Repository of Hochschule Hannover

HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS

Entwicklung eines Verfahrens zur automatischen Sammlung, Erschließung und Bereitstellung multimedialer Open-Access-Objekte mittels der Infrastruktur von Wikimedia Commons und Wikidata

Ina Blümel, Simone Cartellieri, Lambert Heller, Christian Wartena

Suggested citation:

Blümel, Ina, Simone Cartellieri, Lambert Heller, and Christian Wartena. 2014. "Entwicklung eines Verfahrens zur automatischen Sammlung, Erschließung und Bereitstellung multimedialer Open-Access-Objekte mittels der Infrastruktur von Wikimedia Commons und Wikidata." Hannover: Hochschule Hannover. https://doi.org/10.25968/opus-675.

Abstract

Idee zum Projektantrag der TIB Hannover und Hochschule Hannover in der DFG-Ausschreibung: „Open-Access-Transformation" vom 2.6.2014 im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme" (LIS), eingereicht am 30.10.2014

Terms of use

CC BY 4.0

This document is made available under these conditions:
Creative Commons - CC BY - Namensnennung 4.0 International
For more information see:
https://creativecommons.org/licenses/by/4.0/deed.de

Entwicklung eines Verfahrens zur automatischen Sammlung, Erschließung und Bereitstellung multimedialer Open-Access-Objekte mittels der Infrastruktur von Wikimedia Commons und Wikidata

Idee zum Projektantrag der TIB Hannover und Hochschule Hannover in der DFG-Ausschreibung: „Open-Access-Transformation“ vom 2.6.2014 im Bereich „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (LIS), eingereicht am 30.10.2014

1 Ausgangslage und Vorarbeiten

Multimediale Objekte in Open-Access-Publikationen und ihr Potenzial für eine Nachnutzung

Forschungsergebnisse werden auch heute noch überwiegend als Textpublikationen veröffentlicht, die in zunehmendem Maße auch von den zugrundeliegenden Forschungsrohdaten begleitet werden. Ein konstitutiver Bestandteil von Publikationen sind zudem *multimediale Objekte*, deren Funktion in erster Linie darin besteht, die eigentlichen Forschungsergebnisse leichter rezipierbar zu machen und das Verstehen zu fördern. Zu diesem Zweck erstellen Wissenschaftler Abbildungen (z.B. Fotos, aus bildgebenden Verfahren resultierende Abbilder, Skizzen, Diagramme, Visualisierungen von Forschungsdaten), die sie ihren Publikationen beifügen. Daneben können auch Objekte wie Video- und Tondateien in Publikationen eingebunden sein.

Wenn Forschende, Lehrende oder Lernende multimediale Objekte, die Bestandteil von im Closed Access veröffentlichten Textpublikationen sind, nachnutzen möchten, stellt die Lizensierung eine kaum überwindbare Barriere dar. Ganz anders stellt sich die Situation im Bereich Open Access dar: Sobald Textpublikationen im Open Access veröffentlicht werden, sind auch die zur jeweiligen Publikation gehörenden multimedialen Objekte grundsätzlich im freien Internet auffindbar. Aus diesem Grund wird im vorliegenden Antrag von *multimedialen Open-Access-Objekten* (im Folgenden auch MOA) gesprochen. MOA können - losgelöst von ihrer “ursprünglichen” Publikation - in anderen Forschungsarbeiten erneut verwendet werden, oftmals in vollkommen anderen Kontexten. Sie können daneben auch einen hervorragenden Beitrag zur Vermittlung von Forschungsergebnissen leisten, nicht zuletzt als "wissenschaftliches Rohmaterial" für die Erstellung von "Open Educational Resources" (OER). Einen weiteren Einsatzbereich stellt die Verwendung von MOA in Wikipedia-Artikeln dar. An kaum einer Stelle wird der Mehrnutzen von Open Access auch für interessierte Laien so deutlich wie an der Frage der Nachnutzung dieser Objekte.[1]

Herausforderungen bei der Nachnutzung von MOA

Das herausragende Potenzial von MOA für eine Nachnutzung in den angeführten Bereichen wird allerdings noch nicht ausgeschöpft. Denn obwohl der Bedarf nach wiederverwendbaren Abbildungen aus wissenschaftlichen Publikationen sowohl bei Forschenden als auch bei Lehrenden und Lernenden groß ist, gibt es bisher kein multidisziplinäres Repository, das ein gezieltes, umfassendes Retrieval von MOA (für wissenschaftliche Zwecke) erlaubt.

Während die wissenschaftlichen Publikationen durch standardisierte fachliche und fächerübergreifende Indizes über Suchmaschinen und in bibliothekarischen Rechercheportalen gefunden und nachgenutzt werden können, fehlt ein vergleichbarer gezielter Zugriff auf die darin enthaltenen oder sie begleitenden multimedialen Open-Access-Objekte (MOA). Über nach Lizenzen gefilterte Suchanfragen in generellen Websuchmaschinen oder Hosting-Plattformen für

[1] Ergänzend ist anzumerken, dass MOA durchaus auch dann, wenn die jeweiligen Publikationen selbst nicht in Open-Access-Journals (Golden Road oder Hybrid) erschienen sind, unter einen freien Lizenz verfügbar sein können, etwa auf institutionellen oder fachlichen Open-Access-Repositories, wenn die zwischen Verlagen und Autoren vereinbarte Embargo-Frist verstrichen ist. Ergänzende Materialien wie Foliensätze und Vorlesungsmitschnitte sind häufig ebenfalls unter freien Lizenzen auf den Webseiten von Bildungseinrichtungen, und z.T. auch auf Repositories, verfügbar.

Bild- oder Videomaterialien erhält man überwiegend Suchergebnisse ohne direkten Wissenschafts- oder Forschungsbezug.[2]

Erschwerend kommt hinzu, dass der bei weitem überwiegende Teil der Open-Access-Fachliteratur auf den Plattformen zahlreicher Publisher und Repositories verstreut liegt. Zudem ist diese Literatur oft ausschließlich in Gestalt von PDF-Dateien verfügbar, was insbesondere eine Herausforderung für die Gewinnung von Informationen über einzelne multimediale Objekte innerhalb dieser Dateien bedeutet, wie etwa Bildunterschriften und Referenzen zum Bild im Text eines Artikels.

Es besteht daher ein großer Bedarf, standardisierte fachliche und fächerübergreifende Indizes auch für die multimedialen Open-Access-Objekte – insbesondere für Abbildungen als die aktuell am häufigsten in wissenschaftliche Publikationen integrierten MOA – zu generieren und bereitzustellen, um den bereits vorhandenen Bestand besser zu verbreiten und nachnutzbar zu machen. Um diese Aufgabe zu adressieren, müssen innovative Verfahren zur Sammlung, Erschließung und Indexerstellung der MOA entwickelt werden. Auf Herausforderungen der Erschließung von Abbildungen aus Textpublikationen soll im folgenden Abschnitt genauer eingegangen werden.

Extraktion von beschreibenden Metadaten aus dem Multimedia-Kontext: Das Problem der "Semantic Gap"

Ein zentrales Problem bei der textbasierten Suche nach Abbildungen ist die sogenannte semantic gap bzw. die semantische Lücke: die Merkmale, die aus einem Bild extrahiert werden können, lassen sich nicht oder nur schwer mit den Konzepten, mit denen Menschen ein Bild beschreiben würden, in Verbindung bringen (Liu et al. 2006). Der Versuch, diese Lücke zu überbrücken, ist zentral in der textbasierten Bildsuche (Liu et al. 2006, Lew et al. 2006). Hierzu werden hauptsächlich Verfahren des maschinellen Lernens angewandt. Als Trainingsmaterial werden häufig vorhandene Bildunterschriften genutzt (Feng & Lapata 2010). Das Extrahieren von beschreibenden Schlagwörtern aus Bildunterschriften und weiteren Textabschnitten aus der Umgebung des Bildes wird in Leong et al. (2010) beschrieben. Das Verfahren, das sich für die Schlagwortextrahierung durchgesetzt hat, nutzt verschiedene Maße, die etwas darüber sagen, ob ein Wort als Schlagwort geeignet ist. Die Maße beziehen sich dabei sowohl auf die generelle Eignung eines Wortes als Schlagwort als auch auf die Repräsentativität für den ganzen Text. Die optimale Art, die Maße zu kombinieren, wird durch überwachtes Lernen ermittelt (Frank et al. 1999, Turney 2000).

