziert, also bewertet. Je ähnlicher der Kontext eines Begriffs in beiden verglichenen Systemen ist, desto besser wird ein Suchergebnis bewertet.

Zusammenfassend lässt sich sagen, dass im hier vorgestellten Verfahren nicht nach semantisch-epistemologisch eindeutigen Ergebnissen zweier miteinander verglichener Begriffe aus unterschiedlichen Systematiken gesucht wird, sondern in mehreren Stufen zunächst von einfachsten Ähnlichkeitssuchen zu komplexeren Verfahren übergegangen wird. Dadurch entsteht ein robustes Gesamtsystem, dessen Ergebnisse flexibel optimierbar sind. Auf allumfassende Ontologien kann verzichtet werden, da semantische Suchen über eine Heuristik abgebildet werden, die Lösungskandidaten ermittelt und durch Einbeziehung des Kontexts begriffliche Unschärfen reduziert. Kontext führt zur Verbesserung der Ergebnisse und kann somit stufenweise hinzugeschaltet werden. In der ersten Stufe zielt die Suche ausschließlich auf lexikalische Übereinstimmungen (z.B. „Schifffahrt" und „Schiffahrt"), in Stufe zwei auch auf naheliegende semantische Treffer (Synonyme: z.B. „Schifffahrt", „Nautik" und „Seewesen"; Homonyme werden ausgeschlossen). Erst ab Stufe drei werden Hyperonyme und Hyponyme einbezogen (z.B. „Navigation" etc.). Erforderlich sind hierzu zwei Metriken und semantisch verwertbare Daten bei zumindest einer der zu vergleichenden Systematiken. Konkret bedarf es einer Metrik für Distanz, also einem quantifizierbaren Unterschied zwischen zwei Begriffen, einer Metrik für die Zuverlässigkeit einer solchen Aussage und auf Seiten der Konstanzer Systematik (weil diese vor Ort selbst beeinflussbar ist) die Verwendung eines kontrollierten Vokabulars. Fremde Systematiken können auch über computerlinguistische Methoden „normalisiert" werden. Für Konstanz fiel die Entscheidung zur grundsätzlichen Verwendung des GND-Vokabulars, da sich die lokale Sacherschließung daran bereits orientiert. Die Quantifizierung der Distanz erfolgt aus einer Kombination der wie oben beschrieben ermittelten Stufe und der innerhalb einer Stufe festgestellten Unterschiede.

14 In Konstanz werden einzelne Schlagwörter eingesetzt, die für eine Aufstellungssystematik nötig, für eine Normdatei aber unnötig sind. Beispielsweise das Formschlagwort „Autor mit M".

15 Eine Erstreckung ist ein Intervall von Notationen mit definiertem Anfang und Ende. Anfang und Ende können in unterschiedlichen Notationsbereichen sein, beispielsweise spa 82.50 – spa 83 = Spanisch / Sprachunterricht.

