Bild, Beschreibung, (Meta)Text Automatische inhaltliche Erschließung und Annotation kunsthistorischer Daten

Dieckmann, Lisa
lisa.dieckmann@uni-koeln.de
Universität zu Köln, Deutschland

Hermes, Jürgen
hermesj@uni-koeln.de
Universität zu Köln, Deutschland

Neuefeind, Claes
c.neuefeind@uni-koeln.de
Universität zu Köln, Deutschland

Der Vortrag thematisiert die automatische inhaltliche Erschließung und linguistische Annotation der digitalen Repräsentationen kunst- und kulturhistorischer Artefakte innerhalb des prometheus Bildarchivs (http://prometheus-bildarchiv.de), das derzeit 89 Datenbanken aus Museen und Forschungsinstitutionen mit insgesamt über 1,5 Mio. Datensätzen zusammenführt (Dieckmann 2015). Das durch eine bereits abgeschlossene Vorstudie initiierte Projekt verfolgt zwei unterschiedliche, teilweise aufeinander aufbauende Ansätze: Zum einen die Annotation von Freitexten zur strukturierten Erschließung kunsthistorischer Daten, zum anderen die Analyse der Identität von Datensätzen über die Berechnung gradueller Ähnlichkeiten von Objekten. Beide Ansätze dienen erstens einer Verbesserung des Retrievals, zweitens einer nachhaltigen Sicherung der Daten durch die Verknüpfung mit Normdaten; drittens sollen die zusätzlich erschlossenen Informationen längerfristig als Grundlage für weiterführende (fachspezifische) Fragestellungen eingesetzt werden, etwa zur Rekonstruktion von Künstlergruppen durch die Erstellung von Personen-Netzwerken. Das Projekt wird an der Universität zu Köln in enger Zusammenarbeit zwischen Fachwissenschaftlern der Kunstgeschichte und der Sprachlichen Informationsverarbeitung (http://www.spinfo.phil-fak.uni-koeln.de/) durchgeführt, deren Schwerpunkte u.a. auf Systemen zur syntaktischen und semantischen Analyse und Verarbeitung textueller Daten (Hermes 2012, Schwiebert 2012) sowie zur Annotation nicht-standardisierter Daten (Neuefeind 2013) liegen.

Metadaten und Referenzobjekte

Die digitalen Repräsentationen der in prometheus zusammengeführten kunst- und kulturhistorischen Artefakte stellen insofern eine besondere Herausforderung für die automatisierte Erschließung inhaltlicher Informationen dar, als dass die Metadaten und Texte strukturell und inhaltlich sehr heterogen und in unterschiedlichen Kontexten vorliegen. Die Datensätze der einzelnen Bilddatenbanken sind zwar stets in ein eigenes Metadatenschema eingepasst, jedoch erfolgt die Erschließung der Werke an den jeweiligen Institutionen nicht nach einer einheitlichen Methodik, was u.a. datenbank- oder sammlungsspezifische Gründe hat. Zum einen liegt innerhalb der Klassifikationen der jeweiligen Datenbanken eine Vielzahl an Texten vor, die bislang nicht strukturiert erschlossen sind, sondern derzeit nur über eine einfache Volltextsuche miteinbezogen werden (es handelt sich hierbei oftmals um unstrukturierte Freitextfelder, die z.B. Angaben über Standort(e), die Publikationsgeschichte oder ausführliche Bildbeschreibungen enthalten können). Zum anderen wird selten ein bestimmter Metadatenstandard zugrunde gelegt oder auf Fachvokabulare und Terminologieressourcen zurückgegriffen, was dazu führt, dass zum Teil stark variierende Schreibweisen u.a. bei Künstler- oder Ortsnamen existieren. In der kunsthistorischen Forschung haben sich zudem selten einheitliche Bezeichnungen für Werktitel durchgesetzt. So liegt bspw. das Werk “Bonaparte überquert den großen Sankt Bernhard” von Jacques-Louis David (Malmaison, 1801) in prometheus in mindestens sieben verschiedenen Titelbezeichnungen vor, die zumeist in Teilen, unter Umständen aber auch vollständig voneinander abweichen (etwa “Napoleon überquert die Alpen” gegenüber “Bonaparte auf dem großen Sankt Bernhard”). Eine Verknüpfung mit Normdaten wie der Gemeinsamen Normdatei der Deutschen Nationalbibliothek (GND, http://d-nb.info/gnd/1067141367) ist auf dieser Grundlage nicht möglich. Diese wäre aber nötig, um eine automatische Zusammenführung der Einzelabbildungen zu Objekten vornehmen und die Objekte eindeutig und damit nachhaltig identifizieren zu können, was zugleich die Grundlage für eine weitere Anreicherung mit GND-verknüpften Daten oder weiteren Normdaten (z.B. VIAF, http://viaf.org ; Wikidata, https://www.wikidata.org) bilden würde.

