

Semantische Suche nach wissenschaftlichen Videos – Automatische Verschlagwortung durch Named Entity Recognition

Foto: Carsten Schick



Sven Strobel

Foto: Christian Bierwagen



Margret Plank

Das Kompetenzzentrum für nicht-textuelle Materialien an der Technischen Informationsbibliothek hat gemeinsam mit dem Hasso-Plattner-Institut für Softwaresystemtechnik ein Portal für AV-Medien entwickelt, das den Zugang zu wissenschaftlichen Videos aus dem Bereich Technik und Naturwissenschaft optimiert. Die Videos des TIB/AV-Portals sind nur mit wenigen manuell-intellektuell erstellten Metadaten assoziiert, die das gesamte Video beschreiben. Die automatische Erschließung des TIB/AV-Portals beschreibt die Videos auf Segmentebene und ermöglicht dadurch eine zielgenaue Suche innerhalb der Videos. Die deutschsprachigen Videos des TIB/AV-Portals werden mit Sachbegriffen der Gemeinsamen Normdatei (GND) annotiert, die über semantische Relationen miteinander verknüpft sind. Die englischsprachigen Videos des TIB/AV-Portals werden mit Hilfe von englischsprachigen Bezeichnungen verschlagwortet, die durch ein Mapping der GND-Sachbegriffe auf andere Normdaten (u.a. DBpedia) ermittelt wurden.

The Competence Centre for Non-Textual Materials at the German National Library of Science and Technology together with the Hasso Plattner Institute for IT Systems Engineering has developed a portal for AV media which optimises access to academic videos in the fields of technology and science. Each video in the TIB/AV Portal only has a small amount of metadata input by expert library staff. These metadata describe the video as a whole. The automatic indexing of the TIB/AV Portal describes the videos at the segment level, thereby permitting targeted searches within the videos. The German videos in the TIB/AV Portal are annotated with subject headings from the Integrated Authority File (GND) which are interlinked via semantic relations. The English videos in the TIB/AV Portal are given subject headings using English labels which have been determined by mapping the GND subject headings to other authority files (including DBpedia).

1. EINLEITUNG: NEUE BEDARFE, NEUE LÖSUNGEN

Wie findet man eigentlich etwas in Videos? Diese Frage stellt sich beispielsweise der Studierende, der eine bestimmte Sequenz in der Vorlesungsaufzeichnung aus der letzten Woche sucht, die Filmschaffende, die ihre Reportage mit Bildern untermauern möchte, oder der Mitarbeiter eines Museums, der Anschauungsmaterial für eine Ausstellung benötigt. Um einen Film zu finden, durchsuchen die Nutzerinnen und Nutzer die Filmarchive mittels der vorhandenen Metadaten. In den meisten Fällen handelt es sich dabei um Metadaten, die das gesamte Video beschreiben wie Titel, Autor oder Format. Es gibt kaum Beschreibungen einzelner Videosegmente oder Shots. Bibliotheken und Archiven fehlt es schlicht an Ressourcen, um diese Arbeit bewältigen zu können. Angesichts der stark wachsenden Anzahl an digitalen AV-Medien und der Notwendigkeit, diese auf Segmentebene zu erschließen, gibt es einen großen Bedarf an Lösungen für eine

automatische Erschließung. Dabei geht es nicht darum, auf die manuell-intellektuelle Erschließung zu verzichten, sondern die autoritativen Metadaten mit automatisch generierten, feingranularen Metadaten zu ergänzen und somit die Suchmöglichkeiten für die Nutzerinnen und Nutzer in und nach Videos zu verbessern.

Das Kompetenzzentrum für nicht-textuelle Materialien (KNM)¹ an der Technischen Informationsbibliothek (TIB)² arbeitet an grundlegenden Verbesserungen der Zugangs- und Nutzungsbedingungen für audiovisuelle Medien, 3D-Objekte und Forschungsdaten aus den sechs TIB-Kernfächern Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik. Ziel der Entwicklungsaktivitäten im KNM ist es, eine Infrastruktur zu schaffen und Werkzeuge zu entwickeln, die Nutzerinnen und Nutzer (Rezipienten und Produzenten) im wissenschaftlichen Arbeitsprozess aktiv unterstützen. Nicht-textuelle Objekte sollen so einfach publiziert, gefunden und dauerhaft bereitgestellt werden können wie textuelle Dokumente.