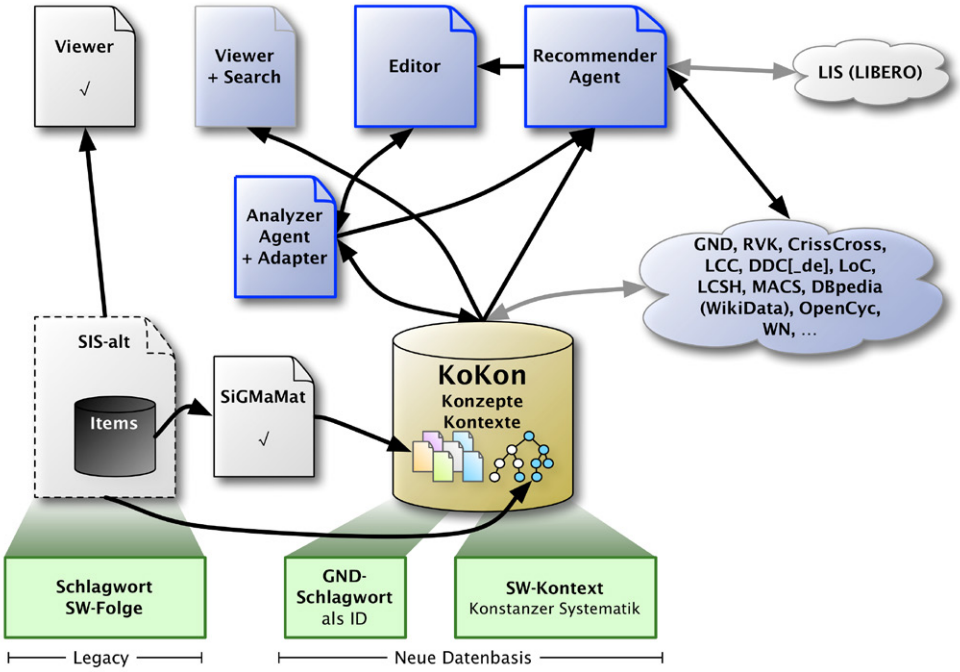

Abb. 1: Architektur des Gesamtsystems (eigene Darstellung)

4. Systemarchitektur

Abbildung 1 beschreibt die Strategie des gesamten Systems, welches in der Arbeit „KoKon – kontextsensitiver Vergleich für Klassifikationen" näher ausgeführt ist;[16] im Folgenden werden nun ihre Elemente anhand einer konkreten Infrastruktur aufgezeigt. Im Zentrum steht die Datenbasis „KoKon" (Zylinder in der Darstellung), welche sowohl die *Konzepte* der Konstanzer Systematik (die Systemstellen in Form semantisch verwertbarer Schlagwörter und Schlagwortfolgen) als auch den *Kontext* in Form einer Baumdarstellung der hierarchischen Systematik beinhaltet. Die in grau dargestellten Bereiche „SIS-alt", „Viewer" und „LIS (LIBERO)" bezeichnen die schon früher am KIM (Kommunikations-, Informations-, Medienzentrum) der Universität Konstanz existierenden Systeme, die als Quellen dienen. Dabei wird die Baumstruktur direkt aus der vorhandenen Systematik in KoKon eingespeist. Die Schlagwörter und Schlagwortfolgen werden über den – speziell zur Datenübernahme aus dem vorhandenen System entwickelten – SiGMaMat (s.u.) in semantisch verwertbare Identifizierer überführt. Die in blau dargestellten eckigen Objekte stehen für Anwendungen, die mit den Daten in KoKon arbeiten. Folgende Komponenten sind geplant:

16 Schelling, Bernd: KoKon. Kontextsensitiver Abgleich für Klassifikationen. Masterarbeit im Rahmen des weiterbildenden Fernstudiums, Berlin, 2014, s. Kap. 4.3, S. 42ff. (bislang unveröffentlicht).

- Ein Rechercheinstrument *Viewer + Search*,
- ein *Editor* für die Systematik (zum Bearbeiten des Strukturbaums, der Schlagwortfolgen sowie zur Pflege eigener Schlagwörter),
- ein *Analyzer Agent + Adapter*, der die notwendigen generischen Mechanismen zum Arbeiten mit den beiden Metriken für Distanz und Zuverlässigkeit für KoKon enthält und
- ein *Recommender Agent*, in dem andere Systematiken für die Verwendung mit dem *Analyzer Agent + Adapter* konfiguriert werden.

Die blau dargestellte Wolke bezeichnet mögliche Quellen, mit denen die Konstanzer Systematikdatenbank interagieren oder angereichert werden soll.

Die Gesamtarchitektur ermöglicht neben einer semantischen Verschränkung der hauseigenen Systematik mit anderen Systemen auch eine Verarbeitung der vorhandenen Daten wie die Pflege der Systematik und die Verarbeitung neu entstehender Sacherschließungsdaten. Die Werkzeuge SiGMaMat und Viewer, welche in diesem Kontext zur Vorbereitung der Arbeit mit verschränkten Systematiken bereits entstanden sind, werden im Folgenden näher erläutert.

5. Voraussetzungen und Werkzeuge für die automatisierte Sacherschließung

Damit eine (teil-)automatisierte Sacherschließung gelingen kann, müssen einige Voraussetzungen erfüllt sein bzw. werden. Alle Sacherschließungsdaten, die für das Programm genutzt werden sollen, müssen interoperabel sein und idealerweise als linked data vorliegen. Die Daten müssen zudem eindeutig sein, was sowohl für die Registereinträge der Klassifikationen gilt als auch für die hierarchische Struktur der Klassifikationen.

