Zwischen Precision und Recall
Information Retrieval in Bildersuchmaschinen am Beispiel von *prometheus*

Lisa Dieckmann

Abstract „Precision" und „Recall" bezeichnen zwei Messgrößen, die die Qualität von Information-Retrieval-Systemen bewerten. Das Ziel ist die Gewinnung von relevanten Informationen und der gleichzeitige Ausschluss von nicht-relevanten Informationen zu einer spezifischen Anfrage. Idealerweise wird eine Erhöhung des „Recall" (Anzahl der gefundenen Dokumente im Verhältnis zu allen relevanten Dokumenten) durch die Anwendung bestimmter Verfahren innerhalb der Suchmaschine erreicht, ohne die „Precision" (Anzahl relevanter Dokumente im Verhältnis zu allen gefundenen Dokumenten) zu verschlechtern. In der Realität ist dies aber eher ein Prozess des Austarierens zwischen beiden Größen. Der Beitrag möchte am Beispiel des *prometheus*-Bildarchivs einen Einblick in die Retrieval- und Bewertungsmechanismen von Suchmaschinen geben und die jeweiligen Entscheidungen für den Einsatz von Verfahren hinsichtlich des Verhältnisses zwischen „Precision" und „Recall" verdeutlichen.

Keywords Information Retrieval, Suchmaschine, Precision, Recall, Relevanz, Ranking

Information Retrieval meint die durch eine Suchmaschine erreichte Gewinnung von Informationen aus Datenquellen zu spezifischen Anfragen.[1] Dabei ist das Ziel, relevante von nicht-relevanten Informationen zu unterscheiden und letztere bei der Auslieferung der Dokumente auszuschließen. Da die Beurteilung von Relevanz jedoch vorwiegend aus Annahmen besteht, kann es dabei allerdings nur um Annäherung gehen. Das bedeutet, „ein gutes IR-System unterscheidet sich von einem schlechten System […] dadurch, dass es einerseits *mehr* relevante Dokumente zum Informationsbedürfnis liefert und andererseits den Anfragenden mit *weniger* irrelevanten Dokumenten im Ergebnis belastet"[2] [beide Hervorhebungen von der Verfasserin]. Das zentrale Kriterium für die Qualität von Information-Retrieval-Systemen heißt also *Relevanz*. Formuliert man nun das spezielle Ziel für Bildersuchmaschinen, wie z. B. vom *prometheus*-Bildarchiv,[3] ist es das Ziel, mithilfe eingegebener Suchbegriffe möglichst viele für diese Begriffe relevante Kunstwerke/kulturelle Objekte mit Abbildungen und Metadaten zu finden und möglichst viele irrelevante Kunstwerke/kulturelle Objekte auszuschließen.[4] Die Qualitätsbewertung des Information Retrieval-Systems erfolgt dabei mithilfe der Messgrößen „Recall" und „Precision", der Anzahl der zurückgelieferten Treffer für eine bestimmte Suchanfrage im Verhältnis zu allen potenziell möglichen Treffern und der Anzahl der relevanten Dokumente innerhalb der Treffermenge. Wenn das *prometheus*-Bildarchiv 1.000 Abbildungen von Albrecht Dürer enthielte, sollten diese bei einer Suchanfrage auch zurückgegeben werden, um einen „Recall" von 100 % zu erreichen. Werden nur 800 gefunden, ergibt das einen „Recall" von 80 %. Werden allerdings 1.200 zurückgegeben, liegt die „Precision" nur bei 80 %. Der Idealfall wird also offensichtlich durch eine Abfrage mit einem „Recall" und einer „Precision" von je 100 % repräsentiert.[5]

Lisa Dieckmann: Zwischen Precision und Recall – Information Retrieval in Bildersuchmaschinen am Beispiel von *prometheus*, in: Effinger et al. (Hrsg.): *Von analogen und digitalen Zugängen zur Kunst – Festschrift für Hubertus Kohle zum 60. Geburtstag*, Heidelberg: arthistoricum.net 2019, S 445–454. https://doi.org/10.11588/arthistoricum.493.c6724

Vor nicht allzu langer Zeit schrieb uns ein Nutzer, der regelmäßig im *prometheus*-Bildarchiv recherchiert, dass, wenn man in den Suchschlitz einen Suchbegriff einträgt, gelegentlich Ergebnisse angezeigt werden, die keinen Sinn ergäben. Für den Nutzer war in diesem Fall klar, dass er, wenn er über die Volltextsuche die Begriffe „Michelangelo" und „David" eingibt, Abbildungen von der gleichnamigen Skulptur von Michelangelo Buonarotti in der Galeria Dell'Academia in Florenz angezeigt bekommt. In den Augen des Nutzers machten einige Beispiele keinen Sinn – ich werde hier zwei unterschiedliche Beispiele herausgreifen:

1. Honoré Victorin Daumier: *La Laveuse* (siehe Abb. 1)[6]
2. Paul Delaroche: *Portrait du comte James-Alexandre de Pourtalès-Gorgier* (siehe Abb. 2)[7]

Betrachtet man die Abbildungen und die rudimentären Metadaten der Ergebnisliste, scheint der Nutzer auf den ersten Blick recht zu haben, jedoch ist das Erscheinen der Datensätze in der Trefferliste vollkommen richtig, wenn man nicht nur Künstler und Titel, sondern alle Metadaten betrachtet, die bei einer Volltextsuche ja schließlich auch berücksichtigt werden. Denn für das erste Beispiel findet sich (etwas versteckt) in der Vollansicht des Datensatzes in dem langen Beschreibungstext folgende Passage: „The focus on humble folk is accompanied by a concern for force and monumentality reminiscent of Michelangelo, showing the spectator a sort of real allegory." Bei den Angaben unter Standort findet sich „David-Weill. Paris (Frankreich)".

