WIRTSCHAFTSUNIVERSITÄT WIEN
Vienna University of Economics and Business

Bachelorarbeit

Deutscher Titel der Bachelorarbeit	Qualität und Kompatibilität von Lizenzinformationen in offenen Datenportalen
Englischer Titel der Bachelorarbeit	Quality and Compatibility of License Information in Open Data portals
Verfasser/in **Familienname, Vorname(n)**	Blaim Mattias
Matrikelnummer	1050802
Studium	Bachelorstudium Wirtschafts- und Sozialwissenschaften
Beurteiler/in **Titel, Vorname(n), Familienname**	Dr. Axel Polleres, Dr. Jürgen Umbrich

Hiermit versichere ich, dass

1. ich die vorliegende Bachelorarbeit selbständig und ohne Verwendung unerlaubter Hilfsmittel verfasst habe. Alle Inhalte, die direkt oder indirekt aus fremden Quellen entnommen sind, sind durch entsprechende Quellenangaben gekennzeichnet.

2. die vorliegende Arbeit bisher weder im In- noch im Ausland zur Beurteilung vorgelegt bzw. veröffentlicht worden ist.

3. diese Arbeit mit der beurteilten bzw. in elektronischer Form eingereichten Bachelorarbeit übereinstimmt.

4. (nur bei Gruppenarbeiten): die vorliegende Arbeit gemeinsam mit

   entstanden ist. Die Teilleistungen der einzelnen Personen sind kenntlich gemacht, ebenso wie jene Passagen, die gemeinsam erarbeitet wurden.

Datum 20/12/2014

Unterschrift

Zusammenfassung

Das OpenData@WU-Projekt beschäftigt sich mit der Qualität von offenen Datenportalen im Internet. Die Projektmitglieder haben Zugang zu 92 Portalen, die frei zur Verfügung stehende Datensätze verwalten und versuchen, qualitätsfördernde Maßnahmen zu ergreifen. Zwei Problemfelder werden in dieser Arbeit aufgezeigt und versucht, diese zu lösen. Erstens die eindeutige Identifizierung von Lizenzen durch ihre Metadaten und zweitens die Bestimmung der Kompatibilität zwischen verschiedenen Lizenzmodellen. Die von den Projektmitgliedern überlassenen Daten wurden im Zuge dieser Arbeit aufgearbeitet und analysiert, um daraus Heuristiken aufzustellen, die den zwei qualitätsmindernden Problemen entgegenwirken sollen.

Abstract

The OpenData@WU project is dealing with the quality of Open Data portals on the internet. Its project members have direct acess to 92 Open Data websites, which offer free data sets available for any prospective user and try to improve their quality on managing the data sets published on them. There are two problems mentioned in this work. First the explicit identification of the license a data set is published and second the compatibility amongst two lincenses. Therefore, the project members provide us with data, which will be analyzed within this work. Consequently, we draw up some heuristics, which should help to improve the quality of the open data websites.

Vorwort

Seit geraumer Zeit, beschäftige ich mich mit dem Thema Open Data. Dieser Forschungsbereich hat mich von Anfang an fasziniert. So kam es, dass ich eine wissenschaftliche Arbeit darüber schreiben wollte. Das spezifische Thema legten mir Professor Dr. Axel Polleres und Dr. Jürgen Umbrich ans Herz. Ziel war es, eine Arbeit zu verfassen, die Heuristiken hervorbringen soll. Diese sollen dann im Rahmen des OpenData@WU-Projekts evaluiert und eingesetzt werden. Ich habe mich dieser Aufgabe sofort angenommen und einige sehr interessante Aspekte herausgefunden.

Ich möchte dieses Vorwort nutzen, um ein paar Personen zu danken. Vor Allem gilt mein Dank meinen zwei Betreuern, die mir stets Hilfestellung gegeben haben. Sie waren immer erreichbar und haben mich unterstützt, wo immer sie konnten. Außerdem will ich meiner Familie und meine Freunden danken, die mich mein ganzes Leben lang unterstützt haben und ohne die ich diese Arbeit niemals hätte schreiben können.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

In den letzten Jahren hat der Begriff „Open Data“ immer mehr an Bedeutung gewonnen. Regierungen, öffentliche Institutionen und Organisationen stellen ihre gewonnenen Daten öffentlich zur Verfügung um Verbesserungen in verschiedenen Lebensbereichen zu erwirken. So benutzen beispielsweise Regierungen, diese von ihnen öffentlich zur Verfügung gestellten Daten, um mehr Transparenz bezüglich ihrer Aktivitäten für ihre Bevölkerung zu gewährleisten. Aber auch private und öffentliche Services sollen durch die Bereitstellung öffentlich zugänglicher Daten verbessert werden, indem sie helfen Entscheidungen zu treffen und Planungsaufgaben unterstützen.