5.1. Normbegriffe dank *SiGMaMat*

Die Schlagwörter der Konstanzer Systematik liegen bislang lediglich als Zeichenketten vor und somit nicht in einem interoperablen Format. Sie waren teils mehrdeutig[17] und hatten lediglich eine interne Identifikationsnummer. Allerdings hat man sich in Konstanz bei der Vergabe der Registereinträge für die Systematik zumeist an die SWD (und zuletzt an die GND) gehalten. Ziel war es, aus den lediglich als Zeichenketten vorliegenden Schlagwörtern klare Konzepte mit einer eindeutigen Identifikationsnummer zu machen. Die lokalen Schlagwörter mussten demnach zunächst auf Begriffe aus Normdateien gemappt werden, so wie dies auch bei der RVK gemacht wurde.[18] Als erste und wichtigste Referenz wurde auch für die Konstanzer Systematik die GND gewählt. Zweitwichtigste Referenz ist die internationale Normdatei VIAF. Da die Konstanzer Systematik rund 180.000 Schlagwörter enthält, konnte das Mapping auf Normdateien nicht allein in Handarbeit

17 Vgl. z.B. das Schlagwort „Antigone". Was ist gemeint? Die Sagengestalt? Das Werk von Sophokles oder das gleichnamige Werk des französischen Autors Jean Anouilh?

18 Vgl. Peisl, Barbara: Register_GND-Projekt. http://rvk.uni-regensburg.de/index.php/34-rvko/inhalt/133-registergnd-projekt (22.09.2015).

erfolgen. Erstellt wurde daher, das Programm SiGMaMat[19] (SIS[20]-GND-Matching-Automat) eine Art Schlagwort-Waschmaschine, in die man die „schmutzigen" Schlagwörter steckt, die dann nach dem Programmdurchlauf als saubere Konzepte mit der eindeutigen ID einer Normdatei wieder herauskommen. Der SiGMaMat vergleicht die Zeichenketten der lokalen Schlagwörter mit denen aus den Normdateien GND und VIAF und mappt diese bei ausreichender Übereinstimmung.[21] Von den Schlagwörtern der Konstanzer Systematik konnten ca. 70% automatisch gemappt werden. Für die 30 %, die übrig bleiben, macht der SiGMaMat (zumeist) Vorschläge für Begriffe aus GND und/oder VIAF, auf die die lokalen Begriffe gemappt werden könnten.

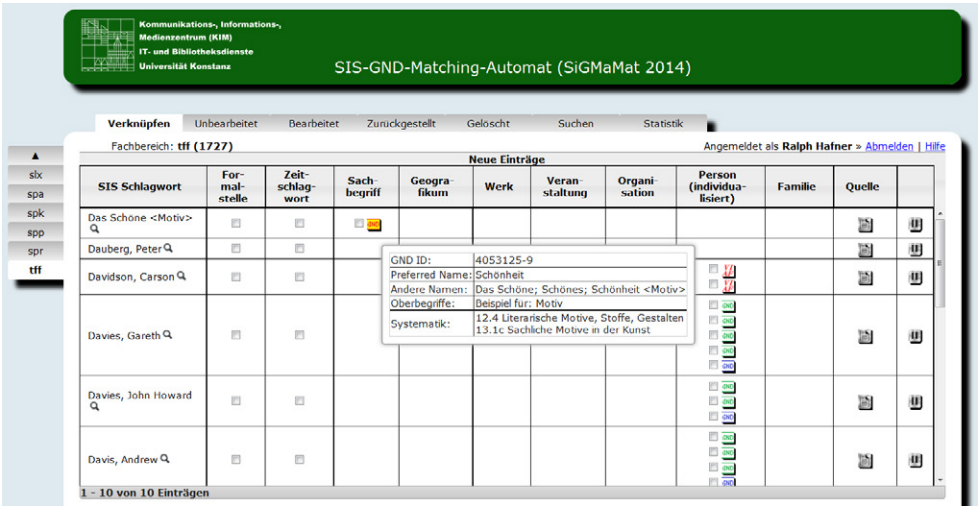

Abb. 2: Mappen lokaler Schlagwörter auf Normbegriffe mit dem SiGMaMat (Screenshot)