Mihalcea et al. (2007) und Medelyan et al. (2008) nutzen Wikipedia als ein kontrolliertes Vokabular für die Verschlagwortung. Einerseits werden mögliche Schlagworte durch das Vokabular eingeschränkt, andererseits wird auch die Thesaurusstruktur zu Bestimmung der Relevanz herangezogen. Die Benutzung eines Vokabulars für die Verschlagwortung von Bildern wurde beispielswiese von Jin et al. (2005) und Srikanth et al. (2005) vorgeschlagen. Da viele Konzepte nicht wörtlich im Text vorkommen und Bildunterschriften meistens nur kurze Texte sind, ist es bei der Verschlagwortung von Bildern erforderlich, festzustellen, ob gefundene Terme Synonyme von Wikidata-Konzepten sind. Für große Mengen potenzieller Terme wurden gute Ergebnisse mit verschiedenen Varianten der distributionellen Semantik erzielt (siehe z.B. Turney und Pantel 2010, Bullinaria und Levy 2012, Saif und Hirst 2012, Kiela und Clark 2014). Wenn es darum geht, zu entscheiden, ob zwei Wörter synonym sind, wurden auch hier die besten Ergebnisse durch die Kombination verschiedener statistischer Ähnlichkeitsmaße mittels überwachtem Lernen erzielt (Bär et al. 2012, Wartena 2013b).

Aktuelle Initiativen und Projekte

Im Folgenden wird beschrieben, welche Initiativen und Projekte die oben angeführten Herausforderungen derzeit angehen und welche Desiderata bestehen bleiben:

Wikimedia Commons als Infrastruktur zur Archivierung und Recherche multimedialer Objekte
Ein Modell für die Sammlung, Erschließung und Verfügbarmachung von MOA ist Wikimedia Commons, ein von der Wikimedia Foundation[3] betriebenes Schwesterprojekt zur Wikipedia, worin

[2] https://www.flickr.com/search/advanced/, http://images.google.de/advanced_image_search

frei lizensierte Multimedia-Dateien samt ihrer Metadaten gespeichert werden und von dort aus allen Sprachversionen der Wikipedia sowie allen weiteren Wikimedia-Projekten (z.B. Wikinews, Wikibooks, Wikiversity) sowie letztlich allen Web-Nutzern zur Verfügung stehen. Wikimedia Commons ist mit mehr als 23 Millionen Mediendateien[4] eine der größten freien Mediensammlungen der Welt. Das Gesamtvolumen der Sammlung beträgt über 45 Terabyte, davon 41 Terabyte Bilder.[5] Metadaten werden auf Wikimedia Commons momentan als semi-strukturierter Text (Hypertext bzw. Wikitext mit Templates) gespeichert. In Zukunft sollen alle Metadaten aber direkt als maschinenlesbare Datenstrukturen gespeichert und gewartet werden. Ein entsprechendes Projekt[6], betrieben von Wikimedia Deutschland[7] gemeinsam mit der amerikanischen Wikimedia Foundation, ist gerade angelaufen.

Wikidata und Wikibase
Wikidata[8] ist ein relativ neues Projekt der Wikimedia Foundation, das der Speicherung von maschinenlesbaren, sprachneutralen Informationen dient, die dann in Wikipedia und anderen Wikimedia-Projekten (und letztlich auch von Dritten) weiter verwendet werden können. Wikidata[9] basiert auf der Software Wikibase, die in wesentlichen Teilen von Wikimedia Deutschland entwickelt und betreut wird. Wikibase wird auch die Grundlage für die Verwaltung von Metadaten auf Wikimedia Commons werden.

WikiProjekt "Open Access", Teil I: Fokussierter Crawler für die Sammlung multimedialer Open-Access-Objekte (MOA) aus PubMed Central
Im Kontext des vom WikiProjekt "Open Access"[10] betriebenen Open Access Media Importer Bots[11] wurde gezeigt, wie multimediale Open-Access-Objekte automatisiert gesammelt und von der Fachöffentlichkeit der Wikipedia-Autoren zur Ergänzung thematisch einschlägiger Lexikonartikel nachgenutzt werden können. Der Bot (eine Software zur Erledigung repetitiver Arbeiten) durchsucht das biomedizinische Literatur-Repositorium PubMed Central[12] nach frei lizensierten Artikeln und diese wiederum auf Audio- und Video-Dateien, welche es dann in ein offenes Format konvertiert und auf Wikimedia Commons hochlädt. Das Projekt, dessen Anfänge 2011 von Wikimedia Deutschland gefördert wurden[13], betritt Neuland hinsichtlich der Nachnutzung von Open-Access-Materialien und wurde dafür 2013 im Rahmen des Accelerating Science Award Program ausgezeichnet[14]. Bislang wurden damit über 17.000 Audio- und Video-Dateien auf Wikimedia Commons übertragen, eine Erweiterung auf Bilddateien ist derzeit in der Testphase.

WikiProjekt "Open Access", Teil II: Open Access Media Importer zur Weiterverarbeitung wissenschaftlicher Artikel im JATS-XML-Format
Der de-facto-Standard für den Austausch von Artikel-Volltexten in maschinenlesbarer Form ist JATS[15], ein NISO-Standard[16] für XML, der von PubMed Central verwendet wird und auch außerhalb der Biowissenschaften zunehmend Anwendung findet (z.B. SciELO, Optical Society of America, Copernicus). Der Open Access Media Importer ist der erste Versuch, dieses Markup großmaßstabig zu verwenden[17] (z.B. zur Kategorisierung der auf Wikimedia Commons hochgeladenen Multimedia-Dateien oder zur Bestimmung ihrer Lizensierung), und sein Betrieb hat bereits einige Probleme hinsichtlich der Nachnutzbarkeit des gegenwärtig produzierten JATS

---

[3] http://wikimediafoundation.org
[4] http://commons.wikimedia.org/wiki/Special:Statistics
[5] http://commons.wikimedia.org/wiki/Commons:MIME_type_statistics
[6] https://commons.wikimedia.org/wiki/Commons:Structured_data
[7] https://www.wikimedia.de
[8] https://www.wikidata.org
[9] http://wikiba.se
[10] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Open_Access
[11] https://commons.wikimedia.org/wiki/User:Open_Access_Media_Importer_Bot
[12] http://www.ncbi.nlm.nih.gov/pmc/
[13] http://blog.wikimedia.de/2011/12/15/wissenswert-2011-wir-gratulieren-den-fuenf-gewinnern/
[14] http://blogs.plos.org/plos/2013/10/announcing-the-recipients-for-the-accelerating-science-award-program/
[15] http://jats.nlm.nih.gov/
[16] http://www.niso.org/apps/group_public/document.php?document_id=10591
[17] https://en.wikipedia.org/wiki/User:Daniel_Mietchen/Talks/JATS-Con_2014/Abstract

aufgedeckt, worauf einige Open-Access-Publisher mit der Bildung einer Arbeitsgruppe zur Verbesserung der Nachnutzbarkeit reagiert haben, welche 2014 ihre Arbeit aufgenommen hat.[18]

Festhalten lässt sich, dass die aufgeführten Projekte "Open Access" (I+II) zwar ein Crawling in Open-Access-Publikationen aus bislang einer Quelle sowie die Extraktion der in ihnen enthaltenen multimedialen Objekte für eine exemplarische Anwendungsdomäne umgesetzt haben, dass aber die für Suche und Bereitstellung dieser Objekte elementare Indexierung / inhaltliche Erschließung bislang nicht geleistet wurde. Des Weiteren findet in diesen Projekten keine Verwendung der bereits gewonnenen Texte zur automatisierten Metadatengewinnung statt. Im Rahmen des in Abschnitt 2 beschriebenen Vorhabens soll unter anderem diese Herausforderung bewältigt werden.

Eigene Vorarbeiten TIB

Die Technische Informationsbibliothek (TIB) in Hannover ist als Deutsche Zentrale Fachbibliothek für Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik ein bedeutender Informationsanbieter in Deutschland. In dieser Funktion treibt die TIB gemeinsam mit ihren Forschungspartnern erfolgreich die kontinuierliche Neu- und Weiterentwicklung von Informationssystemen für die Wissenschaft und zugrundeliegenden Infrastrukturen voran. In GetInfo[19], dem Fachportal für Technik und Naturwissenschaften, werden zusätzlich zu Textpublikationen auch Forschungsdaten, 3D-Modelle und AV-Medien verzeichnet und bereitgestellt. Diese erweiterte Dienstleistung ermöglicht in einem ersten Schritt den auf Metadaten basierenden Nachweis von archivierten und referenzierten Forschungsdaten in einem Fachportal gemeinsam mit Publikationen. Für die Identifizierung, eindeutige Referenzierung und Sicherung der Zitierfähigkeit von Forschungsdaten hat die TIB seit 2005 als weltweit erste nicht-kommerzielle DOI-Registrierungsagentur (DOI: *Digital Object Identifier*) sowohl die nötige Infrastruktur als auch inhaltliche Kompetenzen aufgebaut, die 2009 zur Gründung des internationalen Vereins DataCite, führte. Die DataCite-Geschäftsstelle ist an der TIB ansässig.

