Anreicherung von bibliographischen Metadaten zur Sichtbarmachung zurückgezogener Artikel. Erarbeitung am Beispiel des Suchportals LIVIVO und der "Retraction Watch Database".

Bachelorarbeit zur Erlangung des Bachelor-Grades
Bachelor of Science im Studiengang Data and Information Science
an der Fakultät für Informations- und Kommunikationswissenschaften
der Technischen Hochschule Köln

vorgelegt von: Lucas Vetter

eingereicht bei: Prof. Dr. Konrad Förstner
Zweitgutachter/in: Dr. Eva Seidlmayer

Köln, 28.02.2023

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer oder der Verfasserin/des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Ort, Datum

Rechtsverbindliche Unterschrift

Kurzfassung/Abstract

Die Zahl der wissenschaftlichen Publikationen steigt von Jahr zu Jahr. Der technologische Fortschritt hat es den Forschenden erleichtert, Daten zu sammeln, zu analysieren und zu verarbeiten. Aufgrund von Faktoren wie dem starken Wettbewerb um Forschungsgelder, Arbeitsplätze und Anerkennung wird die Anzahl und Reichweite von Publikationen als ein wichtiger Indikator für die Leistung und den Erfolg von Wissenschaftlern angesehen. Die Zunahme von Publikationen kann aber auch zu einer Zunahme von unzureichend überprüften oder nicht reproduzierbaren Ergebnissen führen. Die Integrität der wissenschaftlichen Literatur wird durch das Zurückziehen (engl. Retraction) fehlerhaft veröffentlichter Publikationen gewahrt. Die Kennzeichnung zurückgezogener Artikel in online verfügbaren Quellen ist daher von großer Bedeutung.

In dieser Ausarbeitung wird die im Suchportal *LIVIVO* verfügbare Literatur auf zurückgezogene Artikel untersucht. Retraction Watch, ein amerikanischer Blog, bietet eine Datenbank zurückgezogener Artikel mit fast 40.000 Einträgen. Um die beiden Datensätze miteinander vergleichen zu können, wird eine lokale Datenbankanwendung entwickelt, die einen Abgleich und eine anschließende Analyse ermöglicht. Für die Anreicherung der Metadaten für *LIVIVO* wird empfohlen, den Digital Object Identifier (DOI), das Datum und die *PubMedID*, unter der der zurückgezogene Artikel veröffentlicht wurde, mit aufzunehmen. Der Abgleich gibt einen ersten Eindruck über das Vorhandensein zurückgezogener Artikel im Suchportal. Es wurden 14.206 Einträge identifiziert, wobei in den letzten Jahren ein stetiger Anstieg in der Anzahl zu verzeichnen ist.

Schlagwörter/Schlüsselwörter: Retraction, Anreicherung von Metadaten, Datenbank, *PostgreSQL*, wissenschaftliche Literatur

Inhalt

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

COPE	Committee on Publication Ethics
CSV	Comma-separated values
DOI	Digitaler Objektbezeichner
ETL	Extraktion, Transformation, Laden
ICEE	International Conference on Emerging Electronics
ID	Identifikationsnummer
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MEDLINE	Medical Literature Analysis and Retrieval System Online
PLoS	Public Library of Science
SQL	Structured Query Language
SSH	Secure Socket Shell
URL	Uniform Resource Locator
VM	Virtuelle Maschine
ZB MED	Deutsche Zentralbibliothek für Medizin

1 Einleitung

Zu Beginn erfolgt eine Einführung in die Thematik, wobei die Motivation und die Ziele der Arbeit erläutert werden. Daraus schlussfolgernd werden am Ende dieses Kapitels die Forschungsfragen dargelegt.

1.1 Problemstellung und Motivation

Im Kontext der Wissenschaft und dem wissenschaftlichen Arbeiten bilden verlässliche und valide Quellen und Informationen das Grundgerüst für alle Forschenden. Um in einem Fachgebiet Kompetenzen zu erlangen, oder die eigenen Ausarbeitungen zu untermauern, werden veröffentlichte Artikel gelesen und auf nützliche Erkenntnisse hin untersucht. Diese Erkenntnisse werden weiterverwendet und zitiert. Jedoch kommt es regelmäßig vor, dass Artikel von entsprechenden Prüfgremien als falsch, oder für nicht belastbar befunden werden. Diese werden aus der wissenschaftlichen Community retracted, also nach der Veröffentlichung zurückgezogen und kenntlich gemacht. Da der deutsche Begriff zu dem englischen Wort Retraction (Rückzug, Zurückziehen) im Zusammenhang mit wissenschaftlichen Veröffentlichungen mehrdeutig ist, wird in dem folgenden Text fortführend Retraction und retracted Artikel verwendet. Studien zeigen, dass vor dem Jahr 2000 circa eine Retraction in 100.000 Veröffentlichungen vorkommt. Im Zeitfenster von 2000 bis 2010 waren es jedoch bereits eine Retraction auf 10.000 Veröffentlichungen. (Grieneisen und Zhang 2012) Es ist ein deutlicher Anstieg festzustellen und dieser untermauert die Relevanz der Ausweisung solcher Artikel.

Ein Artikel, welcher veröffentlicht wurde, weitreichende Auswirkungen auf ein wissenschaftliches Feld und die Öffentlichkeit hat, jedoch später retracted wurde, muss als solcher für alle erkennbar sein, um der weiteren Verbreitung von Fehlinformationen entgegenzuwirken. Eines der bekanntesten Beispiele stellt die Arbeit von Wakefield, Murch et al. dar (Wakefield et al. 1998). Sie wiesen in Ihrer Arbeit einen Zusammenhang zwischen Impfungen zu Masern, Mumps und Röteln und Autismus nach. Obwohl dieser Artikel widerlegt und im Jahr 2010 retracted wurde (Lancet 2010), halten viele Menschen diesen immer noch für valide.

Das *Committee on Publication Ethics* (*COPE*) hat Richtlinien verfasst, wann ein veröffentlichter Artikel retracted werden muss. Gründe hierfür können beispielsweise gravierende Fehler bei Kalkulationen, oder Experimenten sein. Betrug und unethische Vorgehensweisen bei der Erstellung können ebenfalls Gründe sein. (Barbour et al. 2009) Die von *COPE* veröffentlichten Richtlinien sehen ein Kenntlichmachen von retracted Artikeln in allen online verfügbaren Versionen (z.B. in Datenbanken) vor. In Zeiten der Digitalisierung und immer weiter zunehmender Informationsvielfalt stehen Anbieter von

Datenbanken und Suchportalen vor einer großen Herausforderung bei der Identifikation und Sichtbarmachung solcher Artikel, damit sich falsche Informationen nicht weiter verbreiten. *Wright* und *McDaid* (Wright und McDaid 2011) haben drei Datenbanken hinsichtlich der Kenntlichmachung von retracted Artikeln untersucht und konnten feststellen, dass bei *MEDLINE* alle Artikel richtig ausgewiesen wurden, bei *CENTRAL* 80%, während *EMBASE* lediglich 6% aller Fälle markiert hat. Diese Unterschiede zeigen gut die Schwierigkeiten der Aktualisierung und Pflege solcher Informationen.

Das *ZB-MED* Suchportal *LIVIVO* möchte die Informationen zu retracted Artikeln abbilden und steht der Herausforderung der Identifikation und Sichtbarmachung dieser Informationen gegenüber.

1.2 Zielsetzung und Aufbau der Arbeit

Im Rahmen dieser Abschlussarbeit sollen die Metadaten von *LIVIVO* um Informationen zu retracted Artikeln ergänzt werden. Als Ausgangspunkt dient dieser Datenkorpus. Das Projekt ist nach der Idee von *Dr. Eva Seidlmayer* (*ZB MED*) entstanden und wird von ihr betreut. Van der Vet und Nijveen (van der Vet und Nijveen 2016) empfehlen die Verwendung des Services von Retraction Watch, um eine möglichst große und valide Aussage über retracted Artikel zu erhalten. Die Organisation hat im Jahr 2018 eine Datenbank herausgebracht und bietet aktuell einen Datenkorpus mit Informationen zu fast 40.000 Dokumenten (Retraction Watch Database 2022) an. Ziel wird es sein, einen Abgleich zwischen den beiden Datenquellen zu machen und die in *LIVIVO* hinterlegten Retractions zu identifizieren.

Zuerst werden Retractions definiert und die relevanten Richtlinien vorgestellt. Mithilfe einer lokalen Datenbank sollen die Informationen konsolidiert werden. Hierfür werden die beiden Datenkorpusse hinsichtlich ihrer Metadaten erschlossen, untersucht und gemeinsame Identifikatoren für eine Zusammenführung herausgearbeitet. Mögliche Erweiterungen an den Metadaten sind, ob ein Artikel retracted ist, wann er retracted wurde und wo die Information zur Retraction zu finden ist. Abschließend wird das Ergebnis analysiert und vorgestellt. Die Ergebnisse und verwendeten Methodiken können in weiteren Projekten zur Anreicherung der Metadaten mit andere Datenquellen genutzt werden.

1.3 Forschungsfragen

Basierend auf der Zielsetzung lassen sich verschiedene Forschungsfragen entwickeln, die es in diesem Kontext zu beantworten gilt.

F1. Mit welchen Methodiken und Anwendungen lässt sich die Anreicherung der Metadaten vornehmen?
F2. Wie groß ist die Schnittmenge zwischen dem Datenset der *LIVIVO* und dem von Retraction Watch?
    a. Welche Erkenntnisse zu Retractions innerhalb des *LIVIVO* Datensets sind erkennbar?
    b. Lassen sich Trends, oder Muster erkennen?
F3. Welche Inhalte sollten bei der Erweiterung der Metadaten berücksichtigt und ergänzt werden?

2 Retracted Artikel

In diesem Abschnitt erfolgt eine kurze Vorstellung darüber was Retractions sind und die dazugehörigen Richtlinien.

2.1 Der Grund für Retractions in der Wissenschaft

Retraction ist ein Verfahren zur Korrektur wissenschaftlicher Veröffentlichungen und zur Warnung von Lesern vor Artikeln, die derart gravierende Mängel oder fehlerhafte Inhalte oder Daten aufweisen, dass man sich nicht auf deren Ergebnisse und Schlussfolgerungen verlassen kann. Fehlerhafte Inhalte oder Daten können aus ehrlichen Irrtümern, unbedarften Fehlern oder sogar Forschungsvergehen resultieren. Der Hauptzweck einer Retraction besteht darin, die Integrität der wissenschaftlichen Literatur zu gewährleisten, anstatt die Autoren zu bestrafen. Eine Retraktion kann erforderlich sein, um Leser auf verschiedene Verstöße aufmerksam zu machen (Barbour et al. 2009).

2.2 Richtlinien

Die Richtlinien des *Committee on Publication Ethics (COPE)* schreiben vor, wann eine Retraction angebracht ist, was in der veröffentlichten Retraction-Notiz enthalten sein sollte, wer diese herausgeben sollte und wie mit Artikeln umgegangen werden sollte, zu denen nicht hinreichende Beweise für eine Retraction vorliegend sind. (Barbour et al. 2009)

Die Herausgeberinnen und Herausgeber sollten unter den folgenden Bedingungen eine Retraction in Erwägung ziehen. Es gibt klare Hinweise darauf, dass die Ergebnisse unzuverlässig sind, entweder aufgrund eines schwerwiegenden Fehlers (z.B. Rechenfehler oder experimenteller Fehler) oder aufgrund von Fälschung (z.B. von Daten) oder Verfälschung (z.B. Bildmanipulation). Der Beitrag ist ein Plagiat oder enthält Material oder Daten, deren Verwendung nicht genehmigt wurde. Jeder Verstoß gegen das Urheberrecht führt zu einer Retraction (Barbour et al. 2009).

Ethische Kriterien können ebenfalls eine Retraction nach sich ziehen. Sollte der Artikel über unethische Forschung berichten sollte dieser aus der wissenschaftlichen Gemeinschaft entfernt werden. Es wurde ausschließlich auf der Grundlage eines beeinträchtigten oder manipulierten Peer-Review-Prozesses veröffentlicht. Der Autor oder die Autorin haben ein wichtiges konkurrierendes Interesse (Interessenkonflikt) nicht offengelegt, das nach Ansicht des Herausgebers die Interpretation der Arbeit oder die Empfehlungen der Herausgeber und Peer-Reviewern ungerechtfertigt beeinflusst hätte (Barbour et al. 2009).