Der SiGMaMat ist seit Ende Januar 2015 in Konstanz produktiv im Einsatz.[22] Bis Oktober 2015 wurde die Hälfte der verbleibenden Schlagwörter, die nicht automatisch gemappt werden konnten, manuell von Fachreferent/inn/en und Hilfskräften mit Normbegriffen verknüpft. Das manuelle Mappen sollte somit bis spätestens Ende 2016 abgeschlossen werden können.[23]

19 Der SiGMaMat ist eine Eigenentwicklung des KIM der Universität Konstanz. Die Fertigstellung ging von 10/2013 bis 01/2015. Beteiligt waren Entwickler der IT-Abteilung Contentbasierte Dienste und der Sacherschließungsreferent.

20 SIS ist die Abkürzung für Schlagwort-Informations-System, eine Web-Anwendung, mit der bis dato die Konstanzer Systematik gepflegt wird.

21 Kriterien für automatisches Mappen waren z.B. die hundertprozentige Übereinstimmung mit „preferred name" oder „variant name" aus der Normdatei, wenn gleichzeitig diese Zeichenkette nicht mehrfach in der Normdatei vorkam. So kann „Balzac, Honoré de" (GND-Nummer: 118506358) automatisch gemappt werden, aber „Dumas, Alexandre" nicht, da sowohl <père> (1802–1870, GND-Nummer: 118528068) als auch <fils> (1824–1895, GND-Nummer: 11852805X) die gleiche Zeichenkette aufweisen. Weiterhin konnte auch die Übereinstimmung in der Schlagwortkategorie für automatisches Mapping hinzugezogen werden (auch in der Konstanzer Systematik haben (zumindest die meisten) Schlagwörter Kategorien wie „Person" oder „Geographikum" usw.).

22 Der SiGMaMat könnte auch dabei unterstützen, andere hauseigene Systematiken auf GND und VIAF zu mappen.

23 Vor der Arbeitserleichterung durch automatisierte Sacherschließung stand und steht hier also zunächst ein Mehraufwand an. Die Monate bis Ende 2016 sind natürlich nicht als volle Arbeitszeit zu verstehen, sondern als der Zeitraum, den es braucht, um neben den Standardaufgaben des Fachreferats mit dem Mapping fertig zu werden. Der SiGMaMat

5.2. Klare hierarchische Struktur der Klassifikation dank SIS-Visualizer

Eine weitere Voraussetzung für (teil-)automatisierte Sacherschließung ist, dass die Klassifikationen, mit denen gearbeitet werden soll, klare Strukturen mit eindeutigen hierarchischen Beziehungen aufweisen. Fehler in der Struktur der Konstanzer Systematik mussten und müssen korrigiert werden, fehlende Hierarchieebenen mussten und müssen noch ergänzt werden. Um die Struktur der Konstanzer Systematik in ihren Hierarchien darzustellen und damit die Fehler in der Struktur erkennen zu können, die sich in den knapp fünfzig Jahren der Geschichte der Konstanzer Systematik eingeschlichen haben, wurde vor Ort ein Programm zur Darstellung entwickelt – der SIS-Visualizer.[24] Die Webanwendung SIS, mit der bislang in Konstanz die Systematik bearbeitet und gepflegt wurde, konnte die Hierarchien noch nicht darstellen.[25]