Im zweiten Beispiel sind es nicht die Metadaten der Datenbank selbst, sondern die dem Datensatz nachträglich durch das Spiel *Artigo* hinzugefügten Schlagwörter, die via LMU München über eine Programmierschnittstelle (Application Programming Interface, kurz API) abgerufen werden.[8] In dem Kopf des Apoll auf dem Gemälde, in welchem sich der Porträtierte im Rahmen seiner Antikensammlung präsentiert, sah ein *Artigo*-Spieler wohl Ähnlichkeiten mit Michelangelos David (s. Abb. 3).

Auch taucht in der Ergebnisliste z. B. eine Luftaufnahme des Palazzo Vecchio in Florenz[9] auf. Hier ist verständlicherweise „Michelangelo" und „David" verschlagwortet, da sich die Kopie der Skulptur vor dem Palazzo befindet und in der Fotografie abgebildet ist. Dass diese Treffer nicht das Bedürfnis eines Nutzers befriedigen, der Aufnahmen der Skulptur von Michelangelos David finden wollte, dürfte klar sein. Mit diesen Beispielen sind daher schon einige Probleme oder besser Herausforderungen angesprochen: Aus NutzerInnensicht ist der Datensatz nicht relevant, für die Suchmaschine schon – man spricht hier auch von subjektiver und objektiver Relevanz. Die Suchmaschine kennt das Bedürfnis von Nutzerinnen und Nutzern zunächst nicht. Sie kennt nur die beiden Zeichenketten „David" und „Michelangelo" und liefert alle Dokumente zurück, die diese Wörter beinhalten. Jedoch versucht die Suchmaschine (standardmäßig) dennoch, dem Bedürfnis der Nutzerinnen und Nutzer nahezukommen, indem sie die Relevanz bewertet, also einen Score mitliefert, der die Ergebnisse bewertet („rankt").

Die beiden oben genannten Suchergebnisse tauchen demnach in der Ergebnisliste ganz hinten auf als Treffer 383 und 385 von 385 Treffern insgesamt. Also hat die Suchmaschine korrekterweise dem Informationsbedürfnis des Nutzers Rechnung getragen und sie im Vergleich zu anderen Treffern als nicht so relevant bewertet. Jedes Ranking ist aber nur „eine von vielen möglichen algorithmischen Sichten auf die Inhalte […] und beruht grundsätzlich auf Annahmen."[10]

Aber wie funktioniert das eigentlich? Was sind die Retrieval- und Bewertungsmechanismen einer Suchmaschine? Meist ist das Retrieval und das Scoring von Suchmaschinen für Nutzerinnen und Nutzer intransparent.[11] Im Falle von Websuchmaschinen wie Google weiß man eventuell, dass in den Dokumenten die Wortgewichtung oder bei mehreren Suchbegriffen der Wortabstand eine Rolle spielen, dass die Anzahl der Links, die auf das Dokument verweisen, einen Einfluss hat oder aber auch, dass wissenschaftliche Seiten höher gerankt werden, wenn man sich in einem akademischen Kontext befindet.[12]

Google selbst liefert zwar auch einige Informationen zu seinen Suchalgorithmen:

> *„Diesen Rankingsystemen liegt eine ganze Reihe von Algorithmen zugrunde. Damit du die nützlichsten Informationen erhältst, wird eine Vielzahl von Faktoren herangezogen. Dazu gehören unter anderem die in deiner Suchanfrage verwendeten Wörter, die Relevanz und Nützlichkeit von Seiten, die Sachkenntnis von Quellen sowie dein Standort und deine Einstellungen. Die Gewichtung der einzelnen Faktoren hängt von der Art deiner Suchanfrage ab. Zum Beispiel spielt die Aktualität der Inhalte bei der Beantwortung von Fragen zu aktuellen Themen eine größere Rolle als bei Wörterbuchdefinitionen."*[13]

Dennoch bleiben Suchmaschinen und ihre Verfahren meist „Black-Boxes". In Bezug auf die Suchmaschine und die Verfahren, die im *prometheus*-Bildarchiv Anwendung finden, soll nun ein wenig Licht ins Dunkel gebracht werden. Als Suchmaschine verwendet *prometheus*