Die veröffentlichte Notiz zu der Retraction sollte stets auf den Originalartikel referenzieren. Online ist eine Verlinkung der Beiden angebracht. Bei einer solchen Notiz ist eine eineindeutige Zuordnung entscheidend, diese kann durch Angabe des Titels und der Autoren in der Überschrift kenntlich gemacht werden. Retractions müssen klar erkennbar sein und sich von Korrekturen oder Kommentaren unterscheiden lassen, ebenfalls müssen diese für alle frei zugänglich gemacht werden. Sie dürfen nicht hinter Abonnements, oder anderen Zugangsbeschränkungen liegen. Je schneller eine Retraction erfolgt, desto höher ist die Chance Schaden zu minimieren. Die Notiz sollte inhaltlich die Angaben umfassen, wer den Artikel zurückzieht und die Begründung, die dazu führt. Retractions sollten objektiv, sachlich und frei von aufputschender Sprache sein (Barbour et al. 2009).

Unter den folgenden Aspekten sind Retractions oftmals nicht angebracht bzw. können anders abgebildet werden. Die Autorenschaft wird angezweifelt, aber es gibt keinen Grund, die Gültigkeit der Ergebnisse in Frage zu stellen. Die wichtigsten Ergebnisse der Arbeit sind immer noch zuverlässig und Korrekturen könnten Fehler oder Bedenken ausreichend beheben. Ein Herausgeber hat unklare Beweise, um die Retraction zu unterstützen oder wartet auf zusätzliche Informationen, z.B. aus einer institutionellen Untersuchung. Ein Interessenkonflikt wurde dem Journal nach der Veröffentlichung gemeldet, aber aus Sicht des Herausgebers ist es unwahrscheinlich, dass diese die Interpretationen oder Empfehlungen oder Schlussfolgerungen des Artikels beeinflusst haben (Barbour et al. 2009).

3 Ausgangslage

In diesem Abschnitt werden kurz die beiden Organisationen vorgestellt, welche die Datensets bereitgestellt haben. Anschließend werden die Metadaten der beiden Datensets vorgestellt und grob analysiert.

3.1 ZBMED LIVIVO

ZB MED - Informationszentrum Lebenswissenschaften betreibt *LIVIVO*, ist die größte Suchmaschine für Literatur und Forschungsdaten in den Lebenswissenschaften in Europa. Die Plattform vereint mehr als 67 Millionen Datensätze aus den Bereichen Medizin und Gesundheit, Ernährung, Umwelt und Agrarwissenschaften und unterstützt inter- und transdisziplinäres wissenschaftliches Arbeiten.
LIVIVO bietet Zugang zu einer Vielzahl von Datenquellen, darunter *MEDLINE*, *AGRICOLA* und *BASE*, und verbindet die Bestände von *ZB MED* an seinen Standorten Köln und Bonn. Die Suchmaschine verwendet eine leistungsstarke Suchtechnologie basierend auf *Apache Solr* mit linguistischer Anreicherung und semantischer Verknüpfung von Suchbegriffen.
ZB MED Knowledge Environment, eine Datenbank, die speziell für die Verwaltung und Anreicherung von Informationen in den Lebenswissenschaften entwickelt wurde, ist eng mit *LIVIVO* verbunden. In Zukunft wird *LIVIVO* als Informationsplattform für Forschungsdaten fungieren, Nachweisfunktionen übernehmen und internationale Kooperationen ausbauen. *LIVIVO* entwickelt sich damit zu einem One Health-Portal, das die gesamten Lebenswissenschaften abdeckt (ZB MED 2023).

Der Datensatz der *LIVIVO* mit Stand 30.11.2022 bildet die Basis dieser Untersuchung und soll um ergänzende Metadaten erweitert werden. Schauen wir uns im Folgenden die Metadaten und den Datensatz an.

3.1.1 Beschreibung Metadaten LIVIVO

Der Datensatz der *LIVIVO* enthält 61 Datenformate für die Erstellung der Metadaten, die jeden einzelnen Eintrag mit spezifischen Informationen beschreiben.
Die Datenformate lassen sich thematisch zusammenfassen, sodass zehn Kategorien entstehen. Um den Rahmen dieser Ausarbeitung nicht zu sprengen, werden die Kategorien kurz beschrieben und die wichtigsten Eckdaten benannt.

1. Identifikatoren:

Unter den Identifikatoren werden alle eindeutigen Kennungen zusammengefasst. In Summe sind es zwölf verschiedene Standards, die angelegt werden. Von besonderem Interesse ist der sogenannte Document Object Identifier, kurz DOI. Hierbei handelt es sich um einen eindeutigen, dauerhaften digitalen Identifkator für Objekte, insbesondere für wissenschaftliche Netzpublikationen. Der DOI-Name besteht immer aus einem Präfix

und einem Suffix. Diese werden mit einem Schrägstrich getrennt und das Präfix beginnt immer mit „10.“. Definiert ist dieser nach ISO 26324:2012.

2. Formale Eigenschaften:

Unter diesen Eigenschaften werden Informationen zur liefernden Datenbank, dem Dokumententypen und der Sprache, in der der Eintrag verfasst wurde, zusammengefasst. In dieser Kategorie liegen fünf Ausprägungen vor.

3. Titelangaben:

Alle relevanten Angaben zum Titel sind hier enthalten. Titel plus Untertitel und weitere Varianten, falls diese vorliegen. Bei fremdsprachigen Texten werden die übersetzten Titel mit angegeben. Bei Änderungen des Titels werden diese erfasst und ermöglichen die Suche nach jedem zeitlichen Titel. Ergänzend werden Titelangaben zu Beiwerken mit angegeben. In Summe liegen sieben Ausprägungen vor.

4. Urheberangaben:

Personen, Institutionen, Herausgeber und Events die im Zusammenhang mit Urheberangaben stehen werden aufgeführt. Beispielsweise wird hier der Autor des Schriftstücks genannt, weiterführend unter welcher Institution gearbeitet wurde und ob die Publikation im Rahmen eines Kongresses stattgefunden hat.

5. Inhalt und Beschreibung:

Unter dieser Kategorie ist eine kurze Zusammenfassung der inhaltlichen Kernaussagen für den jeweiligen Eintrag hinterlegt. Als Inhalt wird hierfür meistens das Abstract verwendet.

6. Angaben zum Erscheinen:

Das Land der Publikation bis hin zum Ort werden aufgeführt, sowie der beteiligte Verlag und die zeitlichen Angaben an dem der Eintrag veröffentlicht wurde.

7. Quellenangaben:

Als Quellen werden alle Informationen angegeben, um ein einfaches Auffinden des Eintrags möglich zu machen. Für Zeitschriften Artikel werden beispielsweise der Titel der übergeordneten Zeitschrift, die Auflagenummer und die dazugehörige Seitenzahl angegeben. Eine Zuordnung zu dem jeweils übergeordneten Titel ist hierdurch möglich.

8. Sacherschließung:

Die Einträge werden klassifiziert und mit eindeutigen Bezeichnern ausgestattet. Durch normierte Schlagwörter von liefernden Datenbanken liegen zu den Einträgen MeSH-Schlagworte, Beschreibungen der chemischen Substanzen und weitere Notationen vor. Eine grobe Zuordnung zu fachlichen Einordnungen findet ebenfalls statt. Normierte mögliche Themengebiete können Ernährung, Landwirtschaft, Medizin, Umwelt und Rest sein.

9. Links zu Inhalten:

Angabe der URL die zum Volltext bzw. zu Zusatzinformationen zu dem jeweiligen Eintrag führen.

10. Zusatzinformationen:

Zuletzt werden noch weiterführende Informationen angegeben. Aufgeführt werden relevante Zusatzmaterialien, wie Karten und Tabellen. Das Format und die physische Größe, sowie Umfangsangaben zu Seitenzahlen und Anzahl an Abbildungen.

Mit bis zu über 60 beschreibenden Metadaten für einen Eintrag liefert der Datenkorpus der *LIVIVO* sehr detaillierte Informationen. Für die Auswertung sind die eindeutigen Identifikatoren von besonderem Interesse. Mithilfe dieser ist eine genaue Zuordnung von Objekten aus verschiedenen Datensätzen möglich.

3.1.2 Beschreibung Datensatz LIVIVO

Der erhaltene Datensatz der *LIVIVO* hat eine Gesamtgröße von 192GB und liegt als *JSON Lines* Datei vor. Das *JSON Lines* Textformat ist ein durch Zeilenumbrüche strukturiertes Format und bietet deshalb Vorteile bei der Arbeit mit strukturierten Daten. Bei der Strukturierung von *JSON Lines* Dateien gelten drei Bedingungen: Es muss UTF-8 Encoding gegeben sein, jede Zeile muss ein valides *JSON-Objekt* beinhalten und Zeilenumbrüche müssen durch ein *Newline-Statement* (‚\n‘) ausgedrückt werden. (jsonlines.org 2023)

Der Datensatz hat über 74 Millionen Einträge und die nachfolgende Abbildung zeigt anhand eines Beispiels, wie ein Eintrag strukturiert ist. Zur Verbesserung der Lesbarkeit wurden, mithilfe von jq, die *JSON Lines* Daten in eine für Menschen besser lesbare Form gebracht und einzelne Ausprägungen wurden entfernt, da der Fokus auf dem strukturellen Aufbau liegt.

```
{
  "_id": {
    "$oid": "58ecd32926eb6480f5f520f8"
  },
  "liv": {
    "orig_data": {
      "TITLE": [
        "Problems in the measurement of putative serum immune complexes by the method of Beaumont."
      ],
      "DOCTYPE": [
        "ARTIKEL"
      ],
      "AUTHOR": [
        "Goldzieher, J W",
        "Greene, N D",
        "Williams, M C"
      ],
      "DOI": [
        "10.1111/j.1600-0897.1982.tb00158.x"
      ],

      [...]

      "sortyear": [
        "1982"
      ]
    },
    "ts_insert": {
      "$date": "2017-04-11T12:59:14.000Z"
    },
    "ts_update": {
      "$date": "2019-11-02T06:30:33.000Z"
    }
  },
  "ts_update": {
    "$date": "2019-12-12T18:49:02.000Z"
  },
  "avail": {
    "ezb": {
      "crc": "b9a3e23475f1dc390b38adaa3586e106",
      "libs": [
        "UBU"
      ],
      "ts_update": {
        "$date": "2022-11-24T15:56:01.000Z"
      }
    },
    "ts_update": {
      "$date": "2022-11-24T15:56:01.000Z"
    }
  }
}
```


Abbildung 1: Aufbau eines JSON Objektes im LIVIVO Datensatz

Die Daten sind in verschiedenen Schlüssel-Wert Paaren organisiert. Jeder Schlüssel beschreibt eine bestimmte Kategorie von Informationen. *_id* enthält ein weiteres *JSON-Objekt* mit einer eindeutigen ID für den Artikel in der Datenbank. *liv* beinhaltet die Metadaten zum Artikel und drei weitere *JSON-Objekte orig_data*, *ts_insert* und *ts_update*. Die beiden Zeitstempel geben das genaue Datum an, an dem die Metadaten eingepflegt bzw. überarbeitet wurden. *orig_data* beinhaltet die im vorherigen Teil beschriebenen Metadaten, wie Titel, Dokumententyp, Autor und DOI, zu einem Artikel. In dem Schlüssel *avail* wird die Verfügbarkeit des Artikels in verschiedenen Bibliotheken beschrieben, in diesem Beispiel die Verfügbarkeit in der elektronischen Zeitschriftenbibliothek *ezb* und ebenfalls der Zeitpunkt der letzten Aktualisierung. Die Einträge *liv* und *avail* haben darüber hinaus noch jeweils einen Zeitstempel hinterlegt, der die Aktualisierung des Gesamtobjektes festhält (*ts_update*).

Die Daten im Objekt *orig_data* Stellen die Basis für diese Ausarbeitung dar.

3.2 Retraction Watch

Retraction Watch ist ein Blog, der seit 2010 betrieben wird und über Retractions von wissenschaftlichen Artikeln und damit verbundenen Themen berichtet. Geführt wird dieser von *Ivan Oransky* und *Adam Marcus*.