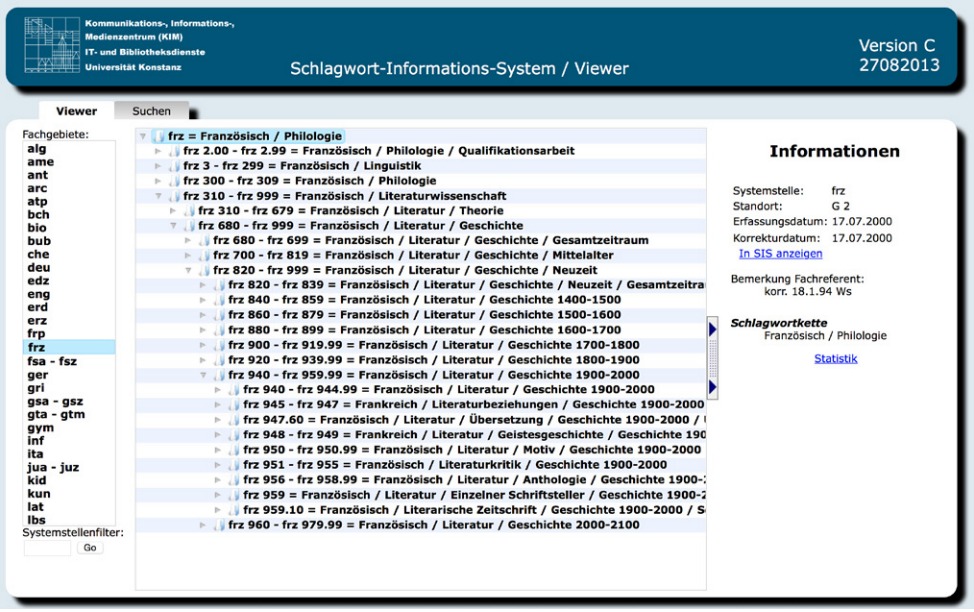

Abb. 3: SIS-Visualizer: Tool zur hierarchischen Darstellung der Konstanzer Systematik

erwies sich dabei als gute Unterstützung. Die Fachreferent/inn/en konnten sich zudem bei dieser Arbeit durch zusätzlich eingestellte studentische Hilfskräfte unterstützen lassen.

24 Der folgende Titel bietet eine detaillierte Beschreibung der Entwicklung des SIS-Visualizers: Kasprzik, Anna: Projektbericht: Implementierung eines Hierarchisierungsalgorithmus' für die Konstanzer Systematik. http://nbn-resolving.de/urn:nbn:de:bsz:352-241667.

25 Das „Schlagwort-Informations-System" (http://sis.ub.uni-konstanz.de/sis/notation.html). Beispiel aus der Sprachwissenschaft: http://sis.ub.uni-konstanz.de/cgi-bin/sis/stammsatz_notation.pl?fachgebiet1=spr&fachgruppe1=86&fachgruppe2=114&trefferzahl=100&suchart=alle. (04.10.2015)

6. Automatisierte Erschließung neuer Titel

Die im Folgenden beschriebene Vorgehensweise ist bisher nur ein Konzept, das Programm zu seiner Umsetzung ist erst in Vorbereitung. An zwei Beispieltiteln soll im Folgenden gezeigt werden, wie die automatisierte Erschließung eines in der Bibliothek neu erworbenen Titels aussehen könnte. Zunächst ein sehr einfaches Beispiel:

Said Sahel, Ralf Vogel, Einführung in die Morphologie des Deutschen, Darmstadt 2013. 978-3-534-24953-4.

Im ersten Schritt sucht das Programm anhand der ISBN in verschiedenen Bibliothekskatalogen nach vorhandenen Sacherschließungsinformationen – GND-Schlagwörtern, RVK-Notationen, DDC-Notationen – und sammelt diese. Es findet dabei u.a. folgende Schlagwortfolge mit GND-Begriffen: *Deutsch; Morphologie <Linguistik>; Einführung*.

Dann prüft das Programm anhand der GND-ID jedes einzelnen Schlagwortes, ob es auch in der Konstanzer Systematik vorkommt. Im vorliegenden Fall stellt es fest, dass alle drei Schlagwörter dort vorkommen und an einer Notationsstelle sogar genau diese drei gemeinsam und zudem in der gleichen Reihenfolge:

deu 140:n = Deutsch / Morphologie <Linguistik> / Einführung

Damit konnte das Programm für diesen Titel eine zu 100% passende Notation in der Konstanzer Systematik finden. In diesem Fall wäre die Übertragung der bereits vorhandenen Sacherschließungsdaten auf die Konstanzer Systematik trivial und eine vollautomatische Sacherschließung möglich.