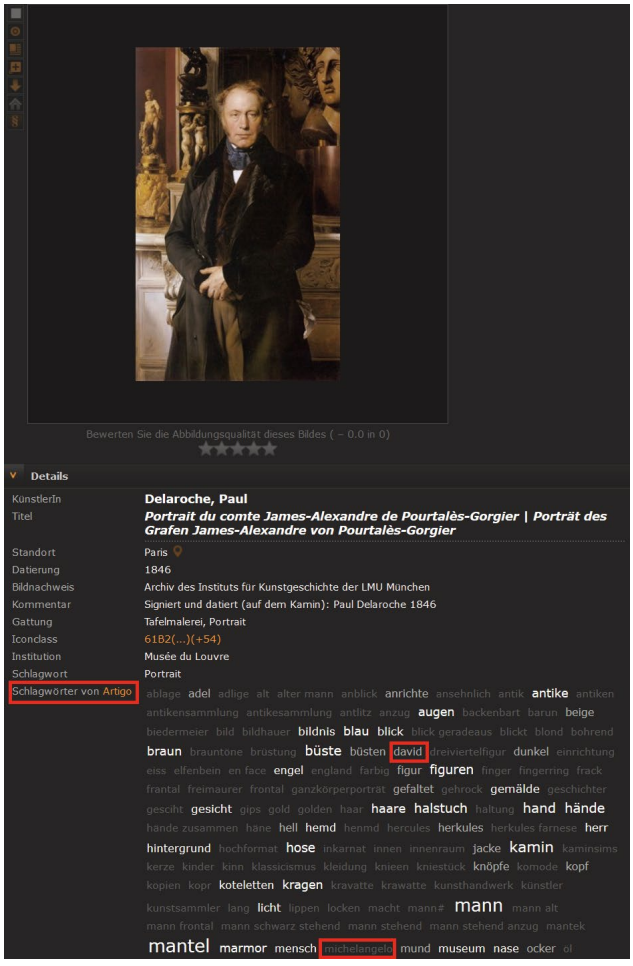
1

2

Abb. 1 Ergebnisliste *prometheus*-Bildarchiv: Datensatz zu Honoré Victorin Daumier: *La Laveuse*, <https://prometheus.uni-koeln.de/de/image/daumier-4cbec45e780cf1cc89854fe5f3d15b995f17dc8a> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang).

Abb. 2 Ergebnisliste *prometheus*-Bildarchiv: Datensatz zu Paul Delaroche: *Portrait du comte James-Alexandre de Pourtalès-Gorgier*, <https://prometheus.uni-koeln.de/de/image/artemis-92b97a4067caf7e2c805f3e8bac57b2128ofb9b4> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang).

3



4

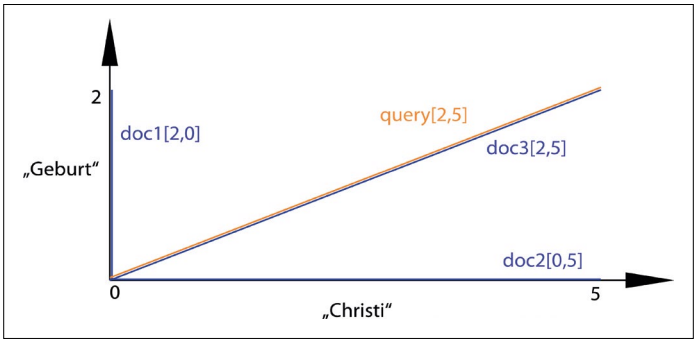


Abb. 3 Vollansicht des Datensatzes zu Paul Delaroche: *Portrait du comte James-Alexandre de Pourtalès-Gorgier*, <https://prometheus.uni-koeln.de/de/image/artemis-92b97a4067caf7e2c805f3e8bac57b21280fb9b4>, (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang) [Markierungen im Bild von der Verfasserin].
Abb. 4 Abfragevektor und Dokumentvektoren im Verhältnis.

„Elasticsearch", eine seit 2010 existierende Suchmaschine auf Basis von „Lucene", die ihre Funktionalität per HTTP-Schnittstelle zur Verfügung stellt mit einem besonderen Fokus auf Datenverteilung und Clustering. Grundlage ist der zentrale invertierte Index, der die Suchbegriffe auf Dokumente abbildet. Zu jedem Term ist die Information gespeichert, in welchen Dokumenten er auftaucht und in welchen nicht.[14] Da das sogenannte „Boolsche Retrieval" die Terme nur nach Vorkommen bzw. Nicht-Vorkommen beurteilt, werden noch zusätzliche Faktoren berücksichtigt und zu den Termen abgespeichert, die eine Gewichtung der Terme – abhängig von dem jeweiligen Dokument – ermöglichen. „Volltext-Relevanz-Formeln" kombinieren verschiedene Faktoren, um einen Relevanz-Score für jedes Dokument zu ermitteln.[15] Die drei Faktoren „term frequency" (TF), „inverse document frequency" (IDF) und „field-length norm" werden während der Indexierung berechnet und gespeichert. Bei der „term frequency" wird ein Wert aus der Häufigkeit des Auftretens des Terms in einem Dokument ermittelt, bei der „inverse document frequency" die relative Häufigkeit des Vorkommens in allen Dokumenten des Index. Das Gewicht ist dabei höher, je seltener der Begriff vorkommt. Bei der „field-length norm" wird der Begriff im Verhältnis zur Feldlänge ausgewertet. Ist das Feld kürzer, dann ist das Gewicht höher, denn die Wahrscheinlichkeit steigt, dass es z. B. in einem knappen Titel-Feld tatsächlich um diesen Begriff geht, wohingegen bei einem langen Beschreibungsfeld der Begriff auch keine inhaltliche Bedeutung haben könnte. Zusammen werden diese drei Faktoren verwendet, um die Gewichtung eines einzelnen Begriffs in einem bestimmten Dokument zu berechnen, um diese dann nach Relevanz zu sortieren. Um darüber hinaus selbst zu bestimmen, welches Feld als relevanter zu betrachten ist als andere, kann auch ein „Index-Field-Boost" eingesetzt werden. In *prometheus* haben wir einen geringfügigen Boost[16] für das Titel-Feld hinzugefügt, weil eine Anfrage z. B. nach „Geburt Christi" wohl eher das Titel-Feld betrifft als beispielsweise den Bildnachweis. Das ist allerdings nur eine Annahme, die unter Umständen auch Auswirkungen auf „Precision" und „Recall" hat. Um die Gewichtungen mehrerer Begriffe zu kombinieren, wird das Vektorraummodell verwendet. Hierbei kann eine Multiterm-Abfrage mit einem Dokument verglichen werden. Dabei wird für die Suchanfrage (Query) anhand der Gewichtung der einzelnen Begriffe ein Abfragevektor und für die einzelnen Dokumente ein Dokumentvektor erstellt auf Basis der Relevanz der Begriffe innerhalb des Dokuments (s. Abb. 4). Durch die Darstellung als Vektoren können Winkel berechnet werden zwischen dem Abfragevektor und dem Dokumentvektor. Je kleiner der Winkel, desto relevanter das Dokument.[17] In dem vorliegenden Beispiel mit einer Suchanfrage „Geburt Christi" nehmen wir einfach an, dass „Geburt" eine Gewichtung von 2 hätte und „Christi" eine Gewichtung von 5.[18] Folgende Dokumente stünden zur Auswahl:

- Dokument 1: Geburt Mariae [2,0]
- Dokument 2: Christi Himmelfahrt [0,5]
- Dokument 3: Geburt Christi [2,5]

Das Dokument 3 hätte somit die höchste Relevanz, weil es dem Query am nächsten kommt.

Bevor die Suchmaschine jedoch aufgrund bestimmter Faktoren in der Lage ist, die Dokumente zur eingegebenen Suchanfrage zu ranken, werden die Metadaten der Datenbanken vorverarbeitet, d. h. sie werden für das Retrieval homogenisiert und auch angereichert. Das bedeutet nicht, dass die Daten selbst abgeändert werden, jedoch werden bei der Indexierung verschiedene linguistische Analyseverfahren angewendet und die Daten angereichert, um die Heterogenität der Daten auszugleichen. Dieser Prozess geschieht für die Nutzerinnen und Nutzer unbemerkt und hat Auswirkungen auf „Recall" und „Precision". „Die Suche nach den zu einem Informationswunsch relevanten Dokumenten berücksichtigt dabei die Vagheit und Unvollständigkeit, die sowohl bei der Formulierung des Informationswunsches als auch bei der – ggf. automatischen – Interpretation des Inhalts der betrachteten Dokumente besteht."[19] Welche Verfahren verwendet werden, wird in der Konfiguration der zu verwendenden Filter festgelegt. Bestimmte Normalisierungen sind wir als Nutzerinnen und Nutzer gewohnt und setzen sie deswegen auch bei der Eingabe der Suchbegriffe voraus. Ob Wörter groß oder klein geschrieben werden, ist egal, denn standardmäßig werden alle Wörter mit dem „lowercase-Filter"[20] im Analyseprozess der Suchmaschine in Kleinbuchstaben umgewandelt, sowohl bei der Indexierung der Dokumente als auch bei der Anfrage. Der eingesetzte „Tokenizer" splittet alle Sätze und Anfragen in einzelne Wörter auf – auch das setzen wir bei einer Suchanfrage nach „Raffael Madonna" bereits voraus. Würden die beiden Suchterme als zusammengehörig abgefragt, wäre die Trefferausbeute wohl eher gering. Darüber hinaus werden Umlaut-Filter eingesetzt, so dass es keine Rolle spielt, ob man „Grundriss" oder „Grundriß", oder „Äpfel" oder „Aepfel" sucht. Auch diakritische Zeichen werden normalisiert, so dass „Cézanne" auch gefunden wird, wenn die Nutzerinnen und Nutzer den Künstler ohne Accent aigu eingeben.[21] Außerdem wird der „word_delimiter-Token-Filter"[22] eingesetzt, der zusammengesetzte Wörter in Subwörter

aufteilt. Wenn man z. B. nach „Museum Köln“ sucht, werden auch alle Museen gefunden, die mit Bindestrich geschrieben werden, wie z. B. das „Wallraf-Richartz-Museum“ oder das „Rautenstrauch-Joest-Museum“. Negative Auswirkungen hat die Anwendung aller Filter auf die „Precision“. Die Anwendung des „word_delimiter-Filters“ führte in dem ersten Beispiel „Daumier: *La laveuse*“ – wie oben beschrieben – dazu, dass dieser Treffer als relevant eingeschätzt wurde, weil der Term „David-Weill“ in zwei Wörter getrennt im Index vorliegt.

Darüber hinaus kann man auch sprachspezifische Verarbeitung durch „Stemming“ einsetzen, um eine größere Variation in der Schreibweise zu ermöglichen. Dabei werden die Begriffe auf die Normalform gebracht, so dass eine Suche nach „Baum“ auch Treffer mit dem Begriff „Bäume“ findet. Der „Recall“ wird in diesem Fall erhöht, die „Precision“ jedoch möglicherweise verringert. Darüber hinaus kann es bei diesem Prozess auch zu falschen Wortbildungen kommen und damit auch zu Treffern, die nicht den Erwartungen entsprechen. Standardmäßig werden deshalb nicht so aggressive „Stemmer“ eingesetzt.[23] Auch bei *prometheus* ist der „Light-German-Filter“ aktiv, der auf den morphologischen Regeln der deutschen Sprache basiert. Als „Light-Stemmer“ werden solche Stemmer bezeichnet, die nur morphologische Flexionen entfernen.[24]