In einem gemeinsamen Projekt mit dem Hasso-Plattner-Institut für Softwaresystemtechnik an der Universität Potsdam (HPI)³ wurde in diesem Kontext ein Portal für AV-Medien entwickelt, das den Zugang zu wissenschaftlichen Videos wie z. B. Computervisualisierungen, Lernmaterialien, Simulationen, Experimenten, Interviews, Vorlesungs- oder Konferenzaufzeichnungen aus Technik und Naturwissenschaft optimiert. Das wesentliche Merkmal des Portals ist die Kombination aus State of the Art Multimedia-Retrieval-Verfahren und semantischer Analyse (Plank/Neumann 2014). Die Entwicklung orientiert sich an den 2010 durchgeführten Anforderungs- und Nutzungsanalysen. 2011 ist ein teilfunktionaler Prototyp des TIB/AV-Portals entwickelt worden, 2012–2013 erfolgte die Weiterentwicklung und der Betabetrieb des Systems, und im Frühjahr 2014 ist der Vollbetrieb in der TIB aufgenommen worden.

Die Prozesskette des TIB/AV-Portals beginnt mit dem Ingest der AV-Medien sowie der vom Autor mitgelieferten autoritativen Metadaten wie Titel, Autor, Herausgeber etc. in ein Medienmanagementsystem. Die *Shot Boundary Detection* segmentiert das Video anschließend anhand von Bildmerkmalen in einzelne Shots und erzeugt Keyframes für einen visuellen In-

Prozesskette des
TIB/AV-Portals

dex. Nach dieser strukturellen Analyse werden Text-einblendungen (z. B. auf Folien) von der Optischen Zeichenerkennung (OCR) analysiert und in Form eines OCR-Transkripts gespeichert. Zudem wird, basierend auf der Audiospur des AV-Mediums, mittels automatischer Spracherkennung eine Transkription generiert. Aus den Audio- und OCR-Transkripten extrahiert die *Named Entity Recognition* (NER) Sachbegriffe der Gemeinsamen Normdatei (Sack/Plank 2014). Schließlich werden mit Hilfe der *Visual Concept Detection* Bildinhalte entsprechend vordefinierter fachspezifischer und fächerübergreifender Kategorien wie z. B. Landschaft, Maschine, Zeichnung, Animation und Vorlesung klassifiziert (Hentschel/Blümel/Sack 2013). Die automatische Videoanalyse des TIB/AV-Portals besteht demnach aus einer strukturellen Analyse sowie einer Text-, Audio- und Bildanalyse.

Der vorliegende Beitrag konzentriert sich auf die automatische Verschlagwortung und semantische Suche des TIB/AV-Portals. Zunächst beschreibt Kapitel 2 die automatische Annotation der deutschsprachigen Videos mit GND-Sachbegriffen (GND-Entitäten). Kapitel 3 beschäftigt sich mit der fachbezogenen Annotation der Videos auf der Basis von erweiterten GND-Fachauschnitten. In Kapitel 4 wird gezeigt, wie GND-Entitäten im Fall von Mehrdeutigkeiten bei der Videoannotation disambiguiert werden. Die semantische Suche des TIB/AV-Portals, die auf den GND-Entitäten aufsetzt, ist Thema des 5. Kapitels. Kapitel 6 behandelt die Verschlagwortung der englischsprachigen Videos des TIB/AV-Portals. Der Artikel schließt mit einer kurzen Zusammenfassung und einem Ausblick auf geplante Vorhaben.

2. VERSCHLAGWORTUNG DEUTSCH- SPRACHIGER VIDEOS MIT HILFE DER GND

Die NER des TIB/AV-Portals bewerkstelligt die automatische Verschlagwortung der Videos. NER bezeichnet hier die Zuordnung von Begriffen aus einem Referenzvokabular auf analysierte Textinhalte. Die analysierten Textinhalte sind im Fall des TIB/AV-Portals die autoritativen Metadaten, die von Autoren oder Anbietern der AV-Medien beim Upload mitgeliefert werden, sowie die Audio- und OCR-Transkripte, die automatisch zum gesprochenen und eingeblendeten Text des Videos generiert werden.