Beim nächsten Beispiel-Titel wird es etwas komplizierter:

Ann Marie Fallon, Global Crusoe. Comparative literature, postcolonial theory and transnational aesthetics, Farnham 2011. 978-1-4094-2998-2.

Im ersten Schritt sammelt auch hier das Programm anhand der ISBN die in verschiedenen Bibliothekskatalogen bereits vorhandenen Sacherschließungsinformationen. Das Programm findet u.a. das GND-Werktitel-Schlagwort *Defoe, Daniel / Robinson Crusoe* mit der ID 4281761-4. Es überprüft im nächsten Schritt, ob und wenn ja, wo diese ID in den Registereinträgen der Konstanzer Systematik vorkommt und stößt auf folgende Einträge:

eng 919:d314:kr61 = Defoe, Daniel / Robinson Crusoe / Ausgabe
eng 919:d314:yr61 = Defoe, Daniel / Robinson Crusoe / Sekundärliteratur

Anhand des Kontextes muss das Programm jetzt entscheiden, welche Notationen es empfiehlt und welche nicht – in diesem Beispiel, ob es sich um eine Ausgabe des Werktitels handelt oder um Sekundärliteratur dazu. Das kann bei der Schlagwort-Kategorie „Werktitel" über die Übereinstimmung

oder Nicht-Übereinstimmung des Autors des zu systematisierenden Titels mit dem beim Werktitelschlagwort angegebenen Verfasser geklärt werden. In diesem Fall kommt das Programm zu dem Ergebnis, dass es sich um Sekundärliteratur zu *Robinson Crusoe* handelt. Es empfiehlt die Notation *eng 919:d314:yr61*.

Für die Wahl der treffendsten Notation reicht der Kontext noch nicht aus. Weitere vorhandene Sacherschließungsdaten müssen ausgewertet werden. Der Titel wurde zudem mit folgender Schlagwortfolge versehen:

Crusoe, Robinson (GND-Nummer: 119139243); Rezeption (GND-Nummer: 4049716-1); Postkoloniale Literatur (GND-Nummer: 4428936-4)

In dieser Schlagwortfolge ist *Crusoe, Robinson* nicht gleich dem *Robinson Crusoe* in der ersten Schlagwortfolge. Hier ist es das Personenschlagwort für die literarische Gestalt mit einer anderen ID. In der Konstanzer Systematik kommt es an folgenden Stellen vor:

inf 467:a770 = Literarische Gestalt / Bibliographie[26]
lit 770:r65 = Robinson Crusoe

Die erste Stelle scheidet aus, da es in keinem Katalog einen Hinweis darauf gibt, dass es sich um eine Bibliographie handelt.[27] Die zweite Stelle – die der literarischen Figur in der allgemeinen Literaturwissenschaft – hingegen ist relevant.

Das Schlagwort *Rezeption* kommt mehr als 2000 Mal in der Konstanzer Systematik vor. Als einzelnes Schlagwort hilft es demnach nicht weiter. Im nächsten Schritt kann getestet werden, ob es in Kombination mit einem der beiden Schlagwörter aus der Schlagwortfolge oben gemeinsam in der Konstanzer Systematik vorkommt, also mit *Crusoe, Robinson* oder mit *Postkoloniale Literatur*. Dies ist aber nicht der Fall.

Für das Schlagwort *Postkoloniale Literatur* gibt es in der Konstanzer Systematik Notationsstellen in der amerikanischen, der englischen, der französischen, der italienischen und der spanischen sowie der allgemeinen Literaturwissenschaft. In den bisher ausgewerteten Sacherschließungs-Informationen finden sich Hinweise darauf, dass englische und allgemeine Literaturwissenschaft in Frage kommen. Das wären folgende Stellen:

26 An dieser Notationsstelle wird „Crusoe, Robinson" als „siehe unter"-Verweis aufgeführt. Daher wird diese Stelle bei der Suche nach Crusoe angezeigt.