Es gibt noch andere Möglichkeiten, die Daten so anzureichern, dass die Vagheit oder Unsicherheit der Suchanfrage ausgeglichen wird. So kann man einen Phonetischen „Token-Filter“ verwenden, der die Aussprache eines Terms (Meyer, Meier) kodiert.[25] Wir haben uns gegen den Einsatz dieses Filters entschieden, da gerade bei den gezielten Suchen nach Künstlernamen die Eindeutigkeit wichtig ist, denn der „Recall“ würde zwar zunehmen, die „Precision“ sich aber stark verschlechtern. Ein Argument für die Einbindung des Filters wäre die Unsicherheit über die Schreibweise eines Namens. Doch dazu gibt es die Möglichkeit, eine unscharfe Suche, die mithilfe der „Levenshtein-Distanz“[26] als Ähnlichkeitsmaß ermittelt wird, durchzuführen, indem man dem Suchbegriff eine Tilde hinzufügt. „Gaugin~“ findet auch alle Datensätze mit der korrekten Schreibweise „Gauguin“. Das heißt, die Nutzerinnen und Nutzer entscheiden selbst, wann Präzision und wann Unschärfe gewünscht ist. Auch haben wir keine Stoppwörter-Liste integriert, bei der sprachspezifische einzelne, häufig vorkommende Wörter entfernt werden, wie Artikel oder Konjunktionen (z. B. „der“, „eine“, „und“). Gerade bei kurzen Titeln kann der Artikel für das Kunstwerk identifizierend sein, wie z. B. für „*Der Schrei*“.

Neben dem Einsatz von Filtern, die durch verschiedene linguistische Analyseverfahren die Daten selbst vorverarbeiten, gibt es in Elasticsearch die Möglichkeit, die Daten mit Synonymen anzureichern. In *prometheus* werden neben der Integration eines englischen Wörterbuchs auch den Künstlerdaten bei der Indexierung Synonyme hinzugefügt, um so variierende Schreibweisen auszugleichen.[27] Hierzu wird die „prometheus Künstler-Namensansetzungs-Datei“ (PKND)[28] verwendet, die neben der Ansetzungsform für einen Künstler auch alle Varianten enthält. Folgende Ansetzungsform „Laer, Pieter van“ wird angereichert durch die im <sub>-Element stehenden Varianten

```
<term>
  <name>Laer, Pieter van</name>
    <sub>
      <name>Laer, Pieter Jacobsz. van</name>
      <name>Lara, Pieter Jacobsz. van</name>
      <name>Bamboccio</name>
      <name>Bamboots</name>
    </sub>
</term>
```

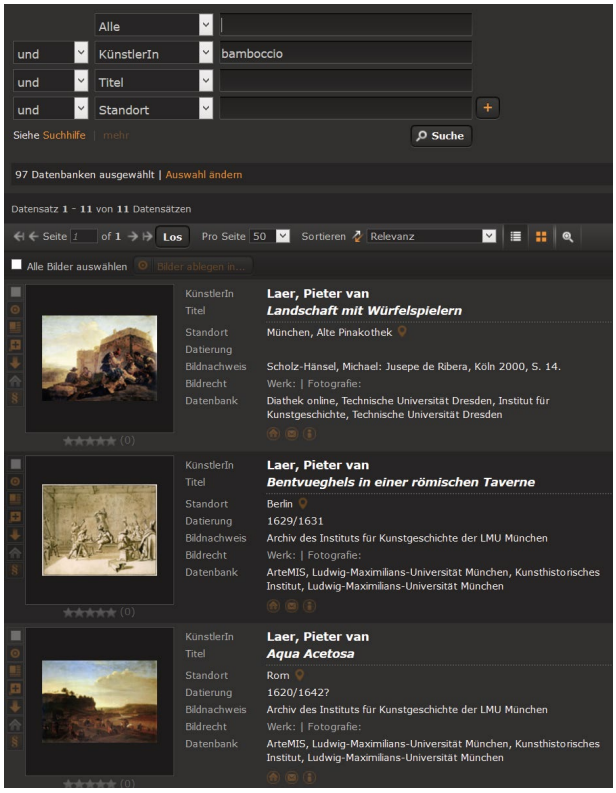
Bei der Suche nach dem Pseudonym „Bamboccio“ findet man die relevanten Datensätze, auch wenn in den Datensätzen der Name selbst nicht auftaucht (s. Abb. 5).

Auch für Anreicherung der Metadaten durch Wörterbücher und Thesauri gilt, dass sich der „Recall“ deutlich verbessert, gleichzeitig aber mit einer verschlechterten „Precision“ einhergehen kann, indem auch irrelevante Datensätze zurückgeliefert werden. Die Namensanreicherung hat etwa z. B. zur Folge, dass bei einer Suche nach „Peter von Cornelius“ auch Abbildungen von Raffael auftauchen, da Peter von Cornelius' Pseudonym „Raphael“ war und daher auch in der PKND als Variante auftaucht. Der Anreicherung mit englischen Begriffen beziehungsweise der grundsätzlichen Mehrdeutigkeit des Wortes ist es geschuldet, dass bei der Sucheingabe „see“ auch Datensätze mit „lake, sehen, sea, meer“ als Ergebnisse auftauchen (s. Abb. 6).

Aber auch hier haben Nutzerinnen und Nutzer Instrumente, um Datensätze auszuschließen, welche nicht ihrer Anfrage entsprechen, da die Suchbegriffe mit „Boolschen Operatoren“ verknüpft werden können. Wenn man zum Beispiel „den See“ meint und nicht „die See“, kann man durch die Suchanfrage „see“ AND NOT „meer“ die Treffermenge reduzieren. Dazu muss natürlich eine gewisse Transparenz über die verwendeten Features gegeben sein und die Nutzerinnen und Nutzer müssen die Funktionalitäten kennen und die Suche beherrschen.