Die TIB ist Gründungsmitglied des strategischen Forschungsverbundes Science 2.0 der Leibniz-Gemeinschaft, dessen Ziele der Aufbau von Expertise sowie die Entwicklung neuartiger Werkzeuge im Bereich der digitalen Wissenschaft sind.

Die TIB hat bereits in verschiedenen Projekten an der grundlegenden Verbesserungen von Zugangs- und Nutzungsbedingungen für nicht-textuelle Materialien gearbeitet, vgl. (Brase et Blümel, 2010), vor allem für wissenschaftliche AV-Materialien und 3D-Modelle, die automatisch erschlossen und bereitgestellt werden sowie text- und inhaltsbasiert durchsucht werden können. Beispielhaft genannt seien hier das Projekt PROBADO (Blümel 2010, Berndt et al. 2010, Wessel, Blümel, Klein 2009) und AV-Portal (Hentschel, Blümel, Sack 2013). Das Kompetenzzentrum für nicht-textuelle Materialien (KNM) an der TIB entwickelt kontinuierlich weitere vielfältige Werkzeuge, Konzepte und Strukturen für die Sammlung, Erschließung und Bereitstellung sowie die Standardisierung und Archivierung nicht-textueller Materialien. Diese Aktivitäten werden seit 2013 ergänzt durch das Open Science Lab (OSL) der TIB, das sich der Entwicklung von Tools und Methoden zum vernetzten und kollaborativen wissenschaftlichen Arbeiten widmet.[20]

Eigene Vorarbeiten HsH

Forschungsschwerpunkte der Professur Sprach- und Wissensverarbeitung an der Hochschule Hannover sind die Verwendung von kontrolliertem Vokabularen und Thesauri für die automatische Verschlagwortung sowie die Erfassung der Bedeutung von (Thesaurus)-Termen, Tags und Wörtern mit Hilfe der distributionellen Semantik. In Wartena & Brussee (2008) wurde die distributionelle Semantik angewandt, um ein Mapping zwischen kollaborativen Tags aus del.icio.us und Wikipedia-Kategorien herzustellen. Distributionelle Semantik von Tags wurde auch im Projekt “MyMedia” im im Rahmen des FP7 untersucht (Wartena, Brussee & Wibbels 2009, Wartena & Wibbels 2011). In Wartena, Brussee & Slakhorst 2010 und Wartena, Slakhorst & Wibbels 2010 wurde die distributionelle Semantik für eine automatische Verschlagwortung wissenschaftlicher Aufsätze bzw. die Verschlagwortung von Fernsehsendungen verwendet. Da es für viele praktische

[18] http://jats4r.github.io/
[19] www.getinfo.de
[20] http://blogs.tib.eu/wp/opensciencelab/

Anwendungen wichtig ist, gerade die Bedeutung selten vorkommender Wörter oder Terme automatisch zu erfassen, wurde zuletzt untersucht, wie distributionelle Ähnlichkeit für niedrig frequente Wörter und Wortpaare mit sehr unterschiedlichen Frequenzen verbessert werden kann (Wartena 2013a, Wartena 2013b, Wartena 2014).

In weiteren Vorarbeiten wurde Erfahrung mit automatischer Verschlagwortung mit festen Vokabularen gesammelt. In Wartena, Brussee, Gazendam & Huijssen (2007) wurde ein Open-Source-Tool zur Erkennung von Thesaurustermen in Texten entwickelt. In Gazendam, Wartena & Brussee (2010) wird darüber hinaus die Struktur des Thesaurus für die Bestimmung der Relevanz der Terme für einen Text benutzt. In Wartena & Sommer (2012) und Wartena & Garcia-Alsina (2013) werden extrahierte Thesaurusterme für die weitere Klassifikation der Dokumente herangezogen.

1.1 Projektbezogene Publikationen

1.1.1 Veröffentlichte Arbeiten aus Publikationsorganen mit wissenschaftlicher Qualitätssicherung, Buchveröffentlichungen sowie bereits zur Veröffentlichung angenommene, aber noch nicht veröffentlichte Arbeiten

- Berndt, R.; Blümel, I.; Clausen, M.; Damm, D.; Diet, J.; Fellner, D.; Fremerey, C.; Klein, R.; Krahl, F.; Scherer, M. (2010). The PROBADO project-approach and lessons learned in building a digital library system for heterogeneous non-textual documents. *Research and Advanced Technology for Digital Libraries*. S. 376-383, Springer.
- Brase, J.; Blümel, I. (2010). Information supply beyond text: non-textual information at the German National Library of Science and Technology (TIB)–challenges and planning. *Interlending & Document Supply* 38 (2). S. 108-117, Emerald Group Publishing Limited.
- Hentschel, C.; Blümel, I.; Sack, H. (2013). Automatic Annotation of Scientific Video Material based on Visual Concept Detection. *Proceedings of 13th International i-know Conference*, ACM Publishing.
- Wartena, C. (2014). On the effect of word frequency on distributional similarity. *Proceedings of KONVENS 2014*. S. 1-10.
- Wartena, C. (2012). Estimating Semantic Similarity of Words and Short Phrases with Frequency Normalized Distance Measures. *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Atlanta, Georgia, USA*.
- Wartena, C. und Sommer, M. (2012). Automatic classification of scientific records using the German Subject Heading Authority File (SWD). *Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA 2012)*. S. 37-48.
- Larson, M.; Soleymani, M.; Serdyukov, P.; Rudinac, S.; Wartena, C; Murdock, V.; Friedland, G.; Ordelman, R.; Jones, G. (2011). Automatic Tagging and Geotagging in Video Collections and Communities. *ACM International Conference on Multimedia Retrieval (ICMR 2011)*.
- Wartena, C. and Brussee, R. (2008). Instance-Based Mapping between Thesauri and Folksonomies. *International Semantic Web Conference, Vol. 5318 of LNCS*. S. 356-370.
- Wartena, C.; Brussee, R.; Gazendam, L.; Huijsen, W.-O. (2007). Apolda: A Practical Tool for Semantic Annotation. DEXA Workshops: IEEE Computer Society. S. 288-292.
- Wessel, R.; Blümel, I.; Klein, R. (2009). A 3D Shape Benchmark for Retrieval and Automatic Classification of Architectural Data. *Eurographics 2009 Workshop on 3D Object Retrieval*. S. 53-56.

2 Ziele

2.1 Voraussichtliche Gesamtdauer des Projekts

36 Monate

2.2 Ziele

Ziel des beantragten Projektes ist es, basierend auf der Infrastruktur von Wikimedia Commons und Wikidata, ein Verfahren zur automatischen Sammlung (Harvesting), Erschließung und Bereitstellung multimedialer Objekte aus qualitätsgesicherten Open-Access-Journals zu entwickeln und über einen Suchservice verfügbar zu machen. Das Projekt verfolgt damit das übergeordnete Ziel, ein Informationsangebot aufzubauen, welches den Zugriff auf und die Nachnutzung von frei zugänglichen multimedialen Objekten zur Visualisierung von Forschungsergebnissen verbessert und nachhaltig absichert.

Bilddateien aus multidisziplinären, heterogenen Quellen für die gezielte Suche und Nachnutzung erschließen

Unstrukturierte Textdokumente aus heterogenen Quellen stellen eine Herausforderung für die Extraktion von Informationen über einzelne Abbildungen innerhalb dieser Dateien dar. Die bereits aus PubMed Central gewonnenen und - im Rahmen des hier vorgeschlagenen Projekts - auch aus anderen Quellen und in anderen Formaten zu gewinnenden Texte - sollen bereinigt und zur Beschreibung der Objekte verwendet werden. Verfahren zur automatischen Erfassung der Bedeutung von Deskriptoren und Thesaurustermen, die eine Zuordnung von Deskriptoren zu multimedialen Objekten ermöglichen, müssen hierzu auf den jeweiligen Anwendungsfall adaptiert und Verfahren der distributionellen Semantik auf eine derartige Anwendung ausgeweitet werden.