27 Das wirft die Frage auf, wie viele der Schlagwörter aus einer Schlagwortfolge, die die Klassenbezeichnung einer Notation ausmachen, übereinstimmen müssen, damit eine Notation vom Programm empfohlen wird. In diesem Fall ist die Übereinstimmung in der Klassenbezeichnung gleich Null. Eine 50%-Übereinstimmung wird in einer siehe-unter-Verweisung erzielt. Zudem befindet sich die Stelle in einem Fach (inf wie Informationsliteratur), auf das wir keine weiteren Hinweise haben.

eng 944:n = Englisch / Postkoloniale Literatur
lit 108 = Kolonialroman / Gattungsgeschichte[28]

Die Notationen aus den anderen Philologien sortiert das Programm aus, da sie aus „falschen" Systematikzweigen kommen.

Neben den GND-Schlagwörtern enthält der Titel *Global Crusoe* noch weitere Sacherschließungsinformationen, u.a. DDC-Notationen, von denen eine hier exemplarisch ausgewertet werden soll:

Die DDC-Notation: *809.387* ist die Stelle für den Abenteuerroman. Sie ist mit dem GND-Sachschlagwort *Abenteuerroman* (GND-ID: 4000089-8) verknüpft. Sucht man wiederum damit in der Konstanzer Systematik, stößt man auf folgende Systemstellen:

lit 464 = Abenteuerroman
deu 588:a14 = Deutsch / Abenteuerroman / Geschichte
eng 588:a14 = Englisch / Abenteuerroman / Geschichte
frz 588:a14 = Französisch / Abenteuerroman / Geschichte
spa 588:a14 = Spanisch / Abenteuerroman / Geschichte

Durch einen Vergleich der Strukturbäume der Konstanzer Systematik und der DDC lässt sich ermitteln, welche der Konstanzer Notationen am besten zu dieser DDC-Notation passt. Die Notation *809.387* ist Teil des Bereichs *800–809 = allgemeine Literaturwissenschaft*. Damit wäre die Konstanzer Notation *lit 464* die beste Entsprechung. Literatur speziell zum englischen Abenteuerroman würde sich bei Dewey im Notationsbereich *820–829* befinden. Trotzdem könnte das Programm zusätzlich die Konstanzer Notation eng *588:a14* für den Abenteuerroman in der englischen Literatur vorschlagen, da es aus anderen Sacherschließungsdaten Hinweise auf die englische Literatur hat. Die Notationsstellen aus den anderen Literaturen sortiert das Programm wieder aus.

Das Programm hat nun eine Reihe von Ähnlichkeiten zwischen den bereits gegebenen Sacherschließungsdaten und Systemstellen in der Konstanzer Systematik zum Titel *Global Crusoe* gefunden.

Es wird im nächsten Schritt die Ergebnisse bewerten: Hierfür wird es sich die im Abschnitt *Konzept* erwähnte Metrik für Distanz zunutze machen. Die Übereinstimmungen mit der geringsten Distanz werden am besten bewertet. Dabei wertet das Programm eine Übereinstimmung in einer Schlagwortfolge höher als die Übereinstimmung einzelner Schlagwörter. Ein eindeutiges Ergebnis, das zu einer automatischen Erschließung führen kann, lässt sich aber mangels maschinenverwertbaren Wissens bei diesem zweiten Beispieltitel nicht ermitteln. Die Bewertung mehrerer übereinstimmender Einzelschlagwörter aus unterschiedlichen Schlagwortfolgen oder ein Vergleich von Notationen kann mit diesem Werkzeug zwar nicht automatisch umgesetzt, aber so aufbereitet werden, dass die Entscheidung durch den Menschen schneller zu treffen ist als ohne dieses Hilfsmittel. Bei diesen

28 Auch dieser Treffer erscheint wegen einer der Notation zugeordneten Verweisungskette, die „Postkoloniale Literatur" enthält.

komplexeren Fällen schlägt das Programm Konstanzer Systemstellen vor, die ein/e Fachreferent/in als gut bestätigt oder als unpassend ablehnt. Auch wählt in diesen Fällen der / die Fachreferent/in die Systemstelle für die Aufstellung selbst aus.