Die genannten Beispiele sind aber tatsächlich nur Extrembeispiele, die aufzeigen sollen, wie sensibel man

5



6

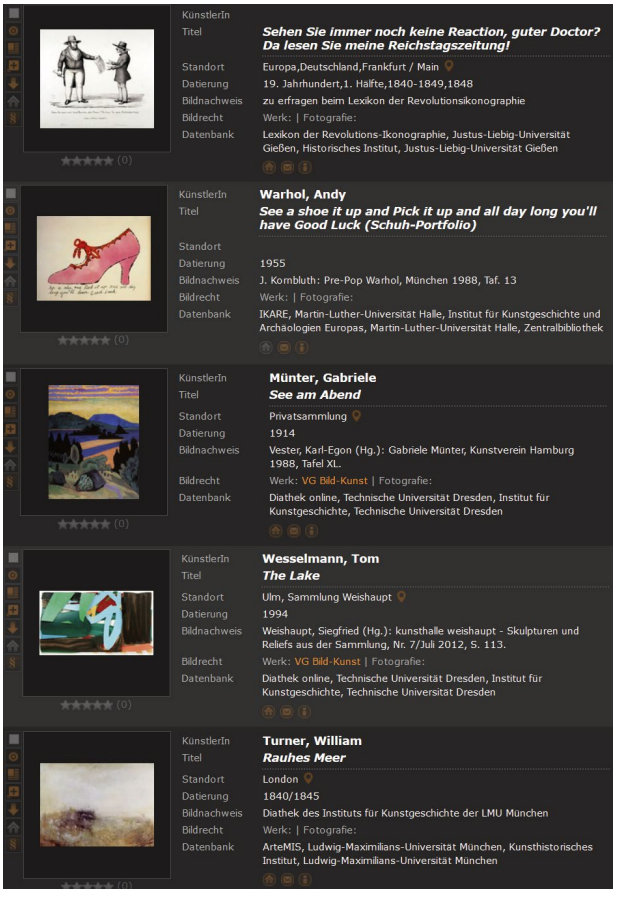


Abb. 5 Ergebnisliste bei Suchanfrage „Bamboccio“, <https://prometheus.uni-koeln.de/de/searches?utf8=%E2%9C%93&search_value%5B0%5D=bamboccio&commit=Daten+absenden> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang).

Abb. 6 Zusammengestellte Beispiele bei Suchanfrage „see“, <https://prometheus.uni-koeln.de/de/searches?utf8=%E2%9C%93&search_value%5B0%5D=see&commit=Daten+absenden> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang)

mit dem Retrieval umgehen muss. Grundsätzlich ist die Integration der Vokabulare in den meisten Fällen sehr nützlich (siehe oben). Hinzukommt, dass die Nutzerschaft von *prometheus* nicht nur deutschsprachig ist und vor allem auch viele Datenbanken integriert sind, die Originaltitel verwenden oder vorwiegend englische Metadaten bereitstellen. Ein großer Teil der relevanten Datensätze würde so gar nicht gefunden.

Der Beitrag sollte einen Einblick geben in die Retrieval- und Bewertungsmechanismen von Suchmaschinen am Beispiel des *prometheus*-Bildarchivs. So sollte insbesondere herausgestellt werden, dass bei der Verwendung von Suchmaschinen der Einsatz von speziellen Filtern oder Synonymwörterbüchern zum Ausgleich von Schreibweisen Ermessenssache bezüglich eines Austarierens zwischen „Precision" und „Recall" ist, damit das Ziel erreicht werden kann, den „Recall" zu verbessern, so dass Nutzerinnen und Nutzer mehr relevante Treffer erhalten, jedoch unter der Prämisse, die „Precision" nicht zu sehr zu verschlechtern. Um das „Retrieval" weiter zu verbessern, soll zukünftig z. B. auch der „Getty Geographic Names-Thesaurus" (TGN)[29] integriert werden. Da die Erfassung der Orte bei den integrierten Datenbanken sehr heterogen ist, werden alle Schreibweisen eines Ortes den Datensätzen bei der Indexierung hinzugefügt. Darüber hinaus werden derzeit alle Datierungsdaten innerhalb von *prometheus* ausgewertet, um auch eine zeitraumübergreifende Suche zu ermöglichen. Die Herausforderung besteht darin, jegliche Datierungsformate und auch natürlichsprachliche Datierungen zu erfassen und in einen Thesaurus zu überführen, der die Datierungen in nummerische Werte umwandeln kann und gleichzeitig Unschärfe berücksichtigt. Um die Präzision z. B. bei den Künstlernamen/Werken grundsätzlich zu erhöhen, sollen die Datensätze zukünftig automatisch annotiert und auch mit Normdaten (GND, Wikidata, VIAF) angereichert werden, so dass eine Unterscheidung eindeutig möglich ist.[30] Als komplementärer Ansatz zur textbasierten Suche werden darüber hinaus zukünftig auch bildbasierte Verfahren Anwendung finden, die die Bilder visuell analysieren, damit auch mithilfe von bildlichen Informationen nach Bildern und Bildpartien gesucht werden kann.[31]