Grundsätzlich wird die Schwierigkeit um an Daten zu kommen durch Open Data vermindert, da diese über online Portale öffentlich zugänglich gemacht werden. Auch wenn diese öffentlich zur Verfügung gestellt werden und somit von jedem verwendet werden können, gilt auch hier weiterhin das Urheberrecht. Die Urheber von Datensätzen erlauben der Allgemeinheit, ihre Daten zu verwenden und können vorgeben, was mit ihren Datensätzen gemacht werden darf.

Hierfür werden sogenannte Lizenzen verwendet, durch die auf die Rechte eines Urhebers verzichtet werden kann. Um zu gewährleisten, dass diese Daten gemäß der Open Definition auch wirklich „Open Data“ sind, muss die Lizenz den Kriterien der Offenheit genügen. Da diese Lizenzmodelle nicht zentralisiert sind, gibt es eine Vielzahl an Lizenzen, die oftmals rechtlich dasselbe absichern, jedoch durch ihre verschiedenen Bezeichnungen für Verwirrung sorgen. An dieser Stelle treten oft einige Probleme auf.

Ein Problem ist, dass Lizenzen auf den ersten Blick nicht eindeutig zu identifizieren sind. Dies führt dazu, dass man nicht auf Anhieb weiß wie mit den Daten umgegangen werden darf. Inkompatibilitäten führen zu weiteren Problemen in der Nutzung von „Open Data Sets“. Oftmals werden mehrere Datensätze für bestimmte Anwendungen verwendet. Wenn diese Datensätze unter verschiedenen Lizenzen veröffentlicht wurden, kann dies zu Inkompatibilitäten führen. So ist es zum Beispiel möglich, dass die Vermischung von zwei Datensätzen nicht durchführbar ist, weil die eine Lizenz kommerzielle Nutzung verbietet, während die andere sie zulässt.

Genau diesen Problemen soll in dieser wissenschaftlichen Arbeit Abhilfe geschaffen werden, indem folgende Ziele erreicht werden:

1. Das Aufstellen von Heuristiken, die dabei helfen sollen Lizenzen eindeutig zu identifizieren und Kompatibilitäten zu finden, um das Arbeiten mit solchen „Open Data Sets“ zu erleichtern.

2. Eine Anwendung zu schreiben, die die aufgestellten Heuristiken verwendet und die eindeutige Identifizierung dieser Lizenzen übernehmen soll.

Fasst man diese Ziele zusammen, kann man folgende Forschungsfrage formulieren:

Welche Heuristiken können dabei helfen, Lizenzen, in Bezug auf Open Data, eindeutig zu identifizieren und Kompatibilitäten zu finden?

Diese Forschungsfrage wird in weiterer Folge Schritt für Schritt beantwortet, um am Ende vollständige Heuristiken entstehen, die angewendet werden können.

2 Theoretische Grundlagen

Um das weitere Vorgehen zu verstehen, werden jetzt einige theoretische Grundlagen aufgearbeitet. Diese sind notwendig, um die einzelnen Schritte, die in der Arbeit beschrieben werden, zu verstehen.

2.1 Open Data

Wie schon Anfangs erwähnt werden „Open Data“ immer häufiger verwendet. Diese Art von Datensätzen wird der Öffentlichkeit zur Verfügung gestellt, um Verbesserungen in verschiedenen Lebensbereichen zu erreichen. Initiiert wurde diese Bewegung von Regierungen, die erstmals Daten, die in keinen Zusammenhang mit einzelnen Personen gebracht werden können, über ein online Portal[1] veröffentlichten und die Verwendung dieser für bestimmte Zwecke erlaubten. Immer mehr Regierungen, öffentliche Institutionen und Organisationen sind diesem Trend nachgegangen und haben ebenfalls Initiativen gestartet um Datensätze zu erstellen und öffentlich zu machen. Heute gibt es eine Vielzahl an „Open Data Sets“, die frei zugänglich sind und auf verschiedenste Art verwendet werden können.