Methodologie

Die Heterogenität der Daten wird in prometheus bereits teilweise in Anwendung linguistischer Analyseverfahren bei der Indexierung ausgeglichen, wobei der Schwerpunkt hier v.a. auf der orthographischen und morphosyntaktischen Ebene liegt, etwa auf Grundlage sprachspezifischer Wörterbücher (u.a. zur Grundformreduzierung, Phrasenerkennung,

Synonymgenerierung, Kompositazerlegung) sowie durch Anreicherung mit synonymen Künstlernamen (siehe http://prometheus-bildarchiv.de/tools/pkn d). Diese Maßnahmen dienen in erster Linie dazu, das Retrieval zu optimieren und den Recall zu verbessern. In Bezug auf die oben aufgeworfenen Probleme der Normalisierung und Zuordnung von Einzeldarstellungen zu Objekten sind sie jedoch nur als ein erster Schritt anzusehen. Ziel ist vielmehr ein erweiterter Thesaurus, in dem die tatsächlich auftretenden, zum Teil stark variierenden Schreibweisen von Werktiteln und Künstlernamen auf die verfügbaren Normdaten abgebildet werden. Da die Variation in den Schreibweisen keine eindeutige Zuordnung erlaubt, bedarf es hierbei zusätzlicher Kriterien. Im Zuge des Projekts wird hierfür ein semantisch motiviertes Verfahren erarbeitet, das die gesamten zu einem Objekt verfügbaren Informationen berücksichtigt: Neben den bereits erschlossenen Metainformationen (wie Name, Titel, Datierung, Standort, etc.) sollen auch die in den bislang nur unstrukturiert vorliegenden Freitextfeldern (s.o.) enthaltenen Informationen genutzt werden können. Zu diesem Zweck werden die Texte zunächst mittels Informationsextraktion aufbereitet (Annotation von Orts- und Personennamen, Zeitausdrücken, etc). Auf Grundlage dieser neu gewonnenen Informationen werden zusätzliche, das Objekt beschreibende Merkmale erstellt und in Form von Feature-Vektoren kodiert (Features sind z.B. „Personen“, „Orte“, „Material“, o.ä.; Werte sind jeweils die konkreten Nennungen, vgl. Abb. 1).

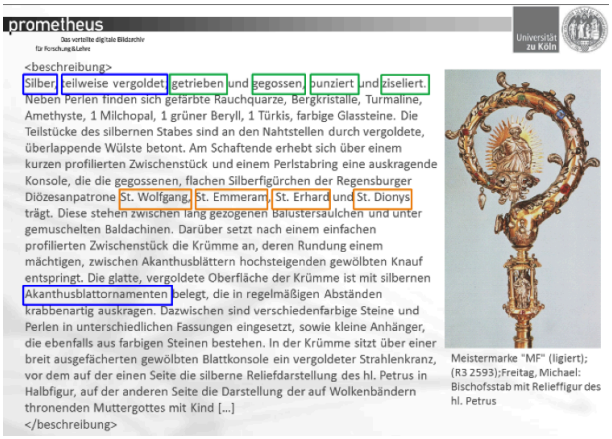

Abb. 1: Beispiel einer Freitextbeschreibung im prometheus-Bildarchiv, in der exemplarisch mittels Informationsextraktion identifizierte Elemente markiert wurden.

Aus den zusätzlichen Merkmalen kann nun, in Kombination mit den bereits vorhandenen Metainformationen, für jedes Objekt ein „semantisches Profil“ bzw. „Fingerprint“ erstellt werden, anhand dessen sich die Ähnlichkeit zwischen Objekten ermitteln lässt. Die Ähnlichkeit wird dabei zunächst in Bezug auf die einzelnen Merkmale ermittelt (u.a. mittels Edit-Distance oder Soundex- bzw. Metaphone-Difference zwischen einzelnen Feldern, Abgleich zeitlicher Angaben, Distanz zwischen Feature-Vektoren zu „Personen“, „Orten“, „Material“, etc.), wobei der Einfluss einzelner Merkmale unterschiedlich gewichtet werden kann. Daraus wird ein kombiniertes Maß der Übereinstimmung zwischen zwei Datensätzen errechnet, das auch bei deutlich abweichenden Schreibweisen eine Aussage darüber erlaubt, ob es sich um das gleiche Objekt handelt. Auf dieser Grundlage können identische Objekte dann auf das jeweilige Referenzobjekt der GND abgebildet werden.