Das Referenzvokabular für die automatische Verschlagwortung der deutschsprachigen Videos im TIB/AAV-Portal sind Sachbegriffe der Gemeinsamen Normdatei (GND). Die GND-Datensätze sind unter der CC0-Lizenz als Linked Data publiziert.⁴ In den Normdatensätzen der GND sind auch Relationen zu ande-

ren Normdatensätzen enthalten. Auf diese Weise ist ein semantisches Netz von miteinander in Beziehung stehenden Datensätzen entstanden, das sich für eine automatisierte semantische Verschlagwortung von Videodokumenten anbietet (GND: www.dnb.de/gnd).

Das Verschlagwortungsvokabular sind GND-Sachbegriffe. Es wird im Folgenden allerdings häufig nicht von GND-Sachbegriffen, sondern von GND-Entitäten die Rede sein. Was ist damit gemeint? *Entität* ist ein Grundbegriff der Ontologie. Entitäten beschreiben Dinge, d. h. Gegenstände, Individuen, Sachverhalte, Ereignisse etc., die in einem konkreten oder abstrakten Sinne existieren (Precht/Burkard 1999, S. 135). Die GND enthält Entitäten für die verschiedensten Dinge, z. B. für PKWs, Zentrifugen, Thermodynamik und Differentialgleichung. Entitäten haben bestimmte *Bezeichnung*. Bezeichner sind die Namen (sprachlichen Zeichen) für die Entitäten. Beispielsweise hat die GND-Entität »Orange« die zwei Bezeichner *Orange* und *Apfelsine*.

Die GND, die diese Entitäten enthält, repräsentiert eine eigene Ontologie. *Ontologie* bedeutet hier in der Begrifflichkeit der Informatik eine »explizite, formale Spezifikation einer gemeinschaftlichen Konzeptualisierung« (Gruber 1995, S. 908), d. h. »ein abstraktes Modell (Konzeptualisierung), das alle relevanten Begriffe innerhalb einer Domäne und deren Beziehungen untereinander abbildet« (Sack 2010, S. 16). Da GND-Entitäten untereinander in semantischen Beziehungen wie Synonymie, Hyperonymie und Hyponymie stehen, kann auf ihnen eine semantische Suche aufgesetzt werden, die die traditionelle Schlüsselwort-basierte Suche verbessert (vgl. Kapitel 5).

Die Videos des TIB/AV-Portals sind in einzelne Segmente unterteilt, die die entsprechenden Ausschnitte der Audio- und OCR-Transkripte enthalten. Die NER extrahiert aus den Transkripten GND-Entitäten und weist diese den einzelnen Videosegmenten zu (vgl. Abb.). Damit ist eine zielgenaue, segmentbasierte Suche nach Wörtern in Audio- und OCR-Transkripten sowie nach Entitäten möglich.

3. FACHBEZOGENE VERSCHLAG- WORTUNG MIT ERWEITERTEN GND-FACHAUSSCHNITTEN

Die deutschsprachigen Videos des TIB/AV-Portals werden mit GND-Sachbegriffen verschlagwortet. Dabei hat jedes der sechs TIB-Fächer seinen eigenen Fachauschnitt aus der GND. Mit anderen Worten: Architektur-Videos werden mit GND-Sachbegriffen aus der Architektur verschlagwortet, Chemie-Videos mit GND-Sachbegriffen aus der Chemie etc. Die sechs GND-Fachauschnitte sind leicht erweitert worden, d. h. sie enthalten Sachbegriffe zum jeweiligen Fächerkern