Für die Gründung waren vier Argumente ausschlaggebend: (Retraction Watch 2010)

1. Die Selbstkorrektur von Wissenschaft und dem wissenschaftlichen Arbeiten: Normalerweise sollten mehr oder andere Daten bzw. Methoden der Grund zur Widerlegung oder Erneuerung von Erkenntnissen sein, nicht Betrug oder Fehler zu einer Retraction führen. Aber wie lange dauert es, bis diese Selbstkorrektur eintritt? Am Beispiel von *Wakefield* wurde die Studie erst 12 Jahre nach ihrer Veröffentlichung retracted, obwohl bereits nach 6 Jahren kritische Fragen von dem Journalisten *Brian Deer* öffentlich wurden.
2. Retractions sind oftmals nicht gut publik gemacht worden. So haben Investoren, ob beispielsweise Steuerzahler oder Unternehmen, nicht immer von der Retraction eines Artikels gehört und somit auf Grundlage falscher Ergebnisse weitergearbeitet. *Retraction Watch* soll daher als informelles Archiv dienen und in Zukunft die Basis einer Datenbank bilden.
3. Mehr Aufmerksamkeit auf Retractions lenken und beispielsweise Journalisten die Möglichkeit geben, Betrug und Missbrauch von Fördermitteln aufzudecken.
4. Die Möglichkeit zu analysieren, wie konsistent Zeitschriften und Herausgeber sind. Zu stellende Fragen sind u.a.: Wie lange wird gewartet, bis eine Retraction gedruckt wird? Wie viele öffentliche Ankündigungen werden gemacht? Lassen sich Ableitungen zur Güte des Peer-Review Verfahrens anhand der Anzahl von retracted Artikeln festlegen?

2018 hat *Retraction Watch*, wie in Punkt 2 bereits angekündigt, eine eigene Datenbank (The Center for Scientific Integrity 2018) mit über 18.000 Einträgen veröffentlicht. Nach eigenen Aussagen stellt die Datenbank die größte Menge an retracted Artikeln auf der Welt dar. Bei der Erarbeitung wurden alle Artikel geprüft und mit Begründungen, die zur Retraction führten, versehen. Hierzu wurde eine eigene Taxonomie geschaffen. (Retraction Watch 2018)

Der aktuelle Datensatz (Stand Januar 2023) der *Retraction Watch* stellt einen essenziellen Bestandteil dieser Untersuchung dar und dient als Grundlage, um die Daten der *LIVIVO*-Datenbank (*ZB MED Knowledge Environment*) mit Informationen zu retracted Artikeln anzureichern. Schauen wir uns im Folgenden die Metadaten und den Datensatz an.

3.2.1 Beschreibung Metadaten Retraction Watch

Der vorliegende Datensatz hat, mit Stand Januar 2023, 20 beschreibende Metadaten. Die folgende Tabelle zeigt einen Überblick über die vorkommenden Ausprägungen und wie diese gepflegt werden.

Ausprägung	Beschreibung	Hinweis
Record ID	Eindeutiger Identifikator	Index, Ganzzahl
Title	Titel des Eintrags	Fließtext
Subject	Themengebiet des Eintrags	Oberkategorien und Unterkategorie (Beispiel: (BLS) Biology – Cancer) – mehrere Zuordnungen möglich (getrennt durch ;)
Institution	Beteiligte Institutionen	Mehrere Zuordnungen möglich (getrennt durch ,)
Journal	Veröffentlichende Zeitschrift	-
Publisher	Herausgeber	Personen oder Verlage
Country	Veröffentlichungsland	Unkown wenn nicht bekannt – Mehrfachnennungen durch ; getrennt
Author	Autor/ Autoren	Nennung des Autors oder der Autoren (Mehrfachnennung durch ; getrennt)
URLS	Weblink	Angabe der URL zum Artikel der Retraction auf Retractionwatch.com
ArticleType	Angabe des Typen	Mehrfachnennung möglich und durch ; getrennt
RetractionDate	Datum Retractionnotiz	Format: TT.MM.JJJJ hh:mm

Ausprägung	Beschreibung	Hinweis
RetractionDOI	DOI Retractionnote	Besteht immer aus Präfix und Suffix wobei beide durch einen Schrägstrich getrennt werden und der Präfix stets mit 10. Beginnt (z.B. 10.14802/jmd.12011)
RetractionPubMedID	PubMed Referenznummer	Leereinträge mit 0, oder nicht gepflegt
OriginalPaperDate	Datum Veröffentlichung	Format: TT.MM.JJJJ hh:mm
OriginalPaperDOI	DOI Artikel	Besteht immer aus Präfix und Suffix wobei beide durch einen Schrägstrich getrennt werden und der Präfix stets mit 10. Beginnt (z.B. 10.14802/jmd.12011)
OriginalPaperPubMedID	PubMed Referenznummer	Leereinträge mit 0, oder als leeres Feld
RetractionNature	-	Einzige Ausprägung ist „Retraction“
Reason	Retraction Grund	Auflistung der Begründung die zur Retraction führte. Mehrfachnennung mit ; getrennt
Paywalled	Kosten	Ausprägungen Yes oder No
Notes	Spezielle Vermerke	Beispiel: „Author banned from Journal for plagiarism“

Tabelle 1: Übersicht Metadaten Retraction Watch

Der Datensatz gibt einen sehr detaillierten Einblick zu den Einträgen. Es ist leicht möglich einem Artikel Kennungen wie der *DOI* oder der *PubMedID* zuzuordnen, sowohl für den Originalartikel oder den retracted Artikel. Ebenfalls sind die wichtigsten Eckdaten, wie beispielsweise der Titel, eine Länderzuordnung, der Herausgeber und die Zeitschrift enthalten. Durch die Angabe des Veröffentlichungsdatum des Originalartikels sowie der Veröffentlichung der Retraction-Notiz lässt sich der zeitliche Verlauf nachvollziehen.
Die Ausprägungen für *Subject* sind in sechs Oberkategorien aufgeteilt *BLS (Biological Science), PHY (Physics), B/T (Business & Technology), SOC (Sociology), ENV (Environmental Sciences) und HSC (Health Science)*. Unter diesen Oberkategorien wird

noch feingliedriger und nach spezieller Fachrichtung unterschieden. Beispiel: *HSC* hat als Unterkategorien *Medicine/ Surgery und Medicine/ Oncology*. Eine Differenzierung in detailliertere Unterkategorien findet für jede Oberkategorie, bis auf *Sociology* statt. Ein Artikel kann mehreren Oberkategorien und zeitgleich auch mehreren Unterkategorien zugeordnet werden (z.B. *(B/T) Computer Science;(HSC) Medicine - Neurology;(HSC) Medicine - Oncology;(HSC) Radiology/Imaging*).

Für das Feld *Reason* ist ebenfalls eine Mehrfachnennung zulässig. Ausprägungen reichen von dem Zitieren von anderen retracted Artikeln, über Fehler in Analyse, Methodik oder Ergebnissen bis hin zu Betrug beim Peer-Review-Verfahren. *Retraction Watch* hat hierzu eine eigene Taxonomie geschaffen und es liegen 102 verschiedene Begründungen für eine Retraction vor. Am Beispiel des Papers „Teaching Mode of Augmented Reality College English Listening and Speaking Supported by Wearable Technology“ von Xiaochun Ma ist die Zuordnung gut ersichtlich. Der Artikel wurde aufgrund von Betrug im Peer-Review Verfahren (Ausprägung + Fake Peer Review) retracted. Ursache für die Prüfung waren Nachforschungen des Herausgebers bzw. der Zeitschrift (Ausprägung +Investigation by Journal/Publisher). Des Weiteren wurde in dem Eintrag der Widerspruch des Autors gegen die Retraction vermerkt (+Objections by Autor(s)). Der Artikel hat somit drei Ausprägungen für das Attribut *Reason*.

In dem Datenfeld für den Autor, das Erscheinungsland und Institution werden mithilfe der Mehrfachnennung alle an dem Paper beteiligten Instanzen aufgelistet.

Für diese Ausarbeitung scheinen die Ausprägungen *URLS, RetractionNature, Paywalled* und *Notes* weniger interessant und werden daher im weiteren Verlauf vernachlässigt.

3.2.2 Beschreibung Datensatz Retraction Watch

Anschließend an die Darstellung der Metadatenstruktur, schauen wir uns im Folgenden die Zusammensetzung und den Inhalt etwas genauer an.

Übergeben wurde der Datensatz in Form einer Excel-Tabelle mit 38.275 Einträgen. Es liegen 35.687 originale Artikel-DOI vor und 37.741 der DOI der Retraction-Notiz. Einträge bei denen der DOI fehlt, werden entweder mit *unavailable*, oder als leeres Feld gekennzeichnet. Es sind Publikationen und die dazugehörigen Retractions von 1940 bis 2023 zu finden. Die meisten Publikationen des Datensatzes stammen aus den Jahren 2010 (5.243) und 2011 (5.182). Für die Jahre 2021 und 2020 liegen bereits jeweils ca. 2700 Artikel vor. Diese Anzahl wird in den kommenden Jahren voraussichtlich steigen. Vermuten lässt dies ein Blick auf den zeitlichen Verlauf der Anzahl des Retraction-Datums. Es ist seit dem Jahr 2012 ein deutlicher Anstieg der Anzahl an Retractions zu erkennen. Von 1.113 ist die Anzahl im Verlauf von zehn Jahren um fast das Fünffache auf 5.003 gestiegen. Für die Jahre 2010 und 2011 liegen analog zu den Veröffentlichungen eine hohe Anzahl an Retractions, mit jeweils ca. 4.900, vor.

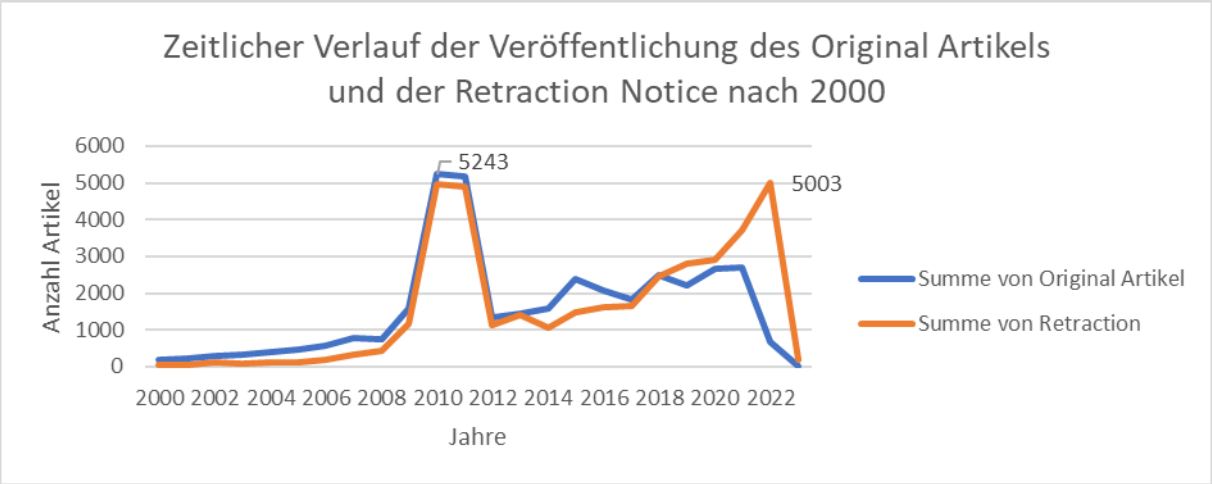

Abbildung 2: Anzahl Retractions und Publikationen ab 2000

Wie kommt es zu den Ausreißern in den Jahren 2010 und 2011? Wieso gibt es so viele Einträge zu diesen Jahren und warum sinkt in den darauffolgenden die Anzahl zunächst stark ab? Antwort auf diese Frage liefert ein Blick auf die Spalten des Herausgebers und des Journals. Das *Institute of Electrical and Electronics Engineers (IEEE)* in New York City hat tausende von Konferenzartikeln retracted. Für 2010 waren es 4.385 und für 2011 4.071 Retractions. Zum Großteil kommen die Autoren aus China und decken eine breite Themenpalette von Physik bis Soziologie ab. Die meisten Texte sind von Konferenzen aus den Jahren 2009 bis 2011, wobei die 2011 *International Conference on E-Business and E-Government (ICEE)* allein 1.280 Retractions nach sich zog. Laut Aussagen vom *IEEE* könne man keine Aussagen zu den genauen Gründen, die zur Retraction führten, machen. Als Begründung wurde das Prüfen alter Kataloge angegeben, bei denen Abweichungen von den Richtlinien für die Artikel erkannt wurden. Der Computerwissenschaftler *Lior Pachter* vom *California Institute of Technology* in *Pasadena* sagt in diesem Zusammenhang, dass diese Spitze die schnellere und weniger intensive Form der Peer-Review widerspiegeln kann, die Konferenzbeiträge im Vergleich zu traditionellen Zeitschriftenbeiträgen häufig durchlaufen (Alison Mccook 2018). Der beschleunigte Zeitplan ermögliche eine schnelle Verarbeitung von Ideen und schnelle Verteilung, aber es könne auch bedeuten, dass Fehler unbemerkt bleiben. Um zukünftige Massen-Retractions zu verhindern, erklärt das *IEEE*, dass es einen Ausschuss aus Mitarbeitern und freiwilligen Experten gebildet habe, um als „Gatekeeper“ für Konferenzmaterialien zu dienen und eine zusätzliche Qualitätskontrolle zu bieten (Alison Mccook 2018). Neben dem *IEEE* sind weitere bekannte Herausgeber in den Top 3. Elsevier ist mit 4.662 und Springer mit 2.981 Einträgen vertreten.