Abbildungsnachweis

Abb. 1 Screenshot aus dem *prometheus*-Bildarchiv: <https://prometheus.uni-koeln.de/de/image/daumier-4cbec45e780cf1cc89854fe5f3d15b995f17dc8a> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang), © Daumier Register, Ascona.
Abb. 2 Screenshot aus dem *prometheus*-Bildarchiv: <https://prometheus.uni-koeln.de/de/image/artemis-92b97a4067caf7e2c805f3e8bac57b21280fb9b4> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang), © Artemis, LMU München.
Abb. 3 Screenshot aus dem *prometheus*-Bildarchiv: <https://prometheus.uni-koeln.de/de/image/artemis-92b97a4067caf7e2c805f3e8bac57b21280fb9b4> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang), © Artemis, LMU München.
Abb. 4 eigene Grafik.
Abb. 5 Screenshot aus dem *prometheus*-Bildarchiv: <https://prometheus.uni-koeln.de/de/searches?utf8=%E2%9C%93&search_value%5B0%5D=bamboccio&commit=Daten+absenden> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang), © Diathek online, TU Dresden; Artemis, LMU München.
Abb. 6 Screenshot aus dem *prometheus*-Bildarchiv: <https://prometheus.uni-koeln.de/de/searches?utf8=%E2%9C%93&search_value%5B0%5D=see&commit=Daten+absenden> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang), © Lexikon der Revolutionsikonographie, JLU Gießen; IKARE, MLU Halle; Diathek online, TU Dresden; Artemis, LMU München.

Anmerkungen

1 Zur Einführung ins Information Retrieval vgl. Andreas Henrich: *Information Retrieval 1: Grundlagen, Modelle und Anwendungen* (Bamberg: Otto-Friedrich-Universität Bamberg 2009), Gobinda G. Chowdhury: *Introduction to Modern Information Retrieval*, 3. Aufl. (London: Facet Publishing 2010), Helmut Jarosch: *Information Retrieval und Künstliche Intelligenz* (Wiesbaden: DUV 2007) <https://doi.org/10.1007/978-3-8350-9444-4> und Harald Klinke, „Information Retrieval", in: Fotis Jannidis/Hubertus Kohle/Malte Rehbein (Hrsg.): *Digital Humanities: Eine Einführung* (Stuttgart: J.B. Metzler 2017), 268–78 <https://doi.org/10.1007/978-3-476-05446-3_19>, alle Links abgerufen am 18.10.2019.

2 Henrich: *Information Retrieval* [wie Anm. 1], 26.

3 https://prometheus-bildarchiv.de. *prometheus* ist ein verteiltes digitales Bildarchiv für Forschung und Lehre in den Kunst- und Kulturwissenschaften und benachbarten bildbasierten Disziplinen. Als Datenbroker verknüpft *prometheus* derzeit 99 Datenbanken von Universitäten, Forschungsinstitutionen, Museen sowie Archiven miteinander und macht so über 2 Millionen hochauflösende Bilder über die Plattform recherchierbar (Stand: Oktober 2019). Die integrierten Bilder decken mit Material aus der Kunstgeschichte, Archäologie, Pädagogik, Geschichte, Theologie, Design- und Architekturgeschichte, Ägyptologie, Umweltgeschichte, Ethnologie und eben auch der Diplomatik ein breites inhaltliches Spektrum ab. Als Suchmaschine findet „Elasticsearch" Anwendung, die auf der freien Software zur Volltextsuche „Lucene" basiert.

4 Eine interessante Studie zur Verbesserung des Retrievals bei Bildersuchmaschinen am Beispiel der *Deutschen Digitalen Bibliothek* (DDB) liefert Jutta Lindenthal: „Datenqualität und Retrieval - Vorschläge zur Verbesserung der Suche in der Deutschen Digitalen Bibliothek" (2016). <http://jl.balilabs.de/DDB/DQ/DDB_Datenqualit%C3%A4t_Retrieval_1.0.pdf>.

5 Zur Qualität von Information-Retrieval-Systemen vgl. Manfred Thaller: „Bemerkungen zu kunsthistorischen Informationssystemen; vornehmlich aus der Sicht der Informatik", in: *zeitenblicke* 2, Nr. 1 (2003), §16 <http://www.zeitenblicke.historicum.net/2003/01/thaller/index.html>.

6 <https://prometheus.uni-koeln.de/de/image/daumier-4cbec45e780cf1cc89854fe5f3d15b995f17dc8a> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang)

7 <https://prometheus.uni-koeln.de/de/image/artemis-92b97a4067caf7e2c805f3e8bac57b21280fb9b4> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang)

8 Bei dem Spiel *Artigo*, das von Hubertus Kohle mitentwickelt wurde, werden mittels „Crowd Sourcing" „Tags" zu Abbildungen der Kunstgeschichte erzeugt. Einen genaueren Einblick in die Funktionalität und die Spielregeln liefert: <http://www.artigo.org>.

9 <https://prometheus.uni-koeln.de/de/image/artemis-20fe758b8058d6d68ba11e6bba9a7c2d88a2475d> (Zugriff nur über lizenzierten Campus bzw. persönlichen Zugang)

10 Dirk Lewandowski: *Suchmaschinen verstehen* (Berlin und Heidelberg: Springer 2018), 93f. <https://doi.org/10.1007/978-3-662-56411-0_5>.

11 Häufig wird Suchmaschinen und den Rankingverfahren deshalb auch eine gewisse Skepsis entgegengebracht: siehe z. B. <https://digitalcourage.de/digitale-selbstverteidigung/es-geht-auch-ohne-google-alternative-suchmaschinen>; <https://crackedlabs.org/studie-kommerzielle-ueberwachung>.