GND-Sachbegriffe als
Entitäten

Verschlagwortung auf
Segmentebene

GND-Fachauschnitte für
die TIB-Kernfächer

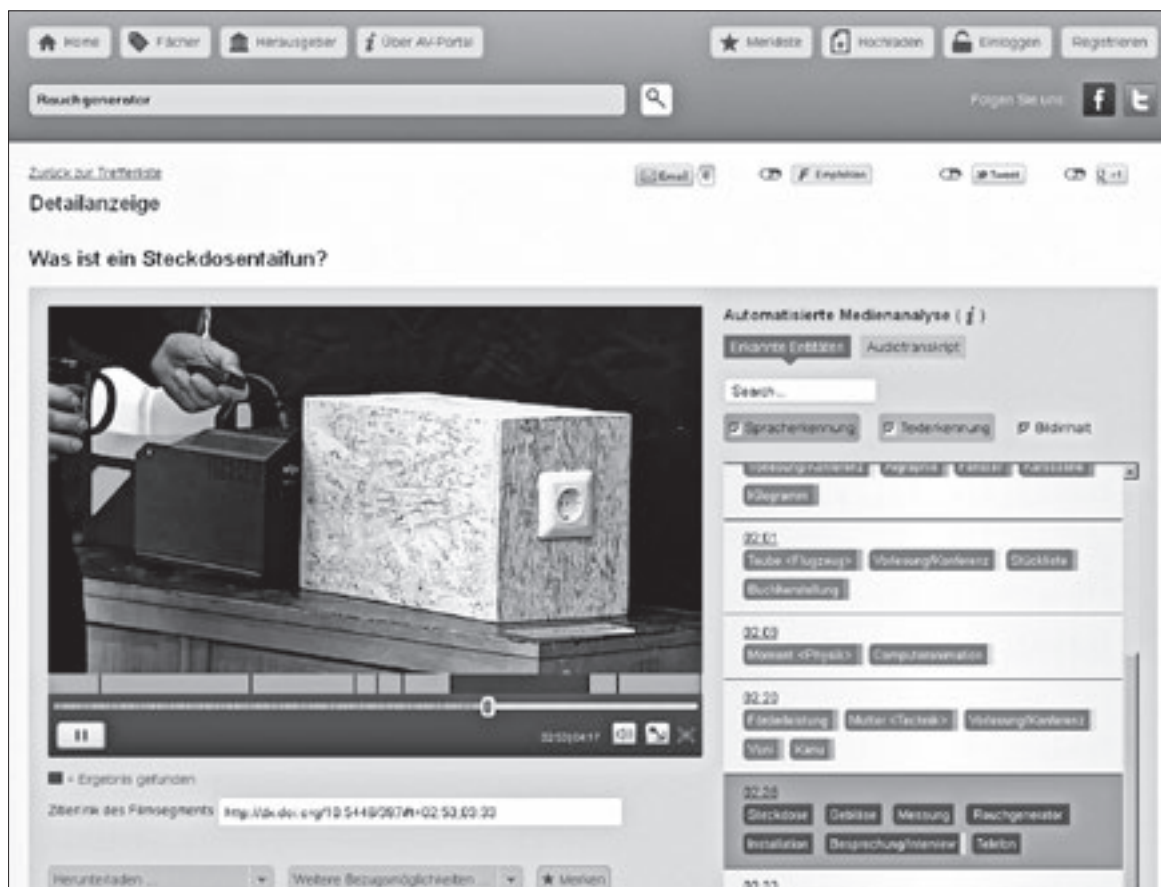


Abb.: Detailanzeige des TIB/AV-Portals der TIB. Auf der rechten Seite sind die GND-Entitäten der einzelnen Videosegmente aufgelistet

plus eine kleine Erweiterung um Klassen, die einen sehr engen Fachbezug haben. Zum Beispiel enthält der erweiterte GND-Fachausschnitt für Physik neben den Sachbegriffen aus der Physik auch Sachbegriffe aus der Astronomie, Elektrotechnik und Physikalischen Chemie. Insgesamt enthalten die sechs erweiterten GND-Fachausschnitte 63.356 Sachbegriffe.

Je mehr man das Vokabular erweitert, mit dem die Videos verschlagwortet werden, desto mehr relevante Begriffe werden erfasst (Verbesserung des Recall). Allerdings nimmt mit der Erweiterung des Vokabulars die »fachliche« Genauigkeit der Begriffe ab (Verschlechterung der Precision). Umgekehrt verhält es sich mit der Reduzierung des Vokabulars. Die fachbezogene Verschlagwortung der Videos auf Basis von erweiterten Fachausschnitten kann als eine Art Mittelweg angesehen werden. Man verliert dadurch voraussichtlich nicht zu viele relevante Begriffe, muss aber auch nicht zu viele »fachfremde« Begriffe bei der Verschlagwortung in Kauf nehmen.