7. Fazit

Die vernetzte Datenwelt macht die Isolation, in der die Konstanzer Systematik steckte, deutlich. Der hier vorgestellte Ansatz zeigt, dass durch maschinelle Unterstützung auch hochgradig an lokale Bedürfnisse angepasste Strukturen den Anschluss an die vernetzte Datenwelt finden können. Innerhalb kurzer Zeit war es möglich, die hauseigenen Signaturen in einer hierarchisierten Baumstruktur darzustellen. Damit wurde über den SIS-Viewer eine Navigation zwischen den Systemstellen möglich. Durch das „Verstehen" und Normalisieren der Schlagwörter und Schlagwortfolgen wird der Bereich „maschinelles Verstehen der eigenen Daten" abgeschlossen. Anschließend folgt das Andocken fremder Datenquellen und die Modellierung der Algorithmen zur Verschränkung der hauseigenen Systematik mit anderen Ordnungssystemen. Damit wird maschinengestützte oder automatisierte Erschließung ermöglicht.

Literaturverzeichnis

- Balakrishnan, Uma; Krausz, Andreas: Cocoda - ein Konkordanztool für bibliothekarische Klassifikationssysteme. https://opus4.kobv.de/opus4-bib-info/frontdoor/index/index/year/2015/docId/1676 (27.08.2015).

- Bösing, Laurenz; Stoltzenburg, Joachim; Thomashoff, Barbara: Regeln für den Aufbau von Buchsignaturen. (Überarbeitete und auf den neuesten Stand gebrachte Fassung der Regeln vom April 1967 (B/Sto/Th) unter Berücksichtigung aller späteren Anhänge und Ergänzungen), Konstanz: Bibliothek der Universität Konstanz, 1969 (Bibliothek aktuell / Sonderheft 1).

- DBpedia. http://wiki.dbpedia.org/ (13.10.2015).

- Fachhochschule Köln; Deutsche Nationalbibliothek: CrissCross. http://ixtrieve.fh-koeln.de/crisscross/index.html (15.10.2015).

- Hafner, Ralph; Schelling, Bernd: Automatisierung der Sacherschließung mit Semantic Web Technologie. http://www.kim.uni-konstanz.de/das-kim/projekte-und-mitgliedschaften/aktuelle-projekte/automatisierte-sacherschliessung/ (08.11.2015).

- Janich, Peter: Wozu Ontologie für Informatiker? Objektbezug durch Sprachkritik. In: Kurt Bauknecht; Wilfried Brauer; Thomas A. Mück (Hg.): Informatik 2001. Wirtschaft und Wissenschaft in der Network Economy - Visionen und Wirklichkeit. Tagungsband der GI/OCG Jahrestagung 2001, 25. - 28. September 2001 Universität Wien, Bd. 2. Wien: Österreichische Computer Gesellschaft, 2001, S. 765–769.

– Kasprzik, Anna: Projektbericht: Implementierung eines Hierarchisierungsalgorithmus' für die Konstanzer Systematik. http://nbn-resolving.de/urn:nbn:de:bsz:352-241667 (10.11.2015).

– Kasprzik, Anna: Automatisierte und semiautomatisierte Klassifizierung - eine Analyse aktueller Projekte. In: Perspektive Bibliothek 3 (2014), S. 85–110. http://journals.ub.uni-heidelberg.de/index.php/bibliothek/article/view/14022 (12.12.2015).

– Lorenz, Bernd: Systematische Aufstellung in Vergangenheit und Gegenwart, Wiesbaden: Harrassowitz, 2003 (Beiträge zum Buch- und Bibliothekswesen 45).

– Müller-Dreier, Armin: Einheitsklassifikation. Die Geschichte einer fortwirkenden Idee, Wiesbaden: Harrassowitz, 1994 (Beiträge zum Buch- und Bibliothekswesen 35).

– OpenCyc.org: OpenCyc for the Semantic Web. http://sw.opencyc.org/ (13.10.2015).

– Peisl, Barbara: Register_GND-Projekt. http://rvk.uni-regensburg.de/index.php/34-rvko/inhalt/133-registergnd-projekt (22.09.2015).

– Princeton University: WordNet. A lexical database for English. https://wordnet.princeton.edu/ (13.10.2015).

– Schelling, Bernd: KoKon. Kontextsensitiver Abgleich für Klassifikationen. Masterarbeit im Rahmen des weiterbildenden Fernstudiums, Berlin, 2014.

– Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page (13.10.2015).