Beeinflusst durch diese Flut an Retractions wird ebenfalls die Auswertung nach den Ländern der Publikationsstandorte. Bei der Analyse nach wissenschaftlicher Beteiligung an der Veröffentlichung nach Ländern steht China, mit 18.340 Artikeln, weit an der Spitze. Darauffolgend die USA mit 4.668 und Indien mit 2.554. Die Differenzierung nach Artikeltyp zeigt, dass über 20.000 Einträge klassische *Research Article* sind, gefolgt von

Conference Abstract/Paper mit 12.247 Artikeln. Bei den Konferenzartikeln sind die oben beschriebenen Artikel aus den Jahren 2009 bis 2011 enthalten.

Thematisch lässt sich der Datensatz, wie oben bereits beschrieben, in sechs Oberkategorien einteilen. Da ein Artikel mehreren Fachrichtungen zugeordnet werden kann, wird in der Abbildung 3 die Verteilung nach Themengebieten nach Vorkommnissen gezählt. Wenn ein Artikel beispielsweise in dem Themengebiet Biologie und Physik vorkommt, wird dieser für Beide gezählt. Hierbei wird die prozentuale Verteilung auf den Korpus betrachtet.

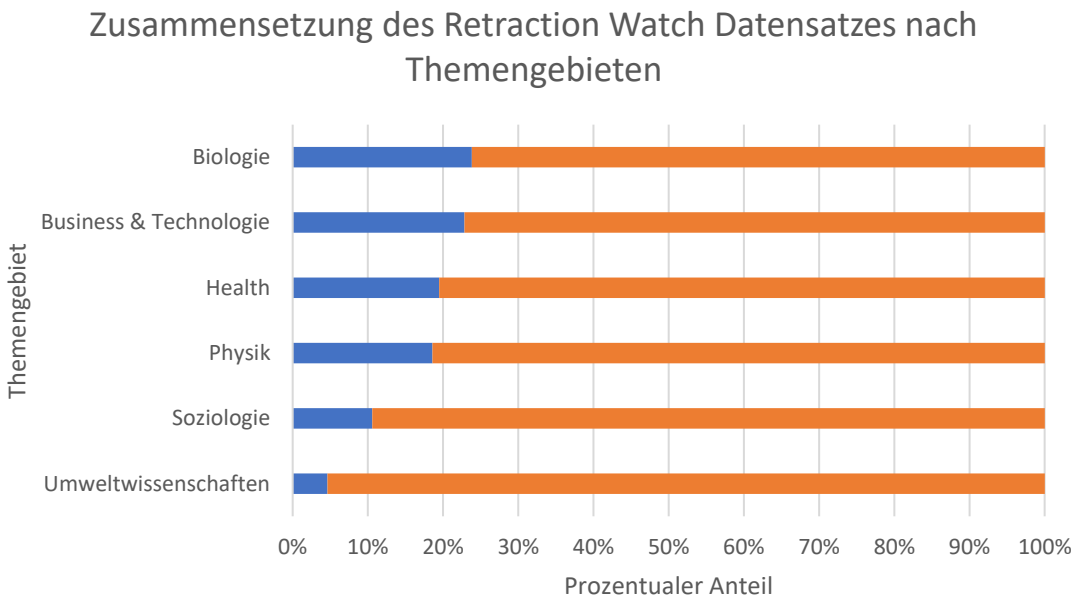

Abbildung 3: Themenverteilung im Datensatz der Retraction Watch

In Abbildung 3 ist gut zu erkennen, dass die Themengebiete Biologie, Business, Health und Physik mit 19% bis 24% am stärksten vertreten sind. Soziologie kommt auf 10% und die Umweltwissenschaften stellen 5% dar. Eine Einschätzung über die Qualität und Arbeitsweisen in den einzelnen Bereichen ist mithilfe dieser Auswertung nicht möglich. Sie stellt lediglich eine grobe Übersicht über das Themenspektrum dar.

Zuletzt schauen wir uns die Ausprägungen zu dem Attribut *Reason* etwas genauer an. Ein Artikel kann mehrere Gründe besitzen, wieso eine Retraction durchgeführt werden musste und weiterführend auch noch Informationen zum Grund der Untersuchung (z.B. initiiert durch den Herausgeber) und dem Einspruch (z.B. durch den Autor). Eine erste Auswertung zeigt, dass in den Top 3 Vorkommnissen über 3.000 Einträge mit dem Verstoß gegen Policen vom Autor begangen wurden. Diese 3.000 Einträge haben jedoch nicht nur diesen einen Verstoß als einzige Ausprägung. Die Mehrfachnennung von Gründen macht die Zuordnung zu einem spezifischen Grund erheblich schwerer. Um die Hintergründe für die Korrekturen besser zu verstehen, wird im Rahmen der Analyse versucht, die Artikel einem Einzigen, dem maßgeblichen bzw. relevantesten Grund zuzuordnen. Mögliche Ausprägungen hierfür sind Betrug, Fehlverhalten, Fehler und Sonstige.

3.3 Erkenntnisse und Ansatz

Beide Datensätze bieten auf unterschiedliche Weise detaillierte Informationen zu den darin enthaltenen Artikeln und Inhalten. Als gemeinsamer, eindeutiger Identifikator steht der digitale Objekt Identifikator (*DOI*) zur Verfügung. Dieser soll genutzt werden, um die Schnittmenge zwischen den beiden Datensätzen herzustellen und eine Aussage über die Anzahl der retracted Artikel innerhalb der *LIVIVO* treffen zu können. Dazu werden die Datensätze analysiert, die benötigten Informationen extrahiert und miteinander verglichen. Aufsetzend auf der Konsolidierung der Datensätze wird das Ergebnis ausgewertet. Von Interesse ist dabei, welcher Grund zu der Retraction geführt hat, wie sich die Retractions auf die Fachgebiete verteilen, welche Verlage und Zeitschriften vertreten sind und ob sich ein Trend über die Jahre erkennen lässt.

4 Pipeline

Im folgenden Abschnitt wird eine Pipeline vorgestellt, genauer gesagt ein Ablauf wie die vorhandenen Daten aufbereitet, verarbeitet und zusammengeführt werden. Außerdem werden notwendigen theoretischen Grundlagen und die verwendeten Werkzeuge (Programme, Anwendungen etc.) vorgestellt, um das Ergebnis zu erzielen.

4.1 Identifikation & Vorstellung benötigter Bestandteile

Betrachten wir zunächst die benötigten Komponenten, die für die Verarbeitung der *LIVIVO*-Daten benötigt werden. Die benötigte Ausprägung des Datensatzes ist der *DOI*. Um die über 190 GB große Datei verarbeiten zu können, wurde von der ZB MED eine virtuelle Maschine (VM) zur Verfügung gestellt. Über das *Deutsches Netzwerk für Bioinformatik-Infrastruktur (de.NBI)* konnte eine *SimpleVM* genutzt werden. *SimpleVM* ist eine einfache Lösung, um eine Maschine aufsetzen zu können. Man kann aus einem Baukastensystem Leistung, Größe, Betriebssystem (verschiedene *Ubuntu Linux* Versionen) und auch Anwendungen auswählen. Die Konfiguration erfolgt automatisch und man kann sich über das *Secure Shell (SSH)* Protokoll verbinden (de.NBI Cloud Portal 2023). Für die *DOI*-Extraktion wird ein *Python*-Skript verwendet, das über die Einträge der *JSON Lines*-Datei iteriert und die benötigten Informationen in einer *CSV*-Datei speichert.

Die Daten der *Retraction Watch* liegen als Excel Arbeitsmappe vor und werden ebenfalls in das *CSV-Format* konvertiert. Diese Formatanpassung ermöglicht die Weiterverarbeitung. Mit Hilfe einer lokalen Datenbank sollen die beiden Rohdaten eingelesen und verarbeitet werden. Zum Einsatz kommt *PostgreSQL*, ein Open-Source-Relationales Datenbank-Management-System (RDBMS). *PostgreSQL* ist eine leistungsfähige und zuverlässige Lösung, um große Datenmengen effizient zu verwalten und zu analysieren. Zur besseren Handhabung wird die *PostgreSQL* Administrationssoftware *pgAdmin 4* eingesetzt. *pgAdmin* bietet eine grafische Benutzeroberfläche und ermöglicht das Erstellen, Verwalten und Überwachen von Datenbanken.

4.2 Theoretische Grundlagen

Da der gute Zustand und die Qualität der Daten für die weitere Verarbeitung sichergestellt werden muss, wird das Prinzip der Landing Zone bzw. Staging Area innerhalb der lokalen *PostgreSQL*-Datenbank verwendet. Dabei handelt es sich um einen temporären Zwischenspeicher, der zum Extrahieren, Transformieren und Laden der Daten genutzt wird. Dieser dreistufige Prozess als wird *ETL*-Prozess abgekürzt. Dieser Bereich wird für eine Reihe von Funktionen zum Bereinigung, Ändern, Kombinieren, Konvertieren,

Entfernen von Duplikaten und Vorbereiten der Quelldaten für die dauerhafte Speicherung und anschließende Verwendung genutzt.

Die drei Phasen des ETL-Prozesses bestehen aus: (Ponniah 2001)

1. Daten Extraktion

Der erste Schritt besteht darin, die verschiedenen Quelldaten zu sichten und das richtige Dateiformat zu wählen. Für die Bereitstellung des richtigen Datenformats können Software oder Programme wie z.B. *Excel* oder *Python* verwendet werden. Anschließend werden die Daten in ihrer Rohform eingelesen.

2. Daten Transformation

Die Datentransformation umfasst eine Reihe von Aufgaben. Zunächst werden die aus jeder Quelle extrahierten Daten bereinigt, was die Korrektur von Datumsformaten oder die Lösung von Konflikten zwischen einzelnen Datenelementen umfassen kann. Darüber hinaus können Standardwerte für fehlende Datenelemente bereitgestellt oder Duplikate entfernt werden. Ein wichtiger Teil der Datentransformation ist die Standardisierung von Datenelementen, bei der Datentypen und Feldlängen für identische Datenelemente aus verschiedenen Quellen standardisiert werden. Daten aus verschiedenen Quellen werden kombiniert, indem Daten aus einem Quelldatensatz oder verwandte Datenelemente aus mehreren Quelldatensätzen zusammengeführt werden. Unnötige Daten werden entfernt.

3. Daten Laden

Im letzten Schritt werden die transformierten Daten in einen permanenten Speicher überführt. Dazu wird ein entsprechendes Schema angelegt und Tabellen mit den zugeschnittenen Datentypen vorbereitet.

Nach erfolgreichem Abschluss dieses Prozesses liegt eine fertige Datenstruktur vor, die für weitere Analysen vorbereitet ist.

4.3 Zusammensetzung und Zielbild

Abbildung 4 zeigt die fertige Pipeline und das grobe Vorgehen, um die Metadaten der *LIVIVO* um die Informationen von *Retraction Watch* zu ergänzen.