4. DISAMBIGUIERUNG BEI DER NAMED ENTITY RECOGNITION

Die NER extrahiert aus den autoritativen Metadaten sowie automatisch generierten Audio- und OCR-Transkripten GND-Entitäten des jeweiligen Fachausschnitts und verschlagwortet damit das Video semantisch. Dabei können häufig mehrere GND-Entitäten einem analysierten Textbestandteil zugeordnet werden, d. h. die Zuordnung ist nicht eindeutig. Es muss in diesen Fällen eine Disambiguierung der Entitäten erfolgen. Der Disambiguierungsalgorithmus gleicht die Kontextinformationen der in Frage kommenden GND-Entitäten mit der textuellen Umgebung des ambigen Textbestandteils ab und versucht über eine Kookkurrenzanalyse die Mehrdeutigkeit aufzulösen.

Wenn beispielsweise sowohl die Entität »Wind« (Luftbewegung) als auch die Entität »Winde« (Mechanik) dem analysierten Textbestandteil *Winde* zugeordnet werden können, müssen die beiden Entitäten disambiguiert werden. Eine Disambiguierung benötigt immer Kontextinformationen. Nehmen wir an, die Kontextinformationen von »Wind« (Luftbewegung) seien Begriffe wie *Luftströmung*, *Meteorologie*

und *Windgeschwindigkeit*; die Kontextinformationen von ›Winde‹ (Mechanik) seien Begriffe wie *Hebezeug*, *Hebel* und *Wellrad*. Stehen nun in der unmittelbaren Umgebung des analysierten Textbestandteils *Winde* die Worte *Wellrad* und *Hebel*, kann der Disambiguierungsalgorithmus die Entität ›Winde‹ (Mechanik) mit hoher Wahrscheinlichkeit bezüglich der kontextuellen Richtigkeit dem Textbestandteil *Winde* zuordnen und somit die Mehrdeutigkeit auflösen.

Die Kontextinformationen der GND-Entitäten der TIB/AV-Portal-Wissensbasis bestehen aus den Oberbegriffen, Alternativbezeichnungen, verwandten Begriffen und Definitionen der GND-Entitäten. Zusätzlich gehören dazu die mit den GND-Entitäten korrespondierenden Wikipedia-Artikeltexte und die induzierten Linkgraphen der korrespondierenden DBpedia-Entitäten, die aus dem Mapping der GND-Entitäten auf die DBpedia resultieren (vgl. Kapitel 6).

5. SEMANTISCHE SUCHE IM TIB/AV-PORTAL

In der traditionellen Schlüsselwort-basierten Suche werden nur diejenigen Dokumente zurückgeliefert, die das eingegebene Suchwort enthalten. Die semantische Suche rekurriert hingegen auf eine Wissensbasis und kann darüber hinaus Videodokumente zurückliefern, die z. B. Synonyme, Ober- oder Unterbegriffe des eingegebenen Suchworts enthalten.

Die semantische Suche des TIB/AV-Portals basiert auf GND-Entitäten. Die Verschlagwortung mit Entitäten ermöglicht eine Verbesserung der Schlüsselwort-basierten Suche sowohl durch eine Erweiterung der Suchergebnisse als auch durch eine Präzisierung der Suchergebnisse. Erweitert werden die Suchergebnisse im TIB/AV-Portal durch den Einbezug von Synonymen und englischen Übersetzungen. Das kann an dem Beispiel der GND-Entität ›Kernenergie‹ illustriert werden. Die GND-Entität ›Kernenergie‹ hat den Hauptbezeichner *Kernenergie*, u. a. die Alternativbezeichner *Atomenergie*, *Nuklearenenergie* und *Kernkraft* sowie den englischen Bezeichner *Nuclear Energy*. Wird nun einer der Bezeichner der Entität ›Kernenergie‹ als Suchwort eingegeben, werden alle übrigen Bezeichner mitgesucht. Damit wird die Vollständigkeit der relevanten Videodokumente in der Treffermenge erhöht.