2.1.1 Definition

Um Datensätze als offen zu bezeichnen, muss folgende Definition herangezogen werden:

> “A piece of data or content is open if anyone is free to use, reuse, and redistribute it.” [5]

[1] z.B. `https://www.data.gv.at/`, `http://data.gov.uk/`

Entsprechen Daten dieser Definition, kann man sie zurecht als „Open Data“ bezeichnen. Die Hauptattribute von „Offenheit“ sind folgende:

1. Verfügbarkeit und Zugang
2. Nachnutzung und Weiterverbreitung
3. Universale Beteiligung

Diese Attribute werden dann, laut der vollen Definition von der Open Knowledge Foundation, in detailliertere Kriterien unterteilt und beschrieben, um mit Sicherheit feststellen zu können, dass es sich um Open Data handelt.

2.1.2 Kriterien

Nun gehen wir auf die einzelnen Kriterien und deren Beschreibungen ein, die laut der Open Definition [5] erfüllt sein müssen. Die Open Knowledge Foundation hat 11 Kriterien definiert die unbedingt erfüllt werden müssen, damit Daten gemäß der bereits erwähnten Open Definition als „offen“ gelten. Diese lauten:

1. Zugang: „Das Werk soll als Ganzes verfügbar sein, zu Kosten, die nicht höher als die Reproduktionskosten sind.“
2. Weiterverbreitung: „Die Lizenz darf niemanden hindern, das Werk entweder eigenständig oder als Teil einer Sammlung aus verschiedenen Quellen zu verschenken oder zu verkaufen.“
3. Nachnutzung: „Die Lizenz muss Modifikationen oder Derivate erlauben, genauso wie die Weiterverbreitung dieser, unter den selben Lizenzbedingungen des ursprünglichen Werks.“
4. Keine technischen Einschränkungen: „Das Werk soll in einer Form zur Verfügung gestellt werden, die keine technischen Hindernisse für die oben genannten Nutzungen beinhaltet.“
5. Namensnennung: „Die Lizenz kann als Bedingung für Weiterverbreitung und Nachnutzung des Werkes, die Nennung der Namen, seiner Urheber und Mitwirkenden verlangen.“
6. Integrität: „Die Lizenz kann als Bedingung für die Verbreitung des Werkes in modifizierter Form verlangen, dass das Derivat einen anderen Namen oder eine andere Versionsnummer als das ursprüngliche Werk erhält.“
7. Keine Diskriminierung von Personen oder Gruppen: „Die Lizenz darf keine Einzelpersonen oder Personengruppen diskriminieren.“

8. Keine Einschränkung der Einsatzzwecke: „Die Lizenz darf niemanden daran hindern, das Werk zu einem beliebigen Zweck einzusetzen.“

9. Lizenzvergabe: „Die rechtlichen Bedingungen, denen ein Werk unterliegt, müssen bei der Weiterverteilung an alle Empfänger übergehen, ohne dass diese verpflichtet sind, zusätzliche Bedingungen zu akzeptieren.“

10. Die Lizenz darf nicht an eine spezifische Sammlung gebunden sein: „Die rechtlichen Bedingungen, denen ein Werk unterliegt, dürfen nicht davon abhängen, ob das Werk Teil einer spezifischen Sammlung ist.“

11. Die Lizenz darf die Verbreitung anderer Werke nicht einschränken: „Die Lizenz darf anderen Werken, die mit dem lizenzierten Werk gemeinsam weitergegeben werden, keine Beschränkungen auferlegen.“

2.1.3 Verwendung

Nachdem wir in den vorherigen Abschnitten geklärt haben, wie der Begriff Open Data definiert ist und welche genauen Kriterien ein Datensatz erfüllen muss, um „offen“ zu sein, stellen wir uns jetzt die Frage, wofür diese Daten gut sind. Open Data bringt in vielen Bereichen großen Nutzen mit sich. Die wichtigsten werden in folgenden Punkten aufgezählt [7]:

- Transparenz und demokratische Kontrolle
- Innovationen
- Verbesserungen von bestehenden oder Entwicklungen von neuen Produkten oder Services
- Verbesserung der Effizienz beziehungsweise der Effektivität der von der Regierung angebotenen Dienstleistungen
- Generierung von neuem Wissen und Finden von neuen Einblicken

Der wichtigste Punkt, bei dem offene Datensätze helfen können, ist die Veröffentlichung von Regierungsdaten. Gerade in einer gut funktionierenden, demokratischen Gesellschaft, spielt Transparenz von Regierungstätigkeiten eine große Rolle. Hier ist aber nicht nur der freie Zugang wichtig, sondern vor allem auch die Weiterverwendung und Verbreitung dieser Daten. Beispiele von Projekten, die diese Transparenz gewährleisten, sind, das Projekt „open spending[2]“ und speziell in Großbritannien das Projekt

[2] `https://openspending.org/`

„where does my money go[3]“. Diese Projekte zeigen auf, wie die Steuereinnahmen des Landes verwendet werden und wohin diese fließen.