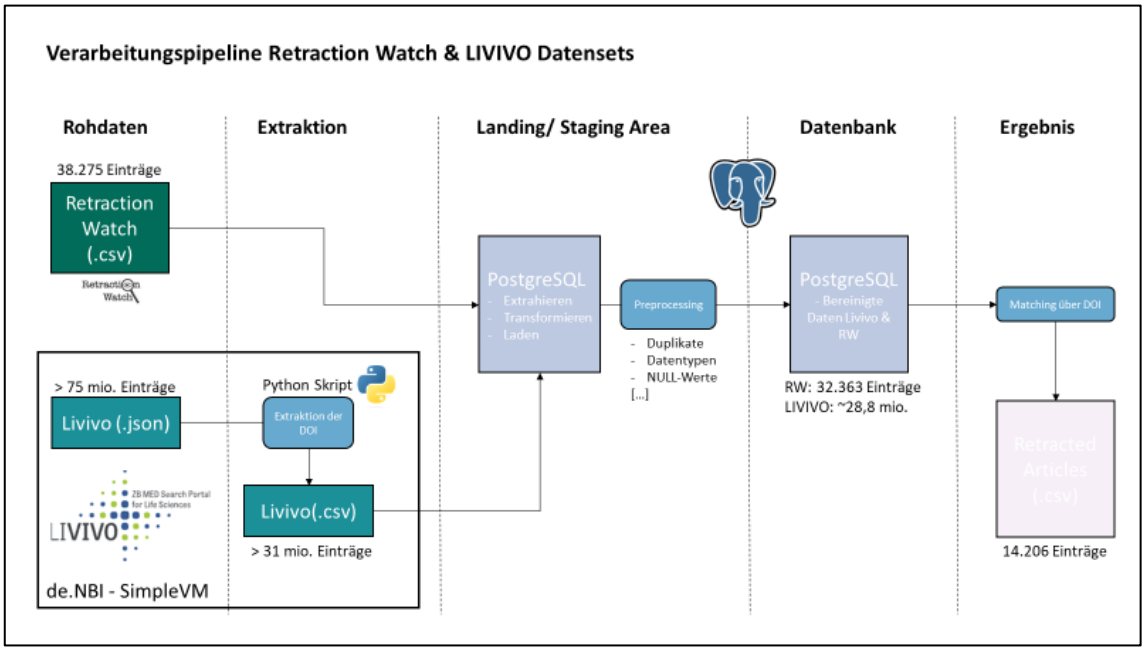

Abbildung 4: Pipeline zur Anreicherung der Metadaten

Die Verarbeitung der *LIVIVO*-Rohdaten wird auf der VM von *de.NBI* stattfinden und beinhaltet einen zusätzlichen Schritt der Datenextraktion aus dem Gesamtkorpus mithilfe eines *Python*-Skripts. Der Datensatz der *Retraction Watch* wird in das *CSV-Format* konvertiert. Anschließend werden die beiden Datenquellen mit Hilfe des oben beschriebenen *ETL*-Prozesses in die *PostgreSQL* Datenbank geladen und miteinander verknüpft. Als Ergebnis wird eine *CSV-Date*i aus *PostgreSQL* heraus generiert, die die Schnittmenge der beiden Datenquellen enthält und für weitere Analysen verwendet werden kann.

5 Umsetzung

Nach der Beschreibung der Datenquellen und der Planung des Verarbeitungsablaufs geht es in die Implementierungsphase. Dieses Kapitel befasst sich mit der Verarbeitung der verschiedenen Datenquellen und der anschließenden Erstellung und Befüllung der lokalen Datenbankanwendung in *PostgreSQL*. Als Ergebnis wird eine Zieldatei angestrebt, die Informationen zu den im *LIVIVO*-Suchportal vorhandenen retracted Artikeln enthält.

5.1 Rohdaten und Extraktion

Für das initiale Einlesen in die lokale *PostgreSQL-Datenbank* müssen die Daten im *Comma-separated values* Format (*CSV-Format*) vorliegen. Für die *Retraction Watch* Datenquelle, welche als Excel-Arbeitsmappe vorliegt, war ein Konvertierung mit UTF-8 Encoding problemlos möglich. Dazu musste die Datei lediglich unter dem gewünschten Format mit der Kodierung abgespeichert werden.

```python
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

"""
File: extract_jsonl
Author: Lukas Galke, Eva Seidlmayer - Adjustment Lucas Vetter
Email: seidlmayer@zbmed.de
Github:
Description: metadata (Extract DOI)
Last Change: Feb 2, 2023
"""

import argparse
import os
import jsonlines
import re
import pandas as pd
from tqdm import tqdm

DEBUG = False


def main():
    """ Extracts DOI from jsonlines file
    """
    parser = argparse.ArgumentParser()
    parser.add_argument("jsonl_file")
    parser.add_argument("-o", "--output", dest='save', default=None, type=str,
                        help="Save files in this directory")
    args = parser.parse_args()

    #open ZB MED data file
    with jsonlines.open(args.jsonl_file) as json:
        flat_gen = ((d['_id']['$oid'], d['liv']['orig_data'])
                    for d in json)


        #set counter
        n_extracted = 0

        # initiate metadata container
        metadata = []

        for identifier, data in tqdm(flat_gen):
            if "DOI" not in data:
                continue

            #define infos: DOI

            doi = data.get('DOI')[0] if 'DOI' in data else None
            metadata.append(doi)

            # add 1 to counter
            n_extracted += 1

    print("Extracted %d papers" % n_extracted)

    # create dataframes from infos
    df_paper_metadata = pd.DataFrame(metadata,
                                     columns=['doi'
                                              ])


    dfs = [
        # dataframe name, dataframe, whether to save index
        ('livivo_doi', df_paper_metadata, False)
    ]

    if args.save is not None:
        # Dumps everything to disk
        os.makedirs(args.save, exist_ok=True)
        for fname, dframe, save_index in dfs:
            dframe.to_csv(os.path.join(args.save, fname + '.csv'),
                          index=save_index)
        with open(os.path.join(args.save, 'args.txt'), 'w') as fh:
            print(args, file=fh)
```

Abbildung 5: Python Skript zur Extraktion der DOI

Für die Extraktion des *DOI* aus dem gesamten Datenkorpus der *LIVIVO* und die anschließende Speicherung als *CSV-Datei* musste ein Zwischenschritt eingebaut werden. Um das gewünschte Ergebnis zu erzielen, wurde ein *Python*-Skript verwendet. Das Skript prüft für jeden Eintrag in der *JSON-Datei*, ob das Schlüssel-Wert-Paar für den *DOI* enthalten ist, arbeitet sich iterativ durch die Datei und speichert alle gefundenen Einträge gebündelt in einer *CSV-Datei*.

Wie in Abbildung 5 zu sehen ist, wurde eine Funktion geschrieben, die mit Hilfe der *Python*-Bibliothek *argparse* verschiedene Eingabeparameter von der Kommandozeile von *Ubuntu* entgegennehmen kann. Die übergebenen Parameter sind die Input-Datei, also die Datenquelle der *LIVIVO* im *JSON-Lines-Format* und der Name des zu erstellenden Ordners, in dem die später generierte *CSV-Datei* gespeichert wird.

Die erwartete Struktur der *JSON-Datei* ist

definiert und unter dem *JSON-Objekt orig_data* befindet sich der *DOI*. Als nächstes wird die Bedingung definiert, dass für den durchsuchten Eintrag ein *DOI* vorhanden sein muss, andernfalls wird der Eintrag übersprungen. Wird ein Eintrag gefunden, so wird dieser in den Metadaten-Container übergeben und anschließend über den nächsten Eintrag innerhalb der *JSON-Datei* iteriert. Schließlich werden alle gefundenen *DOI* in ein *Pandas*-Dataframe übertragen und als *CSV-Datei* gespeichert.

```
(base) ubuntu@accuratecurie-d64eb:/vol/volume/livivo-data$ python3 harvest_lv_doi.py livivo.json -o gesamt
24026953it [15:41, 13818.01it/s]
31129172it [26:20, 8763.20it/s]
74528729it [1:53:42, 10923.89it/s]
Extracted 31314696 papers
```

Abbildung 6:Laufzeit und Ergebnis DOI Extraktion

Wie in Abbildung 6 gut zu erkennen ist, wurden über 31 Millionen Artikel mit hinterlegtem DOI gefunden und gespeichert. Das Skript ist über 74 Millionen Einträge iteriert und benötigte eine Laufzeit von knapp unter zwei Stunden.

Durch diesen Zwischenschritt liegen die Datenquellen im gewünschten Dateiformat vor und können in die Landing bzw. Staging Area der Datenbank geladen und weiterverarbeitet werden.

5.2 Staging Area & Erstellung Datenbank

Um Dateien in eine PostgreSQL-Datenbank laden zu können, wird eine Tabellenstruktur analog zur einzulesenden Datei benötigt. Dazu werden die Spalten anhand der Metadatenbezeichnungen für beide Datensätze definiert. Für die extrahierten *LIVIVO-DOI* wird nur eine einzelne Spalte benötigt. Für den Datensatz von *Retraction Watch* wird ein Schema für die 20 vorhandenen Metadaten erstellt. Um ein reibungsfreies Einlesen zu gewährleisten, werden zunächst alle Spalten als Textformat definiert, alternativ können auch Charaktere variabler Länge verwendet werden. Die Datenformate der Spalten sind in dieser Phase zu vernachlässigen, da dieser Prozessschritt nur eine Zwischenspeicherung darstellt und der Fokus auf dem fehlerfreien Einlesen liegt. Die spezifischen Formate werden nach der Sichtung und Transformation der Daten festgelegt.

Für die verschiedenen Datensätze wurden unterschiedliche Transformationen durchgeführt. In den LIVIVO-Daten wurden mehrfach die gleichen *DOI* gefunden und daher eine Dublettenbereinigung durchgeführt. Am Beispiel des *DOI* „10.1086/BBLv99n2p321“ soll das Auftreten erläutert werden. Dieser *DOI* ist mit insgesamt 96 Vorkommen der am häufigsten vorkommende. Er verweist auf verschiedene Artikel aus *The biological bulletin* von 1950, Ausgabe 99, Heft 2. Ebenso sorgen Einträge wie *notULavail* (201 Vorkommen) und *NO_DOI* (144 Vorkommen) für weitere Dubletten. Nach der Bereinigung reduziert sich die Anzahl der eindeutigen DOI auf ca. 28,9 Millionen.

Bei der Verarbeitung der Retraction Watch Rohdaten wurden zunächst irrelevante Spalten entfernt. Wie bereits in der Beschreibung der Rohdaten erwähnt, sind die Ausprägungen *URLS, RetractionNature, Paywalled* und *Notes* nicht von Interesse und werden nicht in die Datenbank geladen. Die enthaltene *RecordID* wurde ebenfalls entfernt und durch einen eigenen Index ersetzt. Durch das Öffnen der *Excel*-Datei mit dem deutschen Sprachpaket von *Excel* und der anschließenden Umwandlung in eine CSV-Datei liegen die Datumsfelder für den *DOI* und die Retraction im Format TT.MM.JJJJ vor. *PostgreSQL* benötigt für ein Datumsfeld die Formatierung nach ISO 8601, also das Datum nach JJJJ-MM-TT(PostgreSQL Documentation 2011).

Die Spalte *OriginalPaperDOI* ist für das Zusammenführen der beiden Datensets der Primärschlüssel. Um diesen in der Datenbank definieren zu können müssen zwei Bedingungen erfüllt sein. Erstens muss die Eindeutigkeit gewährleistet sein, d.h. es dürfen keine doppelten Werte in der Spalte des Primärschlüssels vorhanden sein. Zweitens dürfen keine Nullwerte, d.h. leere Zellen, enthalten sein. Einträge mit der Bezeichnung *unavailable* und *Unavailable* wurden ebenfalls bei der Dublettensuche identifiziert und entfernt. Der ursprüngliche Datensatz mit über 38 Tausend Einträgen wurde auf 32.363 reduziert. Am Beispiel der Publikation von *Brian Wansink* zu „Can Branding Improve School Lunches?“ (DOI: 10.1001/archpediatrics.2012.999) lässt sich das Auftreten von Duplikaten erklären. Der Eintrag ist zweimal vorhanden und wurde auch zweimal retracted, das erste Mal am 21.09.2017. Der Artikel wurde retracted und direkt durch eine neue Version ersetzt. Im Feld *Reasons* ist dies mit der Ausprägung *Retract and Replace* vermerkt. Auch diese Neuveröffentlichung hatte keinen Bestand und der Artikel wurde am 20.10.2017 erneut retracted. Bei der Bereinigung wurden 2.588 Nullwerte, 3.282 *unavailable* und 44 Dubletten entfernt. Bei den übrigen verbleibenden Metadatenspalten mussten keine Transformationen vorgenommen werden.

Um die transformierten Daten in die finalen *PostgreSQL* Tabellen zu überführen, wurden diese wie folgt definiert:

Format	**Ausprägung**
Datum	RetractionDate, OriginalPaperDOI
Integer	RetractionPubMedID, OriginalPaperPubMedID, Index

Format	Ausprägung
Character Varying	Title, Subject, Institution, Journal, Publisher, Country, Author, ArticleType, RetractionDOI, OriginalPaperDOI (Primärschlüssel), Reason
	-
	DOI LIVIVO Daten (Fremdschlüssel)

Tabelle 2: Datentypen Retraction Watch & LIVIVO

Wie aus Tabelle 2 hervorgeht, sind die meisten Ausprägungen als Zeichen mit variabler Länge definiert, da es sich um Text oder nicht fest definierbare Zeichenketten handelt. Das Veröffentlichungsdatum des Artikels und das Veröffentlichungsdatum der Retraction wurden als Datumstyp definiert. Die Identifikationsnummern wurden jeweils als Ganzzahlen angelegt. Als Primärschlüssel wurde der *DOI* des Originalartikels definiert. Bei der Erstellung der Tabelle für die *DOI* von *LIVIVO* wurden diese als Fremdschlüssel mit Referenz auf den Primärschlüssel angelegt. Damit ist die endgültige Struktur für die Daten gegeben und diese können nach erfolgreicher Transformation in die Tabellen geladen werden.