Mit Wikidata-basierter Erschließung sprachübergreifendes Retrieval sowie die Vernetzung mit anderen Wissensressourcen und Normdaten ermöglichen
Es gibt mehrere Gründe, warum bei der oben angeführten automatischen Erschließung Konzepte aus Wikidata als kontrolliertes Vokabular verwendet werden sollten. Zum einen vereinfacht sich die Nachnutzung der gewonnenen multimedialen Objekte im engeren Kontext der Wikipedia und ihrer Schwesterprojekte. Zum anderen gilt auch außerhalb dieses Kontexts, dass diese Objekte mittels einer Suche, die Suchbegriffe auf einen großen und vielsprachig vernetzten Thesaurus wie das Begriffsnetz der Wikipedia abbildet, international leichter gefunden werden können (Heller 2011a und 2011b). Sie können in "fremden" sprachlichen Kontexten oft verstanden und nachgenutzt werden, da gerade Bilder oft unabhängig von Sprache "funktionieren". Da in zunehmendem Maße auch Normdaten Dritter auf der Plattform Wikidata präsent sind (Martinelli 2014), eröffnet eine auf Wikidata basierende Erschließung zudem ein Potenzial für die Nachnutzung in Suchdiensten und "datengetriebenen" Anwendungen.

Um eine solche Nachnutzung und Erweiterung der hier angestrebten Projektziele durch Dritte zu erleichtern, werden neben der Erschließung mit Wikidata sowie der ausschließlichen Verwendung bestehender Open-Source-Software eigene Skripte und Verarbeitungsschritte im Sinne eines “Selbstbau-Sets” dokumentiert und exemplarisch für die Schaffung eines neuen optionalen Suchziels von TIB GetInfo verwendet.

Wirkungsmessung auf Objektebene vielfältig nutzen
Wikimedia Commons sammelt kontinuierlich Logdaten, aus denen hervorgeht, wann und von wo die Beschreibung eines multimedialen Objekts aufgerufen worden ist. Im Rahmen des Vorhabens soll die vielfältige Nutzbarkeit solcher Daten hinsichtlich der auf Wikimedia Commons gespeicherten wissenschaftlichen Objekte demonstriert werden, indem diese Daten auf Objektebene in aggregierter Form via Wikidata zugänglich gemacht werden, etwa für wissenschaftliche Altmetrics-Dienste. Zudem sollen diese Daten eingesetzt werden, um das Retrieval der Objekte zu verbessern - direkt durch Verwendung als Faktor beim Ranking der Suchergebnisse, indirekt durch das Generieren von Revisionslisten zur intellektuellen Nachprüfung der automatischen Klassifikationsergebnisse. Diese Maßnahme soll dabei helfen, das Problem des erwarteten guten Recalls bei (zu) geringer Precision der automatischen Klassifikationsergebnisse zu kompensieren.

Wikimedia-Plattformen als nachhaltige Infrastruktur für Aufbereitung und Retrieval freier Informationsressourcen einsetzen
Die Wikipedia ist - untrennbar verbunden mit Schwesterprojekten wie Wikimedia Commons und Wikidata - seit 2009 kontinuierlich auf dem sechsten Platz der populärsten Websites weltweit[21]. Um das durch Open Access vervielfachte Potenzial der Nachnutzung wissenschaftlicher multimedialer Objekte optimal zur Entfaltung zu bringen, ist die Konsumentennähe der Wikipedia-Projekte ein vielversprechender Hebel. Die Wikipedia dient, wie Untersuchungen zeigen, inzwischen selbst Wissenschaftlern (Pscheida et al. 2013) vielfach als Anlaufstelle zur fachlichen Orientierung. Zudem passen die Wikipedia-Projekte zum Ziel einer umfassenden Sammlung von Objekten unter Open-Access-Konditionen, da alle diese Projekte strikt Lizenzbedingungen anwenden, die als vereinbar mit dem Open Access im Sinne der Berliner Erklärung gelten.

[21] https://meta.wikimedia.org/wiki/MP

2.3 Arbeitsprogramm und Umsetzung

Das Projekt besteht aus drei aufeinanderfolgenden Phasen - Vorbereitung, Aufbau der Pipeline für Crawling und Erschließung, Finalisierung der Produktionsumgebung - sowie vier Arbeitspaketen (AP), die im folgenden erläutert werden. In jeder Phase wird der Projektfortschritt durch Meilensteine (M1–M6) kontrolliert (siehe auch Abb. 1). Um innerhalb von drei Jahren ein relevantes, retrieval-fähiges Objektkorpus zu generieren, konzentriert sich dieses Projekt auf die Erschließung von **Abbildungen** mittels Analyse der damit verbundenen textuellen Elemente, kombiniert mit einer Bildklassifikation (vgl. AP3), aus englisch- und deutschsprachigen Publikationen.

Vorbereitung

In dieser Phase werden rechtliche Fragestellungen geklärt (AP1), verfügbare Artikel identifiziert und bereits vorhandene Daten zur Vorbereitung der Crawlingphase analysiert (AP2) sowie Pre-Processing-Schritte zur Informationsexktraktion implementiert (AP3).

M1 (Monat 3): Bestandsaufnahme: Datensammlungen, Lizenz, etc. ausgewählt. Webseite aufgebaut, Community-Prozess angestoßen.

Aufbau der Pipeline

Diese Phase umfasst die Auswahl, (Weiter-)Entwicklung, Implementierung sowie Integration geeigneter Verfahren für das Crawling und Informationsextraktion sowie die kontinuierliche Modellierung der Medienbeschreibung in Wikidata (AP 2 und 3). Ebenfalls in diese Phase gehören der Aufbau von Index und Suchservice (AP 4) sowie die Einbeziehung von Nutzungsdaten.

M2 (Monat 9): Rohdaten des ersten Crawls sind verfügbar, Bilder und Texte sind extrahiert, erste Bilder aufbereitet und in Wikimedia Commons abgelegt. Die Extraktion aus Textmaterial liefert die ersten Konzepte, Wikidata ist auf vorhandene Eigenschaften zur inhaltlichen Beschreibung von wissenschaftlichen MOA analysiert. Erste Version der Erweiterung von Wikidata-Konzepten für wissenschaftliche Bildmaterialien verfügbar.

M3 (Monat 15): Anbindung von Extraktion und Mapping auf Wikidata-Konzepte implementiert. Index und Bereitstellung.

M4 (Monat 21): 2. Crawl, weitere Rohdaten verfügbar, erste Version Suchservice und “Construction Set”.

M5 (Monat 30): Veröffentlichung der Umgebung und aller Werkzeuge, Metadatenmodellierung abgeschlossen, Nutzungsmetriken im Suchservice eingebunden.

Finalisierung der Umgebung

In dieser Phase werden weitere Schnittstellen implementiert und die Pipeline vervollständigt. Durch das gewonnene Feedback wird die Pipeline weiter verbessert und für den Produktiveinsatz freigegeben.

M6 (Monat 36): Endabnahme und Abschlussworkshop.

Phase	Vor-bereitung	Aufbau der Pipeline (Sammlung, Erschließung und Bereitstellung)										Finalisierung der Umgebung	
Meilenstein	M1	M2			M3		M4		M5			M6	
Projektjahr	1				2				3				
Quartal	1	2	3	4	1	2	3	4	1	2	3	4	
AP 1: Projektmanagement													
1.1 Koordinierung													
1.2 Dissemination													
1.3 Klärung rechtlicher Fragen													
AP 2: Crawler													
2.1 Bestandsaufnahme													
2.2 Fokussiertes Crawling													
2.3 Aufbereitung der Medienobjekte													
AP 3: Erschließung													
3.1. Modellierung													
3.2 Metadatengewinnung													
AP 4: Nachnutzbare Suchlösungen													
4.1 Metadatenaggregation und -bereitstellung													
4.2 Suchservice													
4.3 “Construction Set” und exemplarische Nachnutzungen													
4.4 Verbesserung des Suchservice durch Nutzungsmetriken													

Abbildung 1: Zeitplanung der Arbeitspakete

2.3.1 **AP 1 – Projektmanagement**

Aufgaben/Beschreibung

Das Projektmanagement gewährleistet die Zusammenarbeit und den Informationsfluss im Projektteam. Dazu gehören die Organisation regelmäßiger Telefonkonferenzen sowie die Durchführung von Projekttreffen. Um eine intensive und effiziente Zusammenarbeit mit Wikimedia Deutschland e.V. zu gewährleisten, ist geplant, regelmäßige gemeinsame Workshops zu organisieren, zu denen auch weitere externe Experten nach Bedarf eingeladen werden. Pro Projektjahr sind 4 fachlich zugeschnittene Workshops geplant. Darüber hinaus stellt das Projektmanagement in enger Abstimmung zwischen den Projektpartnern eine angemessene Dissemination der Projektergebnisse in Form von Publikationen, Konferenzbeiträgen und Workshops sicher.