Ein weiterer wichtiger Punkt, auf den öffentliche Regierungsdaten Einfluss nehmen, ist die Bereitstellung von sozialen und kommerziellen Services.

> “Im digitalen Zeitalter, sind Daten die Schlüsselressource für soziale und kommerzielle Aktivitäten.“ [4]

Regierungen machen ihre gesammelten Daten öffentlich und helfen dabei innovative Anwendungen zu kreieren, die Menschen das Leben erleichtern.

In ökonomischer Hinsicht haben diese Datensätze auch eine enorme Wichtigkeit. Es werden ständig neue Produkte unter Verwendung öffentlicher Regierungsdaten entwickelt. Auch Firmen nutzen diese Daten um Informationen zu gewinnen, wie sie ihre Kosten senken können.

Open Data erzeugt auch einen Wert für die Regierungen selbst. So könnten sie zum Beispiel ihre Kosten senken, da durch die Transparenz die Bevölkerung besser informiert ist und deshalb die Anfragen an bestimmte Behörden vermindert werden.

Auch wenn Open Data jetzt schon sozialen und ökonomischen Wert beigetragen hat, kann man nicht sagen was noch alles möglich ist. Durch Verwendung vieler verschiedener Daten und deren Zusammenschließung, kann neues Wissen generiert und neue Einblicke gewährt werden.

2.1.4 Arten

Wie schon des Öfteren erwähnt, gibt es eine große Menge an frei zugänglichen Datensätzen [6]:

- Kultur: Daten über kulturelle Werke und Artefakte wie z.B. Bibliotheksdatenbanken
- Wissenschaft: Daten die aus der Forschung entstehen
- Finanzen: Daten über Ausgaben von Ländern, Informationen über Finanzmärkte
- Statistiken: Daten die von Statistikbehörden erzeugt wurden
- Wetter: Daten aus denen man zukünftige Wetterlagen voraussagen und verstehen kann

[3] `http://wheredoesmymoneygo.org/`

- Umwelt: Daten die die natürliche Umwelt beschreiben wie z.B. Verschmutzung von Seen
- Logistik: Daten über Standorte, Routenberechnung usw.

2.1.5 Metadaten

Metadaten sind Daten, die andere Daten beschreiben. Im Bereich von Open Data kann man sich das so vorstellen, dass es maschinell lesbare Daten gibt, die einen öffentlich zugänglichen Datensatz beschreiben. Häufige Metadaten von Open Data sind Informationen über den Ersteller, Titel, Beschreibung, Art von Daten, Lizenz, Verfügbare Formate und Tags.

2.2 Open Licenses

Um sich dem zentralen Thema dieser Arbeit langsam anzunähern, wenden wir uns jetzt einem Bereich der Metadaten von öffentlichen Datensätzen zu – den Lizenzen. Stellen wir uns nun folgende Fragen:

1. Wozu werden Lizenzen generell verwendet?
2. Warum sollte man frei zugängliche Daten lizenzieren?

Lizenzen werden verwendet, um geistiges Eigentum vor unrechtmäßiger Verwendung zu schützen. Es werden Lizenzen an Personen vergeben, die diese anfordern und dürfen dann das geistige Eigentum so verwenden, wie es in der Lizenz definiert wird. Um die zweite Frage zu beantworten wird als Argument die Klarheit von Verfügungsrechten hervorgehoben. Auch wenn die Daten ohne weiteres verwendet werden dürfen, gilt noch immer das Urheberrecht. Das sogenannte „Copyright“ ist ein weitreichend geltendes Recht und sichert dem Urheber einer Ressource gewisse Rechte zu. Will man jetzt geistiges Eigentum frei zugänglich machen, sollte genau definiert sein, dass mit der Ressource gemäß der Open Definition umgegangen werden darf. Bei Datensätzen ist es dasselbe Prinzip.