5.3 Anreicherung der LIVIVO-Metadaten

Aufsetzend auf den vorliegenden Tabellen zu *Retraction Watch* und *LIVIVO* können diese jetzt zusammengefügt werden und im Anschluss analysiert werden.

Die Schnittmenge der beiden Tabellen stellt das gewünschte Ergebnis dar. In *SQL* verbindet eine JOIN-Operation Zeilen aus zwei oder mehr Tabellen basierend auf einer verwandten Spalte zwischen ihnen. Das Ergebnis ist eine einzige Tabelle, die Spalten aus allen verknüpften Tabellen enthält. In diesem Anwendungsfall wird ein *INNER JOIN* verwendet. Es werden nur Zeilen zurückgegeben, die übereinstimmende Werte in beiden Tabellen haben.

Die Ergebnistabelle enthält 14.206 Einträge zu gefundenen Retractions im Datensatz von *LIVIVO*. Um das Resultat genauer betrachten zu können, wurden die in Tabelle 3 aufgeführten Metadaten übernommen.

Anzahl an übernommenen Metadaten	Ausprägungen
15	Title, Subject,Institution, Journal, Publisher, Country, Author, Article Type, Retraction Date, Retraction DOI, Retraction PubMedID, Original Paper DOI, Original Paper Date, Original Paper PubMedID, Reason

Tabelle 3: Enthaltene Metadaten der Ergebnisdatei

Als wesentliche Erweiterung der bisherigen Datenstruktur von *LIVIVO* werden die Informationen zur Retractions empfohlen. Dabei handelt es sich um den *DOI*, unter dem die Retraction veröffentlicht wurde, und das entsprechende Veröffentlichungsdatum. Als weiteres Attribut wird ein boolescher Wert empfohlen, der der Wahrheit entspricht, wenn der Artikel zurückgezogen wurde. Die Dokumentation von OpenAlex gibt ein gutes Beispiel, wie dies dargestellt werden kann. Es ist eindeutig zu sagen, ob ein Artikel zurückgezogen wurde, aber wenn der Wert falsch ist, kann der Artikel trotzdem zurückgezogen worden sein. Aufgrund der spärlichen Verfügbarkeit von Open-Source-Informationen zu Retractions ist es schwierig, ein vollständiges Bild aller Retractions abzubilden. (OpenAlex 2023a)

Das Ergebnis wurde in eine CSV-Datei konvertiert und wird im folgenden Abschnitt auf seine Zusammensetzung hin untersucht.

6 Analyse

Dieser Abschnitt befasst sich mit der Beschreibung und Analyse der Ergebnisdatei. Dabei wird die Zusammensetzung der insgesamt 14.206 Datensätze anhand der enthaltenen Metadaten untersucht. Zunächst wird die Methodik der Datenanalyse erläutert. Anschließend erfolgt eine detaillierte Inhaltsanalyse. Durch weitere Recherchen wird versucht, die Ergebnisse in einen logischen Zusammenhang zu bringen und so ein tieferes Verständnis zu erlangen.

6.1 Methodik

Um den Datensatz deskriptiv auswerten zu können, mussten folgende Anpassungen und Festlegungen getroffen werden:

Wie bereits in Tabelle 2 beschrieben wurde, sind bei den Merkmalen Land und Autor Mehrfachnennungen möglich. Fast alle Artikel haben mehrere Autoren, einige sogar mehrere Autoren in mehreren Ländern. Um eine Auswertung zu ermöglichen, wurde jeder Artikel pro Land bzw. Autor mehrfach gezählt.

Retractions werden nach dem Publikationsjahr des Originalartikels ausgewertet, nicht nach dem Publikationsjahr der Retraction-Notiz. Wie das Beispiel von Wakefield zeigt, kann es viele Jahre dauern, bis eine Retraction veröffentlicht wird.

Die meisten Retractions haben mehrere Begründungen. Um eine eindeutige Analyse des Grundes, der zur Retraction geführt hat, geben zu können, wurde die von Retraction Watch erstellte Taxonomie analysiert und in Cluster unterteilt. Ziel ist es jedem Eintrag einen eindeutigen Grund zuordnen zu können. Durch die Recherche der offiziellen Retraction-Notizen von verschiedenen Artikeln, war es möglich, die einzelnen Gründe für die Retraction der Artikel zu einem Cluster zuzuordnen.
Dabei wurde eine hierarchische Struktur gewählt. Es zählt also immer der ranghöchste Verstoß und schließt damit eine Zuordnung zu untergeordneten Kategorien aus. Die genaue Zuordnung ist in Abbildung 7 dargestellt. Die Kategorien setzen sich wie folgt zusammen:

1. Betrug: alle Attribute, die mit wissenschaftlichem Betrug zu tun haben. Plagiarismus, Verfälschung, oder Generierung von beispielsweise falschen Daten, Ergebnissen usw. Beispiel: *Plagiarism of Data*
2. Fehlverhalten: absichtliches und unethisches Fehlverhalten. Beispiel: *Ethical Violations by Author*

3. Mögliches Fehlverhalten: Betrug oder Fehlverhalten wird vermutet, jedoch kann keine konkrete Aussage getroffen werden. Beispiel: *Concerns/Issues About Results*
4. Zuverlässigkeit, Belastbarkeit: die Erkenntnisse sind nicht reproduzierbar oder nicht zuverlässig. Beispiel: *Unreliable Data*
5. Fehler: Fehler in dem wissenschaftlichen Prozess, die nicht in den oberen Kategorien abgebildet werden. Beispiel: *Error in Methods*
6. Sonstige: alle weiteren Retractions, die weitere Gründe, oder keine vorliegend haben. Beispiel: *Doing the Right Thing*

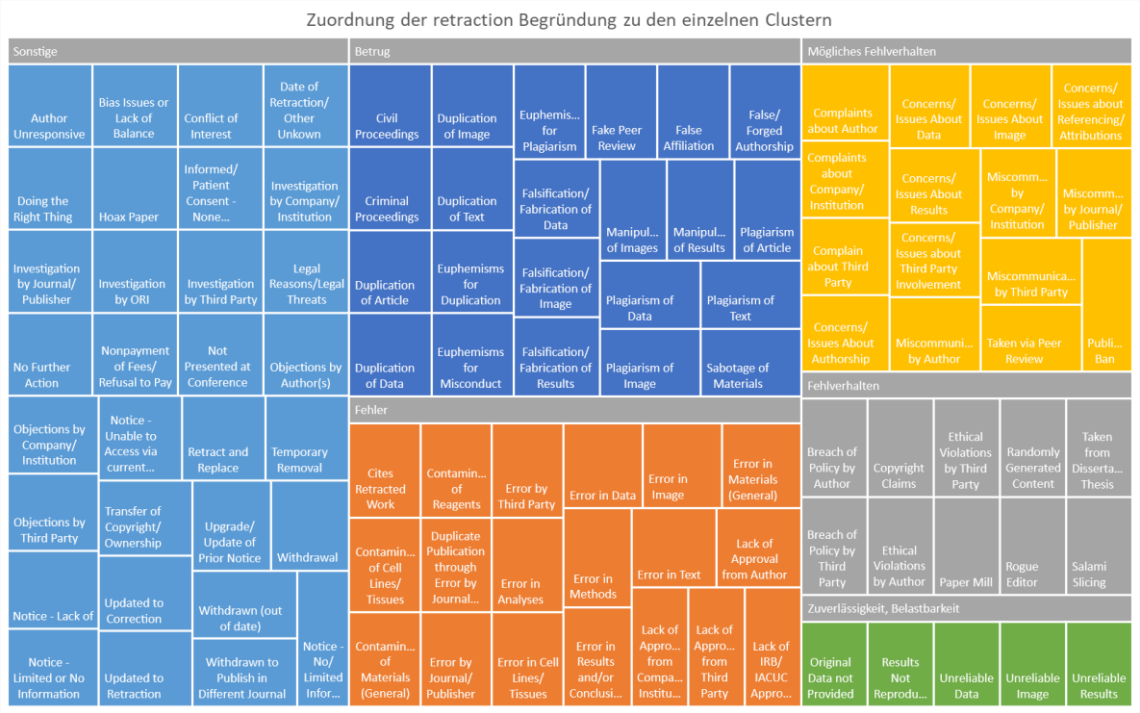

Abbildung 7: Cluster zur Zuordnung der Taxonomie zu einem einzelnen Grund

Für die Berechnung der Zeit zwischen der Veröffentlichung des Originalartikels und der Retraction wurden nur Einträge mit gültigen Datumsangaben für beide Zeitpunkte verwendet.

6.2 Auswertung

An den 14.206 Einträgen sind mehr als 56.000 Autoren aus 139 Ländern beteiligt. Der Artikel mit den meisten Autoren hat 41 Autoren (DOI 10.1007/s00439-017-1759-x). Die größte transnationale Kooperation umfasst Autoren aus 15 Ländern (DOI 10.1155/2011/246412). Es sind 371 Herausgeber und 3.046 Zeitschriften enthalten.

Der älteste Artikel im Datensatz stammt aus dem Jahr 1959. Bis Mitte der 90er Jahre ist die Anzahl mit weniger als 50 Retractions pro Jahr noch sehr gering. Danach ist ein stetiger Anstieg zu verzeichnen, der im Jahr 2020 mit 1.239 Einträgen seinen Höhepunkt erreicht.

Gemessen an der Gesamtheit aller erhobenen *DOI* aus dem *LIVIVO-Datensatz* wurden mit knapp 14.000 aus 28.8 Millionen 0,05% als Retraction identifiziert.

Für die Zuordnung zu den oben beschriebenen Clustern ergibt sich die in Tabelle 4 gezeigte Verteilung.

Nr.	Begründung	Anzahl	Anteil in %
1	Betrug	7.444	52,40
2	Fehlverhalten	1.128	7,94
3	Mögliches Fehlverhalten	1.469	10,34
4	Zuverlässigkeit, Belastbarkeit	1.057	7,44
5	Fehler	1.907	13,42
6	Sonstige	1.201	7,45
	Gesamt	14.206	100,00

Tabelle 4: Zuordnung Retractions zu der Begründung

Über 50% aller identifizierten Retractions mussten aufgrund von Betrug zurückgenommen werden. Knapp 18% aufgrund von Fehlverhalten oder impliziertem Fehlverhalten, 7,44% wurden nach der Prüfung für nicht ausreichend belastbar befunden und bei 13,42% wurden Fehler in den förmlichen Regularien, oder den angewendeten Methoden selbst gefunden. Zu knapp 7% liegen weitere Gründe vor, oder bei der Retraction Notiz wurden keine angegeben.