2.3.2 **AP 2 - Crawler**

Ziele

- Bestandsaufnahme verfügbarer Open-Access-Artikel
- Verfahren zum automatischen Harvesten der Artikel
- Verfahren zum automatischen Extrahieren, Bereinigen und Modellieren der textuellen Inhalte und multimedialen Objekte aus den Artikeln

Aufgaben/Beschreibung

Im Rahmen dieses Arbeitspaketes wird zunächst eine Bestandsaufnahme über die Anzahl von Artikeln aus Open-Access-Journals durchgeführt, wobei zu berücksichtigen ist, ob ein automatisiertes Retrieval möglich ist und zudem Lizenzen verwendet werden, die Content Mining (vgl. AP 3) erlauben. Hauptaufgabe des Arbeitspakets ist die Entwicklung eines Verfahrens zum automatisierten Crawling geeigneter Open-Access-Artikel mit anschließender Extraktion sowohl ihrer strukturierten textlichen Inhalte als auch ggf. enthaltener Bilddateien. Ziel ist die Vorbereitung dieser Objekte gepaart mit ihren Lizenzinformationen sowie beschreibenden Text-Elementen zur automatischen Inhaltserschließung (AP 3).

AP 2.1 Bestandsaufnahme der Quellen, APIs, Lizenzen und Medientypen

Zur Vorbereitung des fokussierten Crawlings nach Open-Access-Artikeln, die MOA enthalten, wird zunächst mittels Stichproben in einem Massenmodell ermittelt,

- wieviele der jährlich publizierten Artikel in qualitätsgesicherten Open-Access-Journals jeweils von welchen Publishern produziert werden, die im Directory of Open Access (DOAJ) Journals nachgewiesen sind (am 16.10.2014 gab DOAJ an, dass aus den 10.037 verzeichneten Journals 5.884 auf Artikelebene durchsuchbar sind und auf diese Weise 1.749.966 Artikel erreicht werden können),
- wieviele Artikel via OAI-PMH, RSS/ATOM-Feeds oder andere APIs bezogen werden können, die vom jeweiligen Publisher angeboten werden, einschließlich multimedialer Objekte, die ggf. als ergänzendes Material ("supplementary material") in den Metadaten der Artikel verlinkt werden,
- wieviele Artikel als Open-Access-Veröffentlichungen im Sinne der Berliner Erklärung oder gemäß der damit kompatiblen Openness-Anforderung von Wikimedia Commons identifiziert werden können (dabei handelt es sich um ein unumgängliches Ausschlusskriterium für die Weiterverarbeitung mit Textmining-Methoden, vgl. AP 3),
- wieviele Artikel in welchen Formaten (XML, PDF, HTML, EPUB etc.) angeboten werden.

Anhand des Massenmodells wird geplant, welche Erweiterungen des bereits produktiven Crawlers für Artikel aus PubMed Central (AP 2.2) erforderlich sind.

AP 2.2 Fokussiertes Crawling

Die Erweiterung des bereits aktiven Crawlers für Artikel aus PubMed Central verfolgt das Ziel, die aktiven in DOAJ nachgewiesenen Journals umfassend automatisiert zu harvesten, einschließlich der folgenden Extraktion und Bereinigung aller geharvesteten Artikel (AP 2.3) zur Vorbereitung der automatischen Erschließung (AP 3).

Die zu entwickelnden Crawler erfüllen zwei Kernfunktionen:

Täglich werden vollständige Kopien aller neuen Artikel aus den geharvesteten Journals auf Wikisource abgelegt und die zugehörigen Media-Dateien auf Wikimedia Commons. Quelle für das Harvesting können sowohl die Webseiten der Publisher als auch fachliche Repositorien wie PubMed Central oder arXiv sein.

In Wikidata wird für jeden gespeicherten Artikel eine Entität angelegt, welche die Ursprungs-URL sowie alle ggf. beim Crawl ermittelten formalen Eigenschaften des Artikels (insbesondere Titel, Autorennamen, Veröffentlichungsdatum, Lizenzinformationen, verwendetes Dateiformat und ggf. persistente Identifier des Artikels) sowie eine Verknüpfung auf das veröffentlichende Journal enthält.[22]

AP 2.3 Aufbereitung der Artikel und Medienobjekte (Extraktion und Bereinigung)

Die Erweiterung der bereits produktiven Crawler in diesem Arbeitspaket verfolgt insbesondere das Ziel, den strukturierten Inhalt der Artikel aus ihren z.T. heterogenen Quellformaten (insbesondere Artikeln, die ausschließlich im PDF-Format veröffentlicht wurden) zu extrahieren, auf ein einheitliches Modell abzubilden und in Wikidata zu beschreiben.

Die zu entwickelnde Pipeline erfüllt schrittweise die folgenden sechs Funktionen:

Der strukturierte Inhalt der Artikel wird mittels einer Kette freier Methoden wie zum Beispiel pdf2htmlEX extrahiert (Garcia, Murray-Rust, Burns et al. 2013, Constantin, Pettifer, Voronkov 2013), einschließlich automatischer Erkennung und Entfernung von Rauschen. Das Rauschen entsteht insbesondere durch die fehlerhafte Ausgabe von Text-Artefakten bei der Extraktion von PDF-Dateien, sowie durch die Extraktion von Bilddateien ohne Bezug zum Inhalt der Artikel (z.B. Verlags-Logos oder Layout-Elemente) oder Bilddateien ohne Bildcharakter (z.B. Formeln oder Tabellen). Es ist zu erwarten, dass aus einer Restklasse von Dokumenten der strukturierte Inhalt nicht mit hinreichender Qualität zu extrahieren sein wird. Für diese Artikel werden die Methoden aus AP 3 auf den unstrukturierten Text des Artikels als Ganzes angewendet werden, während z.B. die Zuordnung von Bildbeschreibungen zum jeweiligen Medienobjekten fehlen wird.

Alle im Text des Artikels als "supplementary material" referenzierten multimedialen Objekte werden mit dem Crawler (vgl. AP 2.2) geharvestet und als Anhang zur Kopie des Artikels ergänzt.

[22] Hierfür brauchen die Vorarbeiten aus https://github.com/mitar/bib2wikidata nur noch geringfügig erweitert zu werden.

Der strukturierte, bereinigte Inhalt wird zur Weiterbearbeitung in einem an JATS orientierten XML-Modell abgebildet. Dies erfolgt durch Weiterentwicklung des vorhandenen JATS-to-MediaWiki Converters[23].

Es wird ein Dubletten-Abgleich zwischen allen extrahierten (oder als zusätzliches “supplementary material” geharvesteten) multimedialen Objekten sowie multimedialen Objekten vorgenommen, die bereits zuvor auf Wikimedia Commons enthalten waren. Nicht nur Bitstream-identische Dubletten werden auf diese Weise ermittelt; mit neuen, speziell auf dem Korpus der Bilder auf Wikimedia Commons entwickelten Methoden[24] wird es darüber hinaus möglich sein, mit hohem Recall eine Liste potenzieller Dubletten zu generieren, die dann intellektuell von Wikipedia-Autoren überprüft werden kann.

Die Dateiformate aller extrahierten, dublettenbereinigten multimedialen Objekte wird gegen die von Wikimedia Commons verwendete Whitelist nicht-proprietärer Formate[25] geprüft und ggf. mittels Wikimedia-eigener Methoden in freie Dateiformate umgewandelt.

Im abschließenden Schritt wird für alle extrahierten multimedialen Objekte eine Wikidata-Entität angelegt. (Im Fall einer erkannten Dublette durch ein Merging mit der Entität des zuvor bestehenden Objekts.) Die Entität enthält eine Verknüpfung auf den Artikel, aus dem das Objekt stammt, sowie automatisch ermittelte formale Eigenschaften des Objekts wie Dateityp, Dateigröße, ggf. formatspezifische Angaben wie Anzahl der Pixel, und ggf. normalisierte[26] verknüpfte Lizenzinformationen.