In einem vorbereitenden Projekt für das laufende Vorhaben wurden zunächst die bestehenden Metadaten der einzelnen Datenbanken des prometheus-Bildarchivs quantitativ ausgewertet, um einen Überblick darüber zu erlangen, wie sich der Umfang der zu erschließenden Daten darstellt. Die meisten der 89 Datenbanken verfügen über noch nicht erschlossene Freitextbeschreibungen der Objekte. Diese erstrecken sich zu einem nicht geringen Teil über mittellange (25-75 Wörter) und lange (>75 Wörter) Texte, die im Zuge des Projekts aufbereitet werden sollen. Abb. 2 zeigt die Verteilung dieser unterschiedlichen Textsorten in ausgewählten Datenbanken. Einige verfügen über keinerlei Freitext-Bildbeschreibungen, z.B. die Datenbank des *Zentralarchivs für Kunstgeschichte* in München (zi_muc). Andere, etwa die Erlanger Datenbank *Zeichnungen der graphischen Sammlung* (erlangen_z), weisen fast ausschließlich kurze Beschreibungen auf, wieder andere enthalten dagegen auch eine Reihe mittellanger und langer Texte.

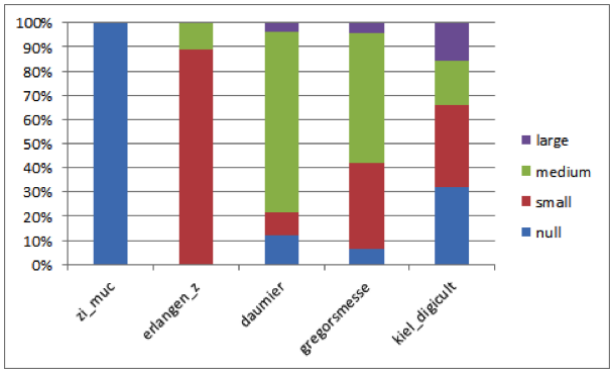

Abb. 2: Verteilung der Freitextlängen über verschiedene Datenbanken des prometheus-Bildarchivs

Zur Nutzung der in den Bildbeschreibungen und Ikonographien enthaltenen Informationen müssen diese zunächst identifiziert und entsprechend ausgezeichnet werden. Dafür wurde zunächst ein Komponenten-Workflow konzipiert und auf Basis des UIMA-Frameworks (Unstructured Information Management Architecture, siehe https://uima.apache.org) implementiert. Im Zuge der Verarbeitung werden die zu annotierenden Informationen in ausgewählten Feldern der Datensätze identifiziert (vgl. Abb. 3). Dabei kommen Standardmethoden

der Informationsextraktion (z.B. Temporal Expression Detection, Named Entity Recognition) genauso zum Einsatz wie informationstheoretische Maße (etwa Log-Likelihood oder tf.idf), um domänenspezifisch relevante Terme zu bestimmen.

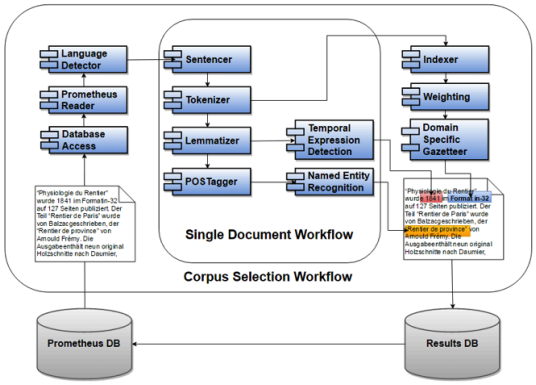

Abb. 3: Workflow zur Informationsextraktion in Freitexten im prometheus-Bildarchiv.

Die erste Projektphase diente vor allem der Evaluierung von Werkzeugen, etwa dem Stanford Named Entity Tagger (siehe http://nlp.stanford.edu/software/CRF-NER.shtml) zur Identifikation von (Orts-)Namen, oder HeidelTime (siehe http://dbs.ifi.uni-heidelberg.de/index.php?id=129) zur Annotation von Zeitausdrücken, um den voraussichtlichen Bedarf an Anpassungen der Werkzeuge für die kunsthistorische Domäne zu ermitteln. Abb. 4 zeigt das Ergebnis der StanfordNER-Komponente, die in einem Text der Datenbank “The Daumier Register” (http://www.daumier-register.org/) Eigennamen auszeichnet: Künstlernamen (“Henry Monnier”), Werktitel (“Séraphita”), Werkstoffe (“China-Papier”), sowie Ortsnamen (“Sevilla”) werden mit verschiedenen Tags (I-PERS, I-MISC, I-LOC) gekennzeichnet. Wie sich zeigt, werden jedoch nicht alle Eigennamen aufgefunden (etwa “Balzac, Honoré de”), was v.a. darauf zurückzuführen ist, dass hier zunächst nur das für das Deutsche verfügbar NER-Modell zum Einsatz kam. In der weiteren Projektlaufzeit muss die Erkennungsrate durch eine Erweiterung und Modifikation der vorhandenen Modelle verbessert werden, damit die Daten möglichst präzise und vollständig auswerten werden können.