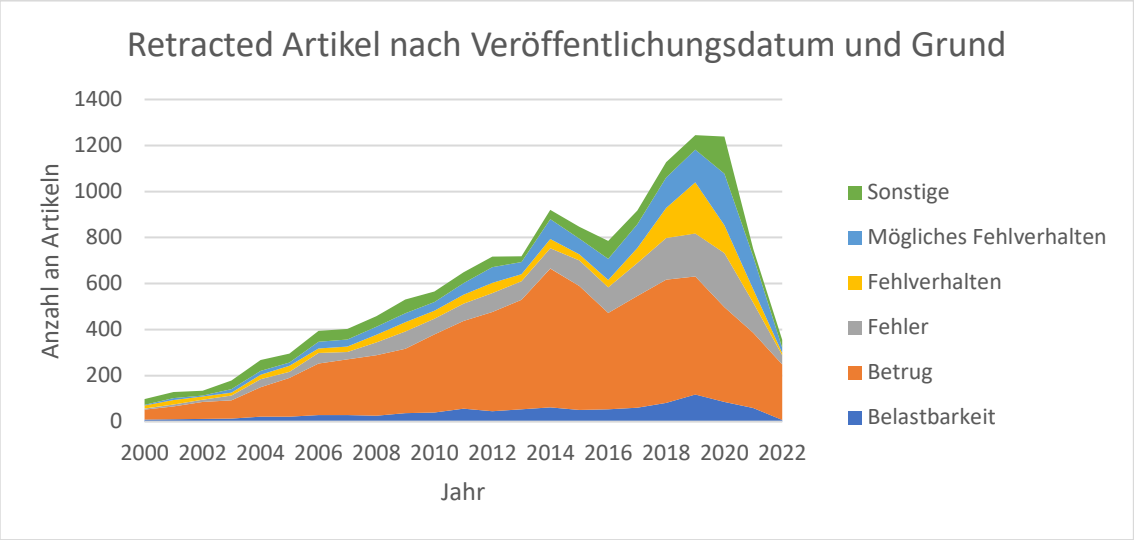

Abbildung 8: Zeitlicher Verlauf von retracted Artikeln ab dem Jahr 2000 aufgeteilt nach der Begründung

Wie in Abbildung 8 gut zu erkennen ist, ist ein stetiger Anstieg der Retracements zu verzeichnen, wobei ein Großteil auf die Kategorie Betrug entfällt. Für die Jahre 2019 und 2020 liegen die Gesamtzahlen bei knapp über 1.200 Artikeln. Nach dem Jahr 2020 gehen die Mengen stark zurück. Dies ist auf die wissenschaftliche Arbeit und den Umgang der Verlage und Zeitschriften mit Retractions zurückzuführen. Nachdem ein potenzielles Problem mit einem Artikel erkannt wurde, werden institutionelle Überprüfungen eingeleitet. Solche Überprüfungen sind langwierige und schwierige Prozesse (Loadsman 2019). Nach der Prüfung und der Feststellung, dass ein Problem vorliegt und Konsequenzen folgen müssen, sollten Verlage und Zeitschriften schnell handeln. Die Kommunikation sollte nicht länger dauern als die Veröffentlichung eines neuen Artikels, aber es gibt große Zeitunterschiede. Studien zeigen, dass PubMed etwa drei Jahre benötigt, um Retractions für ein bestimmtes Jahr korrekt zu indexieren (Decullier et al. 2014). Die Zeitspanne zwischen der Veröffentlichung des Originalartikels und der Retraction Notiz im vorliegenden Datensatz spiegelt diesen Sachverhalt ebenfalls wider. Im Durchschnitt beträgt diese Zeitspanne fast 3,5 Jahre (1258 Tage) und im Median knapp 2 Jahre (732 Tage). Die Anzahl der Rücknahmen für die Jahre ab 2020 wird daher in den nächsten Jahren aller Voraussicht nach noch deutlich zunehmen.

Die Zunahme an Retractions über die letzten 20 Jahre lässt sich auf ein verändertes Verhalten von Autoren und Herausgebern zurückführen. Eine Studie von Steen RG et al. zeigt Gründe für den Anstieg von retracted Artikeln auf. Die Anzahl an Publikationen nimmt stark zu, einhergehend dazu erfolgen vermehrt Retractions. Die Geschwindigkeit, mit der Retractions durchgeführt werden nimmt deutlich zu. Es werden vermehrt Erstvergehen von Autoren festgestellt und das Feststellen von Betrug führt zu einer Neubewertung aller Publikationen eines Autors. (Steen et al. 2013)

Die Zunahme der Publikationen in den letzten Jahren ist auch auf eine Veränderung der Publikationsart zurückzuführen. Vermehrt werden Artikel in Dokumentenservern für Preprints, wie beispielsweise *arXiv* und *bioRxiv* veröffentlicht. Am Beispiel von *arXiv* ist ein Anstieg von 2.365 Artikeln im Januar 2000 auf 17.271 Artikel im November 2022 zu erkennen. (arXiv 2023) Sogenannte Mega-Journale, wie beispielsweise *PLoS One* und *Nature's Scientific Reports*, veröffentlichen Artikel in hoher Frequenz. Sie veröffentlichen peer-reviewed und open-access Artikel, dabei zeichnet Sie eine besonders hohe Annahmequote von >50% aus und der Prozess bis zur Veröffentlichung ist mit 3-5 Monaten signifikant kürzer als bei anderen Verlagen. (Björk 2015)

Diesen Innovationen folgen Betrugstechniken. Es werden systematisch Strukturen und Vorgehensweisen geschaffen, um Artikel mit möglichst geringem Aufwand zu publizieren. Es entstehen so genannte Papiermühlen. Firmen, die auf Bestellung gefälschte wissenschaftliche Manuskripte herstellen (Else und van Noorden 2021). Eine weitere Methode ist das "Salami-Slicing", bei dem Autoren die Ergebnisse eines Artikels bewusst auf mehrere Publikationen verteilen (Tolsgaard et al. 2019).

Bei der Autorenschaft von Artikeln ist ebenfalls ein signifikanter Wandel zu beobachten: Immer mehr Artikel werden von Hunderten oder gar Tausenden von Autoren verfasst. Dieses Phänomen wird als Hyperautorenschaft bezeichnet. (Cronin 2001) In diesem Datensatz liegt ein Artikel mit 41 Autoren vor.

Zusammenfassend lässt sich sagen, dass die Anzahl an Publikationen und Retractions in den letzten 20 Jahren signifikant zugenommen hat und die identifizierten Einträge im *LIVIVO*-Datensatz den allgemein vorherrschenden Mustern in der wissenschaftlichen Gemeinschaft entsprechen.

Nr.	Autor	Anzahl	Anteil an allen Retractions in %
1	Yoshitaka Fujii	136	0,96
2	Joachim Boldt	129	0,91
3	Hironobu Ueshima	119	0,84
4	Hiroshi Otake	102	0,72
5	Hidenori Toyooka	82	0,58

Nr.	Autor	Anzahl	Anteil an allen Retractions in %
	Gesamt	568	4,00

Tabelle 5: Top 5 Autoren nach Anzahl an Retractions

Die Top 5 der am häufigsten beteiligten Autoren (Tabelle 5) repräsentiert mit insgesamt 568 Retractions 4% aller identifizierten Publikationen. Der am häufigsten beteiligte Autor ist *Yoshitaka Fujii* mit 136 Retractions, dicht gefolgt von Joachim Boldt mit 129 Beteiligungen. *Fujii* veröffentlichte die Artikel im Zeitraum von 1993 bis 2011, die erste Retraction erfolgte im Jahr 2011 und die Letzte 2019. *Fujii* ist ein japanischer Anästhesiologe und vermutlich der Forscher mit den meisten Retractions weltweit. Insgesamt wurden 190 seiner in Summe 250 Veröffentlichungen retracted. Der vorwiegende Grund war die Fälschung von Daten. (Dr Geoff 2017) In der Zuordnung zu den verschiedenen Clustern wurden von den 136 gefundenen Artikeln 119 aufgrund von Betrug retracted, 16 aufgrund von Fehlverhalten und 1 Artikel aufgrund von unzureichender Belastbarkeit.

Joachim Boldt hat fast genauso viele Retractions, wie *Fujii*, jedoch setzen sich die 129 Artikel aus anderen Begründungen zusammen. In 101 Fällen wurden Artikel wegen Fehlverhaltens retracted, zumeist wegen unethischer Verstöße. Die erste Retraction erfolgte im Jahr 2010 und die bisher letzte im Oktober 2022.

Nr.	Land	Anzahl	Anteil an allen Retractions in %
1	China	5.322	37,46
2	USA	3.163	22,27
3	Japan	940	6,62
4	Indien	913	6,43
5	Großbritannien	790	5,56
	Gesamt	11.128	78,33

Tabelle 6: Top 5 Länder nach Anzahl an Retractions

Wie Tabelle 6 zeigt, stellen die Top 5 Länder mit China, USA, Japan, Indien und Großbritannien Beteiligungen an insgesamt über 11.000 Artikeln und somit an circa 78% aller im Datenset identifizierten Retractions dar. Ein Grund für diese Verteilung zeigt ein Blick auf die jährlichen Veröffentlichungen der aufgeführten Länder. Laut Auswertungen von SCImago ist China seit 2019 das Land, welches die meisten zitierbaren Dokumente veröffentlicht hat und berichtet für das Jahr 2021 von über 800.000 Publikationen. Die zweitmeisten Veröffentlichungen kommen aus den USA (649.043) und auf Rang drei folgt Indien mit 219.625 Stück. Großbritannien liegt auf Platz 4 (213.389) und Japan auf Platz 7 (135.097). (SCImago 2023)

Wie aus der Studie von Steen RG et al. vorgestellt, herrscht ein Zusammenhang zwischen der Menge an Publikationen und Retractions, jedoch lässt sich aufgrund der vorliegenden Daten diese These nicht ausreichend bewerten.

Nr.	Herausgeber/ Verlag	Anzahl Zeitschriften	Anzahl Retractions	Anteil an allen Retractions in %
1	Elsevier	684	2.523	17,76
2	Wiley	370	1.804	12,70
3	Springer	384	1.165	8,20
4	PLoS	7	584	4,11
5	Taylor and Francis	176	571	4,02
	Gesamt	1.621	6.647	46,79

Tabelle 7: Top 5 Herausgeber nach Anzahl an Retractions und Anzahl an unterliegenden Zeitschriften

Wie in Tabelle 7 gut zu erkennen werden 46,79% aller identifizierten Retractions von den Top 5 Herausgebern publiziert. In Summe liegen 6.647 Retractions aufgeteilt auf 1.621 Zeitschriften vor. Gemessen an der Anzahl an unterliegenden Zeitschriften sind mit *Elsevier*, *Wiley*, *Springer* und *Taylor and Francis* die vier weltweit größten Herausgeber vertreten. Nach einer Analyse (mithilfe von webscraping) von *Andreas Nishikawa-Pacher* ist *Springer* mit 3.763 Zeitschriften der größte Verlag der Welt, gefolgt von *Taylor and Francis* (2.912), *Elsevier* (2.674) und *Wiley* (1.691) (Nishikawa-Pacher 2022). *PLoS* gibt in Summe 12 Zeitschriften heraus.

Fast 94% (548 Stück) der Retractions von *PLoS* stammen aus der *PLoS ONE* Zeitschrift, dem bereits vorgestellten Megajournal. Die Zeitschrift bildet eine Vielzahl von Fachrichtungen ab und überprüft mögliche Publikationen nur nach ihrer wissenschaftlichen und methodischen Qualität, jedoch nicht nach Kriterien wie Originalität oder Neuartigkeit (Björk 2015). Bei den anderen Verlagen ist kein solch gravierender Ausreißer zu erkennen.

7 Diskussion der Ergebnisse

Im Folgenden werden die erzielten Ergebnisse und die aufgeworfenen Forschungsfragen diskutiert und reflektiert. Dabei werden die verwendeten Ansätze und die aufgetretenen Probleme kritisch hinterfragt und Erkenntnisse bzw. Optimierungsvorschläge aufgezeigt. Abschließend werden weitere mögliche Ansätze zur Anreicherung der Metadaten vorgestellt.

Zu F1: Die aufgestellte Pipeline zum Abgleich der Metadaten hat gut ineinandergegriffen und es ist gelungen ein eindeutiges Ergebnis mit 14.206 identifizierten Retractions zu erhalten. Aufgrund der gewaltigen Größe des *LIVIVO*-Datensatzes mit über 190GB wurde ein extra Schritt zum Extrahieren der DOI und dem Umwandeln in das CSV-Format vorgenommen. Dieser Schritt wurde aufgrund von Performanz Problemen auf der virtuellen Maschine so ausgestaltet. Mithilfe des Python Skriptes wurde zunächst probiert noch weitere Metadaten, wie das Erscheinungsjahr und das Land aus dem Datenkorpus zu entnehmen. Die iterative Suche nach den Treffern verlief problemlos, jedoch konnte die Ergebnisdatei nicht generiert werden und der Prozessor brach in diesem Schritt ab. Aus diesem Grund wurde die Extraktion auf das nötigste, also die DOI, beschränkt. Um den Datenkorpus von *LIVIVO* vollumfänglich nutzen zu können, ist eine direkte Verarbeitung des JSON Lines-Formates empfehlenswert, jedoch durch die Größe und die verschachtelte Struktur der Objekte eine Herausforderung. *PostgreSQL* stellt die Möglichkeit bereit das JSON-Format in eine Datenbank zu laden. Hierzu wird eine Tabellenstruktur mit dem Datenformat JSON analog zu den Schlüssel-Wert-Paaren der Ausgangsdatei benötigt. *PostgreSQL* bietet ebenfalls JSON spezifische query Optionen für die spätere Verarbeitung an (PostgreSQL Documentation 2018). Durch das erfolgreiche Einlesen der gesamten Daten könnte die Pipeline verschlankt werden und der ETL-Prozess für beide Datensets in *PostgreSQL* abgebildet werden.