12 Näheres zu anfrageabhängigen und anfrageunabhängigen Ranking-Faktoren bei Websuchmaschinen bei Christian Maaß/Gernot Gräfe: „Suchmaschinen und Informationsqualität: Status quo, Problemfelder, Entwicklungstendenzen", in: Knut Hildebrand/Marcus Gebauer/Holger Hinrichs/Michael Mielke (Hrsg.): *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence* (Wiesbaden: Springer 2018), 158/160 <https://doi.org/10.1007/978-3-658-21994-9_9> und zum Ranking allgemein vgl. Lewandowski: *Suchmaschinen* [wie Anm. 10], 93–129.

13 <https://www.google.com/intl/de/search/howsearchworks/algorithms/>.

14 Einen guten Einstieg in die Funktionalitäten von „Elasticsearch" liefert Florian Hopf: *Elasticsearch: Ein praktischer Einstieg* (Heidelberg: dpunkt 2016). <http://ebookcentral.proquest.com/lib/ubkoeln/detail.action?docID=4347081>.

15 <https://www.elastic.co/guide/en/elasticsearch/guide/current/controlling-relevance.html>.

16 Vgl. Hopf, *Elasticsearch* [wie Anm. 14], 66f.

17 Zum Scoring, den Faktoren zur Berechnung der Gewichtung eines Begriffs und zum Vektorraummodell: <https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html>.

18 Die Grafik ist angelehnt an das Beispiel in der Elasticsearch-Dokumentation: <https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html#img-vector-docs>.

19 Henrich: *Information Retrieval* [wie Anm. 1], 14–15.

20 <https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-lowercase-tokenfilter.html>.

21 <https://www.elastic.co/guide/en/elasticsearch/guide/current/asciifolding-token-filter.html>.

22 <https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-word-delimiter-tokenfilter.html>.

23 Vgl. Hopf, *Elasticsearch* [wie Anm. 14] 41.

24 Mehr zu Light-Stemming-Ansätzen bei Jacques Savoy: „Light Stemming Approaches for the French, Portuguese, German and Hungarian Languages", in: *Proceedings of the 2006 ACM Symposium on Applied Computing*, SAC '06 (New York, ACM 2006), 1031–1035 <https://doi.org/10.1145/1141277.1141523>.

25 Vgl. Hopf, *Elasticsearch* [wie Anm. 14] 49.

26 Vgl. Andrew Cholakian: „How to Use Fuzzy Searches in Elasticsearch", 2013, <https://www.elastic.co/de/blog/found-fuzzy-search>.

27 Der Aufbau der Indexierungswörterbücher erfolgte in einem von der RheinEnergieStiftung von 2007 bis 2009 durchgeführten Projekt: <https://prometheus-bildarchiv.de/de/projects/perseus-a>.

28 <https://prometheus-bildarchiv.de/tools/pknd>.

29 <http://www.getty.edu/research/tools/vocabularies/tgn/index.html>.

30 Das Projekt wird an der Universität zu Köln in enger Zusammenarbeit zwischen Fachwissenschaftlern der Kunstgeschichte und der Sprachlichen Informationsverarbeitung durchgeführt. Die Schwerpunkte der Sprachlichen Informationsverarbeitung liegen unter anderem auf Systemen zur syntaktischen und semantischen Analyse und Verarbeitung textueller Daten. Eine Beschreibung des Vorhabens findet sich unter: Lisa Dieckmann /Jürgen Hermes / Claes Neuefeind: „Bild, Beschreibung, (Meta)Text. Automatische inhaltliche Erschließung und Annotation kunsthistorischer Daten", in: *Abstractband der Jahrestagung des Verbandes für Digital Humanities im deutschsprachigen Raum* (Universität Bern 2017), 105–107 <http://www.dhd2017.ch/wp-content/uploads/2017/02/Abstractband_ergaenzt.pdf>, abgerufen am 15.10.2019.

31 Ein Pilotprojekt wurde bereits von der Computer Vision Group an der Ruprecht-Karls-Universität Heidelberg mit Bilddaten des *prometheus*-Bildarchivs durchgeführt. Die Computer Vision Group des Heidelberg Collaboratory for Image Processing widmet sich der Grundlagenforschung zum automatischen Bildverstehen. Sie entwickelt Algorithmen zur Erschließung von Bildbestandteilen (Segmentierung), diskriminativer Objekt-

erkennung und Szenenvergleich. Sie hat bereits in mehreren Projekten mit der Kunstgeschichte zusammengearbeitet. Siehe auch <http://hci.iwr.uni-heidelberg.de/COMPVIS/>. Für einen tiefergehenden Einblick bereits durchgeführter Projekte, in Methoden und technische Verfahren siehe z. B. Peter Bell/Lisa Dieckmann: „Die Kunst als Ganzes. Heterogene Bilddatensätze als Herausforderung für die Kunstgeschichte und die Computer Vision", in: *DHd 2016. Modellierung – Vernetzung – Visualisierung. Die Digital Humanities als fächerübergreifendes Forschungsparadigma. Konferenzabstracts* (Leipzig 2016), 118–120 <http://dhd2016.de/boa.pdf> und Peter Bell/Björn Ommer: „Training Argus, Ansätze zum automatischen Sehen in der Kunstgeschichte", in: *Kunstchronik 68* (2015), 414–420. Siehe auch Vipin Tyagi: „Content-Based Image Retrieval: An Introduction", in: Vipin Tyagi (Hrsg.): *Content-Based Image Retrieval: Ideas, Influences, and Current Trends* (Singapore: Springer 2017) <https://doi.org/10.1007/978-981-10-6759-4_1>.