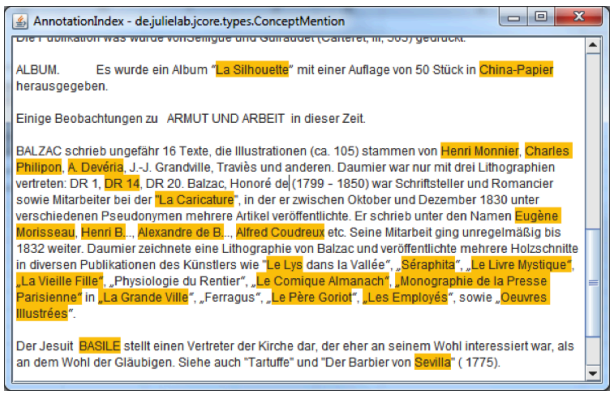

Abb. 4: Exemplarisches Ergebnis der Anwendung eines verfügbaren Standard-Modells zur Named Entity Recognition.

Zusammenfassung und Ausblick

Das beschriebene Vorgehen wird derzeit exemplarisch an ausgewählten Datensätzen entwickelt, um anschließend auf den gesamten Bildpool des prometheus-Bildarchivs angewendet zu werden. Ein wesentliches Ziel des Projekts ist es, eine größtmögliche Automatisierung in der Thesauruserstellung zu erreichen. Das hierfür vorgesehene kombinierte Ähnlichkeitsmaß ist flexibel erweiterbar. So können zum einen zusätzliche Informationen aus externen Quellen herangezogen werden, etwa indem weitere, digital vorliegende kunsthistorische Texte (z.B. das Reallexikon zur Deutschen Kunstgeschichte, RDK, siehe http://www.rdklabor.de), Ausstellungs- und Auktionskataloge (z.B. Getty Art and Architecture, siehe http://www.getty.edu/research/tools/vocabularies/aat/ , UB Heidelberg http://artsales.uni-hd.de) oder auch Wikipedia analysiert und klassifiziert werden und mittels des erstellten Thesaurus mit den Datensätzen in prometheus verknüpft werden. Zum anderen soll das Ähnlichkeitsmaß auch durch komplementäre (z.B. optische) Verfahren des Bildvergleichs erweitert werden. So wurde bspw. zusammen mit der Computer Vision Group Heidelberg bereits ein Projekt zur automatischen Bilderkennung angestoßen (siehe Bell/Dieckmann 2015). Durch die Kombination verschiedener Methoden der Ähnlichkeitsberechnung zu einem gemeinsamen, multidimensionalen Ähnlichkeitsmaß ist der hier vorgeschlagene Ansatz in hohem Maße adaptierbar für vergleichbare Anwendungen. Die im Projekt erarbeitete Vorgehensweise ist somit auf weitere Metadatenpools kulturhistorischer Inhalte übertragbar und dank der Automatisierung beliebig skalierbar.

Bibliographie

Bell, Peter / Dieckmann, Lisa (2015): „Die Kunst als Ganzes. Heterogene Bilddatensätze als Herausforderung für die Kunstgeschichte und die Computer Vision“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* http://dhd2016.de/boa.pdf#118 [letzter Zugriff 23.11.2016].

Bell, Peter / Dieckmann, Lisa / Ommer, Björn / Takami, Masato (2015): *Passion Search. Prototype of an unrestricted image search of the crucifixion*. http://hci.iwr.uni-heidelberg.de/COMPVIS/projects/suchpassion/ [letzter Zugriff 23.11.2016]

Dieckmann, Lisa (2015): „prometheus – das verteilte digitale Bildarchiv für Forschung & Lehre e. V.“, in: Euler, Ellen / Hagedorn-Saupe, Monika/ Maier, Gerald/ Schweibenz, Werner/ Sieglerschmidt, Jörn (eds.): *Handbuch Kulturportale. Online-Angebote aus Kultur und Wissenschaft*. Berlin / Boston: DeGruyter 223–229.

Hermes, Jürgen (2012): *Textprozessierung: Design und Applikation*. Dissertation, Universität zu Köln. http://kups.ub.uni-koeln.de/id/eprint/4561 [letzter Zugriff 23. November 2016].

Neuefeind, Claes (2013): „The Digital Romansh Chrestomathy. Towards an Annotated Corpus of Romansh“, in: Zampieri, Marcos / Diwersy, Sascha (eds.), *Special Volume on Non-Standard Data Sources in Corpus-Based Research* (ZSM Studien 5). Aachen: Shaker 41–58.

Schwiebert, Stephan (2012): *Tesla - ein virtuelles Labor für experimentelle Computer- und Korpuslinguistik*. Dissertation, Universität zu Köln. http://kups.ub.uni-koeln.de/id/eprint/4571 [letzter Zugriff 23. November 2016].