Die Verarbeitung des Datensatzes von *Retraction Watch* war durch die deutlich kleinere Größe und das gelieferte Format gut zu handhaben. Innerhalb von *PostgreSQL* lassen sich Daten leicht anpassen und miteinander verknüpfen. Die Zusammenführung der beiden Korpusse war mithilfe eines *INNER JOINS* der vorliegenden DOI möglich. Als weiteren eindeutigen Identifikator liegt die *PubMedID* vor. Ein Abgleich der Schnittmenge über diese ID könnte noch ergänzende Treffer liefern, da zu manchen Einträgen eine *PubMedID* vorliegt, jedoch keine DOI und umgekehrt.

Zu F2: Die Schnittmenge der beiden Datensätze beträgt 14.206 Einträge. Durch die übernommenen Ausprägungen zur Begründung, den Verlagen und Erscheinungsjahren konnte eine ausführliche deskriptive Beschreibung erfolgen. Es konnte eine eindeutige Steigerung der Anzahl an Retractions über die Jahre festgestellt werden und grob in den

wissenschaftlichen Kontext eingeordnet werden. Durch das Fehlen der Informationen zur Gesamtheit der *LIVIVO*-Daten, war die Auswertung in Relation zwischen Retractions und Gesamtanzahl an vorliegenden Publikationen leider nicht möglich. Um diese Auswertungen möglich zu machen wäre eine Erweiterung der *PostgreSQL* Datenbank nötig. Wie bereits oben beschrieben, könnte dies durch das vollständige Einlesen der *LIVIVO* Daten in die Datenbankstruktur möglich gemacht werden.

Zu F3: Als Empfehlung zur Erweiterung der Metadaten von *LIVIVO* werden, in Anlehnung an die vorgestellten Richtlinien von *COPE,* folgende Ausprägungen, die aus dem *Retraction Watch* Datensatz entnommen werden konnten, empfohlen:

1. Der DOI unter dem die Retraction-Notiz veröffentlich wurde.
2. Das Datum an dem die Retraction-Notiz veröffentlicht wurde.
3. Die *PubMedID* unter der die Retraction-Notiz indexiert wurde.

Des Weiteren wird die Schaffung eines booleschen Wertes empfohlen, wie bereits in Kapitel 5.3 vorgestellt wurde. Dieser entspricht der Wahrheit, sollte der Eintrag retracted worden sein. Optionaler weise sollte eine Referenz in Form eines Links zu der Retraction Notiz erfolgen. Am Beispiel der DOI 10.3892/mmr.2016.5534 ist zu erkennen, dass die referenzierende Retraction Notiz unter der DOI 10.3892/mmr.2022.12643 in *LIVIVO* bereits vorhanden ist. Hierfür wären fortführende Untersuchungen nötig, um die Verfügbarkeit zwischen dem Originalartikel und der Notiz herauszuarbeiten. Geführt werden die Retraction Notizen in *LIVIVO* unter dem Dokumententyp *Retraction of Publication (Yang et al. 2022).*

Aufsetzend auf den Erkenntnissen durch den Abgleich zwischen den beiden Datensätzen können weitere Datenquellen verglichen werden. Als mögliche open source verfügbare Quellen kommen *OpenAlex* und *MEDLINE* in Frage.
OpenAlex bietet innerhalb der API-Dokumentation bereits umfangreiche Informationen zum Erheben und Verarbeiten der von ihnen bereitgestellten Daten. Es werden Snapshots der Daten im JSON-Format angeboten und es besteht eine vollumfängliche Dokumentation. In dieser wird vom Download, über das Umwandeln der JSON in eine CSV Datei, bis hin zum Einlesen in eine *PostgreSQL* Datenbankstruktur alles aufgezeigt (OpenAlex 2023b). Da ebenfalls eine *PostgreSQL* Datenbank genutzt wird, wäre eine Zusammenführung mit geringerem Aufwand möglich.
Die *MEDical Literature Analysis and Retrieval System OnLINE,* kurz *MEDLINE* genannte Datenbank ist bereits zu weiten Teilen in *LIVIVO* integriert und kann nach retracted Artikeln durchsucht werden (ZB MED - LIVIVO 2023). Wie am Beispiel von *Yang et al*. zu erkennen ist, sind veröffentlichte Retraction Notizen enthalten und bekannte retracted Artikel werden unter dem Dokumententyp *Retracted Publication* geführt. Ein Abgleich

dieser Informationen könnte die Menge an identifizierten Retractions noch weiter erhöhen.

8 Zusammenfassung und Fazit

Zusammenfassend kann gesagt werden, dass die erste Identifikation von Retractions im *LIVIVO*-Datensatz einen guten Überblick über die Situation und das mögliche weitere Vorgehen gibt. Mit 14.206 von über 28,8 Millionen Einträgen (0,05 %) wurde eine erste Größenordnung für die Anzahl der Retractions geschaffen, die in weiteren Projekten mit anderen Datenquellen erweitert werden kann.

Durch den Abgleich mit weiteren Datenquellen können sukzessive die in *LIVIVO* vorhandenen Retractions identifiziert und ausgewiesen werden. Durch die Erweiterung um weitere eindeutige Identifikatoren, wie z.B. die PubMedID, kann der Datenkorpus vergrößert werden und es ergeben sich neue Möglichkeiten, diesen mit anderen Quellen abzugleichen.

Literaturverzeichnis

Alison Mccook (2018): What a massive database of retracted papers reveals about science publishing‘s ‘death penalty‘. Online verfügbar unter https://www.science.org/content/article/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty, zuletzt aktualisiert am 07.02.2023, zuletzt geprüft am 07.02.2023.

arXiv (2023): Monthly Submissions. Online verfügbar unter https://arxiv.org/stats/monthly_submissions, zuletzt aktualisiert am 21.02.2023, zuletzt geprüft am 21.02.2023.

Barbour, Virginia; Kleinert, Sabine; Wager, Elizabeth; Yentis, Steven (2009): Guidelines for retracting articles.

Björk, Bo-Christer (2015): Have the „mega-journals“ reached the limits to growth? In: *PeerJ* 3, e981. DOI: 10.7717/peerj.981.

Cronin, Blaise (2001): Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? In: *J. Am. Soc. Inf. Sci.* 52 (7), S. 558–569. DOI: 10.1002/asi.1097.

de.NBI Cloud Portal (2023): de.NBI Cloud - SimpleVM. Unter Mitarbeit von Viktor Rudko David Weinholz Ewgenij Katchko. Online verfügbar unter https://cloud.denbi.de/about/project-types/simplevm/, zuletzt aktualisiert am 09.02.2023, zuletzt geprüft am 09.02.2023.

Decullier, Evelyne; Huot, Laure; Maisonneuve, Hervé (2014): What time-lag for a retraction search on PubMed? In: *BMC research notes* 7, S. 395. DOI: 10.1186/1756-0500-7-395.

Dr Geoff (2017): Yoshitaka Fujii – Japanese anaesthetist and record breaking research fraud. Online verfügbar unter https://drgeoffnutrition.wordpress.com/2017/04/05/yoshitaka-fujii-japanese-anaesthetist-and-record-breaking-research-fraud/, zuletzt aktualisiert am 11.08.2017, zuletzt geprüft am 21.02.2023.

Else, Holly; van Noorden, Richard (2021): The fight against fake-paper factories that churn out sham science. In: *Nature* 591 (7851), S. 516–519. DOI: 10.1038/d41586-021-00733-5.

Grieneisen, Michael L.; Zhang, Minghua (2012): A comprehensive survey of retracted articles from the scholarly literature. In: *PloS one* 7 (10), e44118. DOI: 10.1371/journal.pone.0044118.

jsonlines.org (2023): JSON Lines. Online verfügbar unter https://jsonlines.org/, zuletzt aktualisiert am 08.02.2023, zuletzt geprüft am 08.02.2023.

Lancet, The Editors of The (2010): Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. In: *The Lancet* 375 (9713), S. 445. DOI: 10.1016/S0140-6736(10)60175-4.

Loadsman, J. A. (2019): Why does retraction take so much longer than publication? In: *Anaesthesia* 74 (1), S. 3–5. DOI: 10.1111/anae.14484.

Nishikawa-Pacher, Andreas (2022): Who are the 100 largest scientific publishers by journal count? A webscraping approach. In: *JD* 78 (7), S. 450–463. DOI: 10.1108/JD-04-2022-0083.

OpenAlex (2023a): OpenAlex Documentation. Work object. Online verfügbar unter https://docs.openalex.org/api-entities/works/work-object#is_retracted, zuletzt aktualisiert am 13.02.2023, zuletzt geprüft am 13.02.2023.

OpenAlex (2023b): Load to a relational database. Online verfügbar unter https://docs.openalex.org/download-all-data/upload-to-your-database/load-to-a-relational-database, zuletzt aktualisiert am 23.02.2023, zuletzt geprüft am 23.02.2023.

Ponniah, Paulraj (2001): Data warehousing fundamentals. A comprehensive guide for IT professionals. Chichester, Weinheim: Wiley. Online verfügbar unter http://www.loc.gov/catdir/bios/wiley043/2001024243.html.

PostgreSQL Documentation (2011): Date/Time Types. Online verfügbar unter https://www.postgresql.org/docs/8.2/datatype-datetime.html, zuletzt aktualisiert am 13.02.2023, zuletzt geprüft am 13.02.2023.

PostgreSQL Documentation (2018): JSON Functions and Operators. Online verfügbar unter https://www.postgresql.org/docs/9.3/functions-json.html, zuletzt aktualisiert am 23.02.2023, zuletzt geprüft am 23.02.2023.

Retraction Watch (2010): Why write a blog about retractions? Online verfügbar unter https://retractionwatch.com/2010/08/03/why-write-a-blog-about-retractions/, zuletzt aktualisiert am 27.10.2015, zuletzt geprüft am 03.02.2023.

Retraction Watch (2018): We’re officially launching our database today. Here’s what you need to know. Online verfügbar unter https://retractionwatch.com/2018/10/25/were-officially-launching-our-database-today-heres-what-you-need-to-know/, zuletzt aktualisiert am 25.10.2018, zuletzt geprüft am 03.02.2023.

SCImago (2023): Country Rankings. Online verfügbar unter https://www.scimagojr.com/countryrank.php?order=itp&ord=desc&year=2021, zuletzt aktualisiert am 21.02.2023, zuletzt geprüft am 21.02.2023.

Steen, R. Grant; Casadevall, Arturo; Fang, Ferric C. (2013): Why has the number of scientific retractions increased? In: *PloS one* 8 (7), e68397. DOI: 10.1371/journal.pone.0068397.

The Center for Scientific Integrity (2018): The Retraction Watch Database. New York (ISSN: 2692-465X). Online verfügbar unter http://retractiondatabase.org/, zuletzt geprüft am 03.02.2023.

Tolsgaard, Martin G.; Ellaway, Rachel; Woods, Nikki; Norman, Geoff (2019): Salami-slicing and plagiarism: How should we respond? In: *Advances in health sciences education : theory and practice* 24 (1), S. 3–14. DOI: 10.1007/s10459-019-09876-7.

Wakefield, A. J.; Murch, S. H.; Anthony, A.; Linnell, J.; Casson, D. M.; Malik, M. et al. (1998): RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. In: *The Lancet* 351 (9103), S. 637–641. DOI: 10.1016/S0140-6736(97)11096-0.

Wright, Kath; McDaid, Catriona (2011): Reporting of article retractions in bibliographic databases and online journals. In: *Journal of the Medical Library Association : JMLA* 99 (2), S. 164–167. DOI: 10.3163/1536-5050.99.2.010.

Yang, Xinhua; Liu, Jinge; Long, Yongqi; Liu, Wenjin; Chen, Lei; Zhang, Yulong et al. (2022): [Retracted] MicroRNA-101 inhibits the proliferation and invasion of bladder cancer cells via targeting c-FOS. Online verfügbar unter https://www.livivo.de/doc/M35169853.

ZB MED (2023): LIVIVO-Suchportal. Online verfügbar unter https://www.zbmed.de/recherchieren/livivo/, zuletzt aktualisiert am 08.02.2023, zuletzt geprüft am 08.02.2023.

ZB MED - LIVIVO (2023): LIVIVO - Data base informations. Online verfügbar unter https://www.livivo.de/app/misc/dbinfo?dbid=MEDLINE, zuletzt aktualisiert am 23.02.2023, zuletzt geprüft am 23.02.2023.

Anhang

Die Dokumentation der aufgestellten Pipeline ist verfügbar unter:

GitHub Repository LIVIVO_RW

https://github.com/Vetterino/LIVIVO_RW