2.3.3 AP 3 - Erschließung

Ziele

- Automatische Erschließung und Anreicherung von MOA (Abbildungsdateien)
- Kombination von Kontextdaten (umgebender Text) und Bilderkennungsdaten
- Klassifikation mit Wikidata: Entity Recognition und Mappen der Konzepte auf Wikidata

Aufgaben/Beschreibung

Aufgabe des AP3 ist es, strukturierte Metadaten für die Bilddateien aus vorhandenen Kontextdaten zu extrahieren, die Beschreibung der Bilder in Wikidata um inhaltliche Metadaten zu ergänzen und für die Suche bereitzustellen.

AP 3.1 Modellierung

AP 3.1. legt am Projektanfang fest, welche Aspekte eines MOA beschrieben und wie diese Metadaten gespeichert werden sollen. Die Auswahl der Aspekte soll sich dabei einerseits an den Anforderungen der Anwender orientieren (Auffindbarkeit, Kompabilität zu anderen Daten, (Langzeit)archivierung) aber auch die Umsetzbarkeit und die zu erwartende Fehlerquote für die gewählte Beschreibungsebene berücksichtigen. Die Beschreibung der Daten baut auf bestehenden Standards, wie die COMM (Core Ontology for Multimedia), auf oder lehnt sich daran an. Es werden praktische Richtlinien für die Verwendung erarbeitet. Die Beschreibung der Artikel und Bilder wird auf das Modell für Artikel als Entitäten in Wikidata bzw. auf das Modell zur Beschreibung von Multimedia-Objekten in Wikimedia Commons zugeschnitten, so dass die in AP 3.2 gewonnenen Verknüpfungen nach Wikidata bzw. Wikimedia Commons importiert und entsprechend der Erschließung weitere Wikidata-Eigenschaften und Aussagen generiert werden können.

AP 3.2 Metadatengewinnung

Im AP 3.2 soll eine Reihe effizienter Softwaremodule zur automatisierten Metadatengenerierung implementiert werden. Das Arbeitspaket strebt nicht in erster Linie die Entwicklung neuer Verfahren an, sondern vielmehr sollen vorhandene opensource Komponenten genutzt werden, wissenschaftlich validierte Verfahren in neuen opensource Modulen implementiert und eine

[23] https://github.com/wpoa/JATS-to-Mediawiki
[24] https://outreach.wikimedia.org/wiki/GLAM/Newsletter/September_2014/Contents/Special_story
[25] http://commons.wikimedia.org/wiki/Commons:File_types
[26] https://github.com/wpoa/Open-License-Dictionary

Analysestraße entwickelt werden. Für die Implementierung der Analysestraße soll ein skalierbares und modulares Framework, wie z.B. Apache UIMA (Unstructured Information Management Architecture[27]), genutzt werden.

Im beantragten Projekt sollen beschreibende Metadaten für MOA aus textuellen Kontextdaten gewonnen werden. In der Pilotanwendung (AP 4.2) soll eine Bildersuchmaschine implementiert werden. Hierfür erwartet eine Anwenderin mindestens die Möglichkeiten, die sie auch von der Bildersuche von Google gewöhnt ist. In der Suchoberfläche soll daher eine Einschränkung nach Bildgröße und Bildtyp (Farbfoto, Schwarz-Weiß-Foto, Strichzeichnung, usw.) möglich sein. Hierzu soll mit Hilfe einer gängigen Open Source Bibliothek, wie z.B. OpenCV[28], ein einfacher Bildklassifikator trainiert und eingebunden werden.

Die Generierung der übrigen beschreibenden Metadaten erfolgt in drei Schritten. Im ersten Schritt werden Texte zum Bild gesammelt und hieraus alle Abschnitte, die sich auf das Bild beziehen, identifiziert. Außer den Bildunterschriften kommen hierfür hauptsächlich Sätze, in denen auf das Bild verwiesen wird, in Frage. Im zweiten Schritt werden Kandidaten für beschreibende Terme aus diesen Textabschnitten extrahiert. Diese werden im dritten Schritt klassifiziert und den richtigen Metadatenfeldern und Wikidata-Konzepten zugeordnet.

Im Extraktionsschritt werden also in den Bildunterschriften und verweisenden Texten die Wörter und Phrasen, die als beschreibende Terme in Frage kommen, identifiziert. Das sind erstens alle Bezeichner der Wikidata-Konzepte und Ankertexte zu den Verweisen auf die Konzepte. Dann folgt die Identifizierung aller *Named Entities* und Mehrwortlexeme, für deren Erkennung ein *Conditional Random Fields*-Modell auf bereits annotierten Daten trainiert werden soll, und schließlich weitere Substantive.

Im Klassifikationsschritt sollen zwei Arten der Klassifikation vorgenommen werden. Erstens sollen die Terme möglichst auf Wikidata-Konzepte normalisiert werden. Das heißt, dass für jeden gefundenen Term (oder Named Entity) ein bedeutungsgleicher Term aus dem Wikidata-Thesaurus gefunden werden soll. Als erfolgreiches Verfahren hat sich hier eine binäre Klassifikation für jedes Wortpaar, bei der verschiedene Ähnlichkeitsmaße benutzt werden, erwiesen (Bär et al. 2012, Wartena 2013b). Als Merkmale kommen u.a. distributionelle Ähnlichkeit, n-Gramm-Überschneidung, Häufigkeit der Verwendung als Ankertext und Editierdistanz in Frage. Wenn vorhanden, können auch Merkmale aus der Bilderkennung und nutzergenerierte Daten mit diesem Verfahren auf Wikidata-Konzepte abgebildet werden (Wartena und Brussee, 2008). Die distributionelle Ähnlichkeit zwischen zwei Termen wird bestimmt durch die Übereinstimmung der Kontexte, in denen die Terme vorkommen. Üblicherweise werden als Kontextmerkmale umgebende Wörter genutzt. Im Rahmen dieses Projektes stellt sich die Frage, inwieweit auch nicht-textuelle Merkmale, z.B. aus der Bilderkennung, als sinnvolle Kontextmerkmale dienen können. Als Trainingsdaten für die Klassifikation als Synonym/Nicht-Synonym kommen nicht nur bereits manuell annotierte Daten in Frage, sondern auch Weiterleitungen aus Wikipedia sowie Informationen aus Wikidata: die alternativen Benennungen für ein Konzept, die in Wikipedia aufgenommen sind, für die aber direkt auf einen anderen Eintrag umgeleitet wird, sind genau von der Art, wie die Varianten, die man in einem Text finden könnte.

Alle gefundene Terme und Wikidata-Konzepte können direkt für die einfache stichwortbasierte Suche genutzt werden (im Falle der Wikidata-Konzepte auch sprachübergreifend). Für die Generierung von Metadaten für die detaillierte Bildbeschreibung müssen die gefundene Terme schließlich noch in den ausgewählten Metadatenfeldern (wie z.B. Personen, Fachrichtungen, Orte) eingeordnet werden. Für diese Klassifikation können wieder distributionelle Merkmale der Terme benutzt werden. Als Trainingsdaten sind manuell annotierte Daten und durch User-Feedback korrigierte Daten erforderlich.