Präzisiert werden können die Suchergebnisse des TIB/AV-Portals durch eine inhaltsbasierte Facetten-suche. Im TIB/AV-Portal gibt es Facetten für Fachgebiet, Herausgeber, Erscheinungsjahr, Lizenz, visuelle Konzepte, Personen, Organisationen und Sachbegriffe. Die Facette Sachbegriffe enthält die aus den textuellen Daten extrahierten GND-Entitäten. Sucht ein Nutzer zum Beispiel textuell nach *Beton*, bekommt er

eine erste Treffermenge ausgegeben, die mithilfe der GND-Entitäten ›Stütze‹, ›Tragverhalten‹ oder ›Baustoff‹ aus der Facette der Sachbegriffe inhaltlich gefiltert werden kann. Damit wird die Genauigkeit der Suchergebnisse erhöht. Der Nutzer kann im TIB/AV-Portal zuerst mit einer textuellen Suche einsteigen und dann die Suchergebnisse mithilfe der Facetten kontinuierlich verfeinern.

6. ENGLISCHSPRACHIGE NAMED ENTITY RECOGNITION

Für die Verschlagwortung der englischsprachigen Videos des TIB/AV-Portals fehlte zunächst ein englischsprachiges Begriffsvokabular. Die GND selbst enthält zu den im TIB/AV-Portal verwendeten Sachbegriffen aus Naturwissenschaft und Technik nur sehr wenige englischsprachige Bezeichner. Im Rahmen eines gemeinsamen Projekts mit dem HPI wurden die GND-Entitäten der sechs Fachausschnitte auf andere Normdaten gemappt, um englischsprachige Bezeichner zu den deutschsprachigen Bezeichnern der GND-Entitäten zu ermitteln (Steinmetz/Sack 2013). Zu diesen Normdaten gehören DBpedia⁵, Library of Congress Subject Headings (LCSH)⁶, Daten des Projekts Multi Lingual Access to Subjects (MACS)⁷ und der WTI-Thesaurus »Technik und Management«⁸. Die englischsprachigen Videos des TIB/AV-Portals können nunmehr mithilfe der ermittelten englischsprachigen Bezeichner verschlagwortet werden.

Zusätzlich zu den englischsprachigen Bezeichnern wurden durch das Mapping Kontextinformationen aus Wikipedia und DBpedia für die GND-Entitäten gewonnen, mit denen der Disambiguierungsalgorithmus der NER verbessert werden kann. Je mehr Kontextinformationen für die GND-Entitäten vorliegen, desto besser verläuft die Disambiguierung der Entitäten.

7. ZUSAMMENFASSUNG UND ZIELE: LINKED-OPEN-DATA-POLICY FÜR AV-MEDIEN

Die Videos des TIB/AV-Portals verfügen nur über wenige grobkörnige autoritative Metadaten, die beim Upload von den Anbietern mitgeliefert werden. Die autoritativen Metadaten werden durch eine feingranulare automatische Erschließung auf der Segmentebene des Videos ergänzt. Dies ermöglicht eine zielgenaue, segmentbasierte Suche im Videobestand. Die deutschsprachigen Videos des TIB/AV-Portals werden automatisiert mit GND-Sachbegriffen semantisch verschlagwortet. Mehrdeutigkeiten bei der Zuordnung von GND-Entitäten werden, soweit möglich, durch einen Disambiguierungsalgorithmus aufgelöst, der dazu

Mapping der GND-Sachbegriffe auf andere Normdaten

Erweiterung der Suchergebnisse durch Synonyme und Übersetzungen

Präzisierung der Suchergebnisse durch Facetten

Kontextinformationen aus GND, DBpedia und Wikipedia nutzt. Die semantische Verschlagwortung der Videos mit Entitäten ermöglicht eine Verbesserung der traditionellen Schlüsselwort-basierten Suche sowohl durch eine Erweiterung der Suchergebnisse (Synonyme, englische Übersetzungen) als auch durch eine Präzisierung der Suchergebnisse (Facettensuche). Die englischsprachigen Videos des TIB/AV-Portals werden mithilfe von englischsprachigen Bezeichnern verschlagwortet, die über ein Mapping der GND-Entitäten auf andere Normdaten (DBpedia etc.) gewonnen wurden.