> „Damit Daten offen sind, d.h. damit sie gemäß der Open Definition „Open Data“ sind, müssen sie mit einer Lizenz versehen sein, die den Kriterien der Offenheit genügt.“ [8]

Nun gibt es verschiedene Möglichkeiten Daten zu lizenzieren und der Allgemeinheit preiszugeben. Diese teilt man in drei Kategorien ein [3]:

1. Public-Domain-Lizenz: Unter diesem Lizenzmodell können die Daten ohne Einschränkung verwendet werden
2. Attribution-Lizenz: Ist der Public-Domain-Lizenz ähnlich, jedoch muss bei der Nachnutzung der Daten der Name des Urhebers angegeben werden
3. Share-Alike-Lizenz: Diese Lizenz ist eine Erweiterung zur Attribution Lizenz und besagt, dass neben der Namensnennung, bei der Nachnutzung die entstandenen Daten unter derselben Lizenz veröffentlicht werden müssen

2.2.1 Entstehung

Ein Trend, der wesentlich früher aufkam als die Open Data Bewegung, ist die Veröffentlichung von Open Source Software. Diese Art von Software steht frei zur Verfügung und es muss kein Aufpreis für die Verwendung gezahlt werden. Außerdem kann der source code eingesehen und verändert werden, was zu Innovationen, bezüglich neuer Softwarekomponenten führen kann. Da die Open Data Bewegung dem Open Source Konzept sehr ähnlich ist, bauen deren Lizenzmodelle auf der Idee von Open Source Lizenzen auf.

Den Anfang machte die Creative Commons Initiative, die Lizenzmodelle ausarbeitete, die die Rechtslage von geistigen Eigentum, die öffentlich zugänglich gemacht werden, regeln. Auf Basis dieser Lizenzen haben andere Organisationen ihre eigenen Lizenzen erstellt. Es wurden weitere Initiativen gegründet, die Rechtstexte formulieren und daraus Lizenzen generieren. Diese Lizenzen können von jedem verwendet werden und gewährleisten, dass sie ihr geistiges Eigentum rechtlich abgesichert frei zur Verfügung stellen können und trotzdem ihr Urheberrecht behalten. So hat zum Beispiel die Open Knowledge Foundation ein Projekt gestartet, das Lizenzen ausgearbeitet hat, die einzig und allein zur Lizenzierung von öffentlichen Datenbanken verwendet werden.

Aus dem Open Commons Projekt sind noch zwei weitere Lizenzen entstanden, genauso wie aus vielen anderen Projekten und Initiativen neue Lizenzmodelle entstanden sind. Deshalb gibt es heute eine Vielzahl an Lizenzen, die zur Veröffentlichung von Datensätzen verwendet werden können.

2.2.2 Lizenzen für kreative Inhalte

Um kreative Inhalte (Texte, Fotos, Musik, usw.) öffentlich zur Verfügung zu stellen, werden oft Lizenzmodelle von Creative Commons verwendet. Jedoch sind nicht alle Modelle dieser Organisation dafür geeignet, diese als „offen“ zu deklarieren. Es gibt nur drei die mit der „Open Definition“ völlig konform gehen [3]:

- Creative Commons Attribution Lizenz (CC-BY)
- Creative Commons Attribution Share-Alike Lizenz (CC-BY-SA)
- Creative Commons CCZero (CC0)

Diese genannten Lizenzmodelle werden in der Praxis ziemlich häufig verwendet, insbesondere die CC-BY genannte Attributions-Lizenz. Diese erlaubt es den Datensatz zu teilen und zu bearbeiten. Unter Teilen versteht man die Weiterverwendung bzw. die Weiterverbreitung und das Vervielfältigen der Daten. Das Bearbeiten bedeutet die Veränderung des Originaldatensatzes und die Vermischung mit anderen Datensätzen. Jedoch sind unter dieser Lizenz Pflichten angegeben, die eingehalten werden müssen.