[27] https://uima.apache.org/index.html
[28] http://opencv.org/

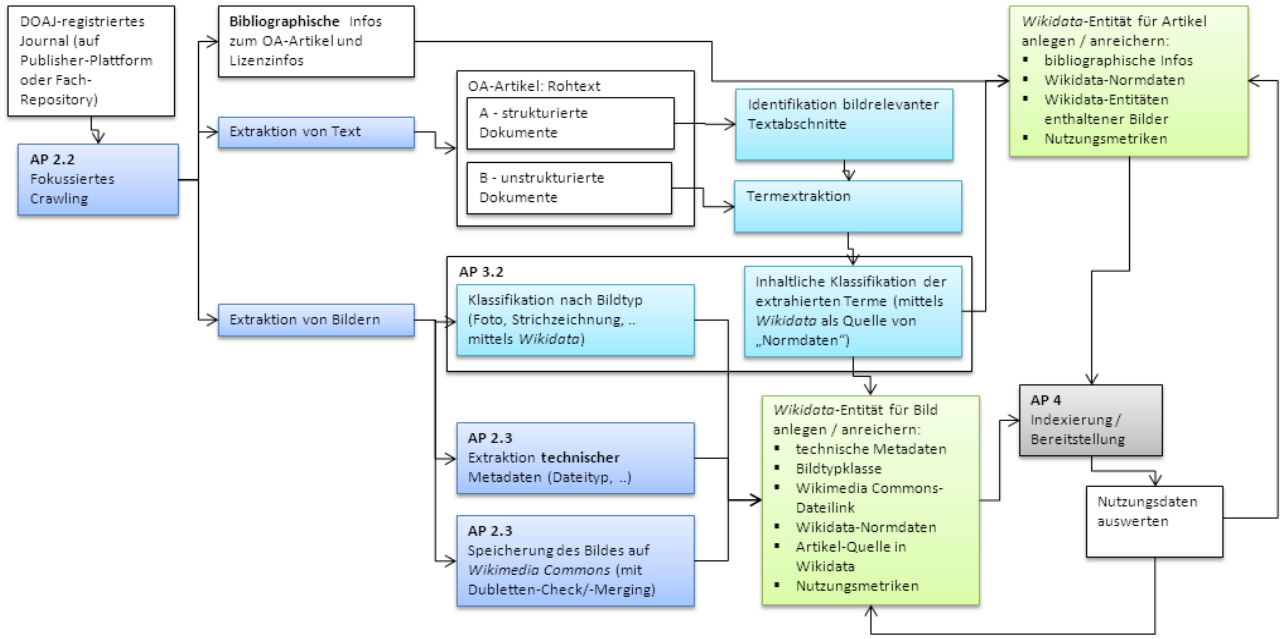

Abbildung 2: Übersicht über die Arbeitsschritte in den Arbeitspaketen 2, 3 und 4

2.3.4 **AP 4 – Nachnutzbare Suchlösungen. Indexerstellung und flexible Bereitstellung der angereicherten MOA**

Ziele

- Bereitstellung eines Abzugs strukturierter Metadaten zur späteren Indexierung
- Suchservice zur facettierten Suche in den MOA sowie mit API
- “Construction Set” mit Skripten zur Reproduktion des Suchservice durch Dritte
- Exemplarische dokumentierte Nachnutzung in TIB GetInfo
- Javascript-basiertes Suchwidget für Wiki-Autoren und -Benutzer
- Verbesserung des Suchservice durch Nutzungsmetriken

Aufgaben/Beschreibung

Die Kernaufgabe des Arbeitspaketes besteht darin, die in den APs 2 und 3 gewonnenen strukurierten Metadaten zur Indexierung durch Dritte sowie für ein dediziertes Suchportal bereit zu stellen. Exemplarisch sollen zudem eine Nachnutzung in einem bibliothekarischen Discovery-Portal sowie eine Nachnutzung via API in einem Suchwidget demonstriert werden.

AP 4.1 Metadatenaggregation und -bereitstellung

AP3 generiert Metadaten zu den in AP 2 geharvesteten Bildern. Während der Projektlaufzeit wird einerseits die Zahl der beschriebenen Objekte wachsen und sich andererseits die Qualität der Metadaten ständig verbessern. Die aktuellen Metadaten werden von AP 3 in einem systemneutralen Datenformat zur Verfügung gestellt. AP 4.1. entwickelt automatische Routinen zum Import der neuen und aktualisierten Metadaten zu den Medienobjekten aus einem Peer Review unterzogener wissenschaftlicher Literatur in das kollaborative Medienarchiv Wikimedia Commons. Das regelmäßig aktualisierte Metadatenpaket wird zwecks Nachnutzung oder Analyse für Dritte unter einer freien Lizenz zur Verfügung gestellt.

AP 4.2 Suchservice

Die Daten aus AP 4.1 sollen als Pilotanwendung mittels einer verbreiteten Open Source Enterprise Search Plattform (wie Lucene in Verbindung mit ElasticSearch oder Solr) durchsuchbar gemacht werden. Dazu wird ein geeignetes Indexierungkonzept erstellt. Neben einer Browser-Suchoberfläche wird mit dieser Plattform eine flexible API bereit gestellt, die von Dritten angesprochen werden kann, z.B. in Widgets, Meta-Suchen oder für Analyse-Anwendungen. Die

Browser-Suchoberfläche soll Benutzererwartungen der Bildersuche von Google entgegenkommen, wozu in AP 2 u.a. formale Metadaten wie die Bildgröße generiert werden und in AP 3 nach Bildtyp (Farbfoto, Schwarzweißfoto, Strichzeichnung usw.) klassifiziert wird.

Die Gestaltung der Suchergebnislisten orientiert sich am Ansatz der “Universal Search” (Sullivan 2007, Quirmbach 2009). Diese konfrontiert den Suchenden nicht mit zusätzlichen Suchräumen, die als Tabs über dem Suchschlitz zur Variation der Suchanfrage angeboten werden, oder (im Fall eines einheitlichen Suchindex z.B. in vielen bibliothekarischen Discovery-Portalen) als Facetten neben der Suchergebnisliste zur Verfeinerung der Suche, und vermeidet damit einen kognitiven Overload beim Benutzer. Stattdessen werden dynamisch die für die jeweilige Suchanfrage als besonders relevant ermittelten Facetten (anhand von Merkmalen wie Bildklasse und fachlicher Zuordnung) ermittelt und in Abschnitten untereinander als Ergebnisliste gezeigt. Damit wird eine vielfältige, subjektiv “interessante” Suchergebnis-Darstellung angestrebt. Bei der geplanten Weiterentwicklung des Wikimedia-eigenen Suchportals kann das Konzept der “Universal Search” ggf. übernommen oder variiert werden.

AP 4.3 “Construction Set” und exemplarische Nachnutzungen

Um Dritten die Möglichkeit zu geben, den in AP 4.2. aufgebauten Suchservice auch selbst zu reproduzieren, selbstständig zu betreiben oder zu erweitern, wird in dem Suchservice weitestgehend mit marktüblichen Open-Source-Tools gearbeitet. Für den gesamten Workflow einschließlich der Indexierung und dem Betrieb des Suchportals (AP 4.2) werden zudem alle verwendeten Skripte unter einer freien Lizenz zur Verfügung gestellt und nachvollziehbar dokumentiert.

An der TIB Hannover wird modellhaft ein fachlich definierter Auszug des Indexes für den Bereich der naturwissenschaftlichen und technischen Fächer abgezogen und als optional vom Benutzer zuwählbare Suchquelle in den Index des Suchdienstes GetInfo eingebunden.

Für Autoren und Benutzer der Wikipedia und anderer MediaWiki-basierter Wikis wird ein Javascript-basiertes Tool erstellt.[29] Das Tool generiert mittels der API des Suchportals (AP 4.2) automatisch eine Vorschau-Galerie mit wissenschaftlichen Bildern, die inhaltlich zur gerade betrachteten Seite passen, auf dieser aber noch nicht enthalten sind. Diese Bilder können dann mit wenigen Mausklicks in die jeweilige Seite übernommen werden, einschließlich Quellenangaben und Zitation der originären Bildbeschreibung.[30]

AP 4.4 Verbesserung des Suchservice durch Nutzungsmetriken

Es wird ein Verfahren zur automatischen Auswertung von Aufrufzahlen und Referrer-URLs in den bereits vorhandenen Logfiles für die Aufrufe von Objekt-Beschreibungsseiten in Wikimedia Commons erstellt, um a) zusätzliche offene Metriken zur (passiven) Benutzung der Objekte in deren Beschreibung anbieten zu können, die z.B. von Altmetrics-Analysediensten nachgenutzt werden können, b) das Ranking der Suchergebnisse in AP 4.2 zu verbessern und c) Revisionslisten zur intellektuellen Nachprüfung von in AP 3 automatisch vorgenommenen Klassifikationen zu generieren. Diese Maßnahme soll dabei helfen, das Problem des erwarteten guten Recalls bei (zu) geringer Precision der automatischen Klassifikationsergebnisse zu kompensieren.

2.4 Maßnahmen zur Erfüllung der Förderbedingungen und Umgang mit den Projektergebnissen

Die aus dem Projekt resultierenden Publikationen ebenso wie einschlägige Dokumentationen werden im Open Access verfügbar gemacht und Dritten zur umfassenden Nachnutzung bereitstehen. Der im AP 4 aufgebaute Suchservice kann durch Dritte nachgenutzt werden, das heißt reproduziert, selbständig betrieben oder erweitert werden. Für den gesamten Workflow von der Generierung der Rohdaten bis hin zur Indexierung und dem Betrieb des Suchportals werden zudem alle verwendeten Skripte unter einer freien Lizenz zur Verfügung gestellt und nachvollziehbar dokumentiert.

---

[29] basierend auf dem in https://meta.wikimedia.org/wiki/GlobalCssJs und https://www.mediawiki.org/wiki/Manual:Interface/JavaScript vorgestellten Frameworks

[30] Eine technisch vergleichbare Anwendung wird unter http://vimeo.com/109435794 vorgestellt.