Die Beschreibungen der Metadaten (autoritative Metadaten, Transkripte, Entitäten) der Videos sind in Form des Resource Description Framework (RDF) in einem Triplestore abgelegt. Diese Beschreibungen sind momentan nur intern verfügbar. Für die Zukunft ist geplant, diese Daten als Linked Open Data zu publizieren, um deren Sichtbarkeit im Web of Data zu erhöhen und eine Nachnutzbarkeit in anderen Projekten zu ermöglichen. Außerdem soll die TIB/AV-Portal-Wissensbasis regelmäßig aktualisiert werden, indem die frei veröffentlichten Updates der GND und DBpedia, die die verwendeten Klassen aus Naturwissenschaft und Technik betreffen, in der Wissensbasis aufgenommen werden. Damit wird die Aktualität der Begriffe gesichert und der Disambiguierungsalgorithmus tendenziell verbessert.

LITERATUR

Gemeinsame Normdatei (GND). Internetseite der Deutschen Nationalbibliothek. www.dnb.de/gnd (abgerufen: 10.04.14).

Gruber, Thomas R. Toward Principles for the Design of Ontologies used for Knowledge Sharing. In: International Journal of Human-Computer Studies, 43/5–6 (1995), S. 907–928.

Hentschel, Christian; Blümel, Ina; Sack, Harald. Automatic Annotation of Scientific Video Material based on Visual Concept Detection. In: Proc. 13th International Conference on Knowledge Management and Knowledge Technologies, No. 16 (2013).

Plank, Margret; Neumann, Janna. TIB's Portal for Audiovisual Media. New Ways of Indexing and Retrieval. In: IFLA Journal, 40 (2014), S. 17–23.

Prechtl, Peter; Burkard, Franz-Peter [Hrsg.]. Metzler Philosophie Lexikon. Begriffe und Definitionen. Stuttgart, Weimar: Metzler, 1999.

Sack, Harald. Semantische Suche. Theorie und Praxis am Beispiel der Videosuchmaschine yovisto.com. In: Hengartner, Urs; Meier, Andreas [Hrsg.]. Web 3.0 & Semantic Web. HMD – Praxis der Wirtschaftsinformatik, Nr. 271. Heidelberg: dpunkt Verlag, 2010, S. 13–25.

Sack, Harald; Plank, Margret. AV-Portal. The German National Library of Science and Technology's Semantic Video Portal. In: ERCIM News, No. 96 (2014).

Steinmetz, Nadine; Sack, Harald. Cross-Lingual Semantic Mapping of Authority Files. Präsentation auf der Konferenz »Semantic Web in Libraries«. SWIB 2013, Hamburg. Video zur Präsentation: www.scivee.tv/node/61558

1 www.tib-hannover.de/de/dienstleistungen/kompetenzzentrum-fuer-nicht-textuelle-materialien-knm

2 www.tib-hannover.de/de/die-tib

3 www.hpi.uni-potsdam.de

4 www.dnb.de/lids

5 <http://de.dbpedia.org>

6 <http://id.loc.gov/authorities/subjects.html>

7 www.dnb.de/DE/Wir/Projekte/Laufend/macros.html

8 www.wti-frankfurt.de/index.php/thesaurus

DIE VERFASSER

Dr. Sven Strobel, Wissenschaftlicher Mitarbeiter, Kompetenzzentrum für nicht-textuelle Materialien (KNM), Technische Informationsbibliothek (TIB), Welfengarten 1B, 30167 Hannover, Tel.: 0511 – 762-14229, E-Mail: Sven.Strobel@tib.uni-hannover.de

Margret Plank, M. A., Leitung, Kompetenzzentrum für nicht-textuelle Materialien (KNM), Technische Informationsbibliothek (TIB), Welfengarten 1B, 30167 Hannover, Tel.: 0511 – 762-4884, E-Mail: Margret.Plank@tib.uni-hannover.de