Dieses Projekt hat Pilotcharakter und liefert mit dem entwickelten Verfahren ein auch fächer- und medienübergreifend nachnutzbares Beispiel zur automatischen Sammlung, Erschließung und Bereitstellung multimedialer Open-Access-Objekte.

Die Projektförderung dieses Vorhabens unterstützt den Aufbau eines längerfristigen, überregionalen Informationsdienstes.

Bereits existierende Standards (Metadatenstandards, Creative-Commons-Lizenzen für Open-Access-Publikationen) werden im Rahmen dieses Projektes berücksichtigt und nachgenutzt.

3 Literaturverzeichnis

Bär, D.; Biemann, C.; Gurevych, I.; Zesch, T. (2012). UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval-2012), S. 435–440.

Berndt, R.; Blümel, I.; Clausen, M.; Damm, D.; Diet, J.; Fellner, D.; Fremerey, C.; Klein, R.; Krahl, F.; Scherer, M. (2010). The PROBADO project-approach and lessons learned in building a digital library system for heterogeneous non-textual documents. *Research and Advanced Technology for Digital Libraries*, S. 376-383, Springer.

Blümel, I., Berndt, R., Ochmann, S., Vock, R., & Wessel, R. (2010). PROBADO3D-Indexing and Searching 3D CAD Databases: Supporting Planning through Content-Based Indexing and 3D Shape Retrieval. *Proceedings of 10th International DDSS Conference*.

Brase, J.; Blümel, I. (2010). Information supply beyond text: non-textual information at the German National Library of Science and Technology (TIB)–challenges and planning. *Interlending & Document Supply*, 38, 2, S. 108-117, Emerald Group Publishing Limited.

Bullinaria, J. A..; Levy, J. P. (2012). Extracting Semantic Representations from Word Co-occurrence Statistics: Stop-lists. *Stemming and SVD* (44), S. 890–907.

Constantin, A.; Pettifer, S.; Voronkov, A. (2013). PDFX: fully-automated PDF-to-XML conversion of scientific literature. *Proceedings of the 2013 ACM symposium on Document engineering*.

Feng, Y.; Lapata, M. (2010a). How many words is a picture worth? Automatic caption generation for news images. *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*.

Feng, Y.; Lapata. M. (2010b). Topic models for image annotation and text illustration. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Frank, E.; Paynter, G.W.;Witten, I.H.; Gutwin, C.; Nevill-Manning, C.G. (1999). Domain-specic keyphrase extraction. Proceedings of the Sixteenth International Joint Conference on Articial Intelligence, IJCAI 99, Stockholm, Sweden, July 31 - August 6, 1999. S. 668-673.

Garcia, A.; Murray-Rust, P.; Burns, G. A.; Stevens, R.; Tkaczyk, D.; McLaughlin, C.; Wabiszewski, M. (2013). PDFJailbreak – a communal architecture for making biomedical PDFs semantic. *Proceedings of BioLINK SIG 2013*. S. 13.

Gazendam, L.; Wartena, C.; Brussee, R. (2010). Thesaurus Based Term Ranking for Keyword Extraction. *DEXA Workshops: IEEE Computer Society*. S. 49-53.

Hentschel, C.; Blümel, I.; Sack, H. (2013). Automatic Annotation of Scientific Video Material based on Visual Concept Detection. *Proceedings of 13th International i-know Conference*, ACM Publishing.

Heller, L. (2011a). Vortrag zur Verbundkonferenz des GBV und gleichzeitig zur Jahrestagung der Deutschen Gesellschaft für Klassifikation (Gfkl). [veröffentlicht auf slideshare.net]

Heller, L. (2011b). Sacherschließung von Literatur in und mit der Wikipedia — einfach anfangen? [veröffentlicht auf biblionik.de]

Jin, Y. et al. (2005). Image annotations by combining multiple evidence & Wordnet. *Proceedings of the 13th annual ACM international conference on Multimedia*.

Kiela, D.; Clark, S. (2014). A Systematic Study of Semantic Vector Space Model Parameters. *2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*.

Larson, M.; Soleymani, M.; Serdyukov, P.; Rudinac, S.; Wartena, C., Murdock, V.; Friedland, G.; Ordelman, R.; Jones, G. (2011). Automatic Tagging and Geotagging in Video Collections and Communities. *ACM International Conference on Multimedia Retrieval (ICMR 2011).*

Leong, C. W.; Mihalcea, R.; Hassan, S. (2010). Text mining for automatic image tagging. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*.

Lew, M. S., et al. (2006). Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)* 2.1. S. 1-19.

Liu, Y. et al. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40.1. S. 262-282.

Martinelli, L. (2014). Wikidata: A New Way to Disseminate Structured Data. Faster, Smarter and Richer. Reshaping the library catalogue, Roma (Italy), 27-28 February 2014.

Medelyan, O.; Witten, I. H.; Milne, D. (2008). Topic indexing with Wikipedia. *Proceedings of the AAAI WikiAI Workshop*.

Mihalcea, R.; Csomai, A. (2007). Wikify!: linking documents to encyclopedic knowledge. Proceedings of the sixteenth ACM conference on Conference on information and knowledge management.

Pscheida, D.; Albrecht, S.; Herbst, S.; Minet, C.; Köhler, T. (2013). *Nutzung von Social Media und onlinebasierten Anwendungen in der Wissenschaft*. ZBW – Deutsche Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft.

Quirmbach, S. (2009). Universal Search: Kontextuelle Einbindung von Ergebnissen unterschiedlicher Quellen und Auswirkungen auf das User Interface. *Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und* Praxis. S. 175-203.

Saif, M.; Hirst, G. (2012). Distributional Measures of Semantic Distance: A Survey.

Srikanth, M. et al. (2005). Exploiting ontologies for automatic image annotation. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.

Sullivan, D. (2007). Google Launches “Universal Search” & Blended Results. [http://searchengineland.com/google-20-google-universal-search-11232].

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information Retrieval* 2(4). S. 303-336.

Turney, P. D.; Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37. S. 141–188.

Wartena, C.; Garcia-Alsina, M. (2013). Challenges and Potentials for Keyword Extraction from Company Websites for the Development of Regional Knowledge Maps. *Proceedings of the 5th International Conference on Knowledge Discovery and Information Retrieval (KDIR 2013), Vilamoura, Portugal*. S. 514-519.

Wartena, C. (2013a). Distributional Similarity of Words with Different Frequencies. *Proceedings of the 13th edition of the Dutch-Belgian information retrieval Workshop (DIR 2013),* 2013a, p. 8-11.

Wartena, C. (2013b). Estimating Semantic Similarity of Words and Short Phrases with Frequency Normalized Distance Measures. *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA*.

Wartena, C. (2014). On the effect of word frequency on distributional similarity. *Proceedings of KONVENS 2014*. S. 1-10.

Wartena, C.; Brussee, R.; Wibbels, M. (2009). Using Tag Co-occurrence for Recommendation. *ISDA: IEEE Computer Society*. S. 273-278.

Wartena, C.; Brussee, R., Gazendam, L.; Huijsen, O.-W. (2007). Apolda: A Practical Tool for Semantic Annotation. *DEXA Workshops: IEEE Computer Society*. S. 288-292.

Wartena, C.; Brussee, R. (2008). Instance-Based Mapping between Thesauri and Folksonomies. *International Semantic Web Conference*, *Vol. 5318 of LNCS*. S. 356-370.

Wartena, C.; Brussee, R.; Slakhorst, W. (2010). Keyword Extraction using Co-occurrence. *DEXA Workshops: IEEE Computer Society.* S. 54-58.

Wartena, C.; Slakhorst, W.; Wibbels, M. (2010). Selecting keywords for content based recommendation. *CIKM*. S. 1533-1536.

Wartena, C.; Sommer, M. (2012). Automatic classification of scientific records using the German Subject Heading Authority File (SWD). *Proceedings of the 2nd International Workshop on Semantic Digital Archives (SDA 2012).* S. 37-48.

Wartena, C.; Wibbels, M. (2011). Improving Tag-Based Recommendation by Topic Diversification. *Advances in Information Retrieval, Lecture Notes in Computer Science* 6611. S. 43-54.

Wessel, R.; Blümel, I.; Klein, R. (2009). A 3D Shape Benchmark for Retrieval and Automatic Classification of Architectural Data. *Eurographics 2009 Workshop on 3D Object Retrieval*. S. 53-56.