Zu Beginn des Kapitels 2.2 wurde schon erwähnt, was bei einer Attribution-Lizenz beachtet werden muss. Wird ein Datensatz verwendet, der unter der CC-BY Lizenz veröffentlicht wurde, muss bei der Nachnutzung der Daten, der Name des Urhebers angegeben werden. Ist die Veröffentlichung mit der CC-BY-SA Lizenz geschehen, kommt eine zusätzliche Pflicht hinzu. Hierbei muss nicht nur der Urheber bei der Nachnutzung angegeben werden, es muss auch der dadurch entstandene Datensatz unter der selben Lizenz, wie der Originaldatensatz, veröffentlicht werden. Ganz anders ist dies beim CC0 Lizenzmodell. Hier werden die Daten der Allgemeinheit zu Verfügung gestellt, ohne das diese Pflichten nachgehen muss. Man hat den Anspruch auf die selben Rechte wie bei der CC-BY oder der CC-BY-SA Lizenz, müssen bei der Nachnutzung jedoch keine weiteren Pflichten beachten.

Immer wieder werden noch weitere Lizenzen von der Creative Commons Initiative verwendet. Diese sind jedoch nicht mit der Open Definition konform, da sie Verbote beinhalten, die die Verwendung der Daten wesentlich einschränken. Hier sind diese aufgezählt:

- Creative Commons Attribution-NoDerives (CC-BY-ND): Bei der Nachnutzung der Daten, die unter dieser Lizenz veröffentlicht wurden, dürfen keine neuen Anwendungen, durch Weiterverwendung dieser, entstehen
- Creative Commons Attribution-NonCommercial (CC-BY-NC): Bei der Nachnutzung der Daten, die unter dieser Lizenz veröffentlicht wurden, dürfen die daraus entstandenen Anwendungen, durch Weiterverwendung dieser, nicht kommerziell zur Verfügung gestellt werden
- Creative Commons Attribution-NonCommercial-NoDerives (CC-BY-NC-ND): Dieses Lizenzmodell ist eine Kombination aus den vorherigen und stellt das höchste Maß an Verboten dar

2.2.3 Lizenzen für Datensätze

Es muss erwähnt werden, dass für Daten genauso gut Creative Commons Lizenzen verwendet werden können, um diese öffentlich zugänglich zu machen. Im Zuge des Open Data Commons Projekts sind jedoch Lizenzen generiert worden, die eigens zur Lizenzierung von Datensätzen verwendet werden. Diese lauten:

- Open Data Commons Public Domain Dedication and License (ODC-PDDL)
- Open Data Commons Attribution License (ODC-BY)
- Open Data Commons Open Database License (ODC-ODbL)

Sieht man sich die Rechtstexte bzw. die Zusammenfassung der Rechte und Pflichten der Lizenzen an, erkennt man, dass sie den Creative Commons Lizenzmodellen sehr ähnlich sind. So ist die ODC-PDDL-Lizenz der CC0-Lizenz gleichzustellen. Das Selbe gilt bei den zwei anderen Lizenzen, die auch gleichgestellt werden können. Die ODC-BY-Lizenz und die CC-BY-Lizenz enthalten die gleichen Rechte und Pflichten, genauso wie die ODC-ODbL-Lizenz und die CC-BY-SA-Lizenz sich sehr ähnlich sind. Grund dafür ist das Ausgangsschema in 2.2. Beide Organisationen haben darauf geachtet, jeweils eine Attribution, eine Share-Alike und eine Public-Domain Lizenz zu erstellen, die das "freie zugänglich machen" von Daten erleichtern soll.

2.2.4 Weitere Lizenzen

Häufig werden auch andere Lizenzen für die Veröffentlichung von Open Data verwendet. Großbritannien hat z.B. ein eigenes Lizenzmodell aufgebaut und stellt dieses der Öffentlichkeit zur Verfügung. Auch Deutschland ist der Bewegung gefolgt und hat eine selbst erstellte Lizenz veröffentlicht, die häufig genannt wird. Um diese zu verbessern, um das Arbeiten mit ihnen zu erleichtern, werden die Rechtstexte häufig überarbeitet. Deshalb gibt es von einer Lizenz oft mehrere Versionen, wobei die neueste Version bei der Veröffentlichung bevorzugt wird. Hier werden einige der gebräuchlichsten Lizenzen aufgezählt:

- Open Government Licence - United Kingdom (OGL-UK)
- Datenlizenz Deutschland Namensnennung (dl-de-by)
- GNU General Public License (GLP)
- Ordnance Survey Open Data Licence (OS-OpenData)

Weitere Lizenzen, die mit der Open Definition konform sind, davon jedoch einige nur noch selten verwendet werden, oder veraltet bzw. stillgelegt sind, werden hier aufgezählt:

- Against DRM (Against-DRM)
- Design Science License (DSL)
- EFF Open Audio License
- Free Art License (FAL)
- GNU Free Documentation License (GNU FDL)
- Higher Education Statistics Agency Copyright with data.gov.uk rights (hesa-withrights)
- Local Authority Copyright with data.gov.uk rights (localauth-withrights)
- MirOS License (MirOS)
- Open Government License - Canada (OGL-Canada)
- Other Attribution (other-at)
- Other Open (other-open)
- Other Public Domain (other-pd)
- Talis Community License (Talis)
- UK Crown Copyright with data.gov.uk rights (ukcrown-withrights)

2.2.5 Kompatibilität von Lizenzen

Um einen weiteren Punkt in dieser Arbeit zu betrachten, muss geklärt werden, was Kompatibilität in Bezug auf Lizenzen bedeutet. Diese ist besonders wichtig bei der Weiterverwendung von offenen Datensätzen. In diesem Szenario sollte man wissen, ob die Lizenzen zweier Datensätze zusammenpassen und kompatibel sind. Will man jetzt zwei Datensätze kombinieren, muss die verwendete Lizenz des einen Datensatzes mit der des Anderen kompatibel sein. Der daraus entstandene Datensatz, muss in weiterer Folge unter einer Lizenz veröffentlicht werden, die keine Probleme mit den vorhergegangenen Lizenzen verursacht.

Man sieht schon, dass hierbei eine Menge Probleme auftreten können. Deshalb ist es besonders hilfreich schon im Vorhinein zu wissen, welche Lizenzen kompatibel sind. Nun

stellen wir uns die Frage, was Kompatibilität überhaupt ist? Hierfür gibt es verschiedene Definitionen. So definiert die Creative Commons Initiative, Lizenzen als kompatibel mit ihren Modellen, wenn

> [...] sie ein minimum an Bedingungen enthalten, die dem selben Zweck dienen, die selbe Bedeutung haben oder den gleichen Effekt haben, wie die Schlüsselelemente einer bestimmten Creative Commons Lizenz [...] [2]

Diese Definition bedeutet, dass verschieden Lizenzmodelle in ihrer Grundausrichtung die selben Absichten haben müssen, um kompatibel zu sein.

Warum stellt sich nun die Frage nach der Kompatibilität? Hier gibt es verschiedene Punkte, die betrachtet werden sollten:

- Lizenzübereinstimmung: Die Autoren einer bestimmten Lizenz wollen sicherstellen, dass bei der Zusammenführung zweier Datensätzen, die Lizenzen miteinander harmonieren
- Lizenzauswahl: Die Ersteller von Datensätzen wollen sicherstellen, dass bei deren Weiterverwendung, die Benutzer höchst mögliche Freiheit haben und mit anderen kombinieren können
- Lizenzwechsel: Die Ersteller von Datensätzen wollen eine Alternative zu anderen Lizenzen haben, um diese bei Bedarf zu wechseln
- Lizenzierung von Weiterverwendungen: Die Ersteller von Arbeiten müssen verstehen, wie sie deren Arbeiten lizenzieren und verbreiten können, wenn sie Datensätze weiterverwenden

Man sieht, dass das Wissen über die Kompatibilität von Lizenzen, in Hinblick auf das Arbeiten mit offenen Datensätzen, von großem Nutzen sein kann. Weitere Grundlagen werden in 4.3 besprochen.

3 Eindeutige Identifizierung von Lizenzen

Da nun jetzt ausreichend über die theoretischen Grundlagen dieses Forschungsgebiets informiert wurde, beginnt hier der zentrale Punkt dieser Arbeit, nämlich die eindeutige Identifikation von Lizenzen.

3.1 Problemaufriss und Zielsetzung

Wie schon erwähnt, werden offene Datensätze mit sogenannten Metadaten[4] beschrieben. In diesen Daten wird auch angegeben, unter welcher Lizenz der Datensatz veröffentlicht wurde. Die Lizenz wird in den Metadaten durch vier Felder beschrieben (mehr Informationen in Abschnitt 3.4). Normalerweise versucht man die Lizenz zu identifizieren, indem man sich diese vier Felder ansieht. Durch die darin enthaltenen Angaben, können mögliche Lizenzmodelle eingegrenzt werden. Je besser bzw. je vollständiger die Angaben in diesen Metadatensätzen sind, desto besser kann auf die Lizenz geschlossen werden, unter der ein Datensatz veröffentlicht wurde. Leider ist es nicht immer der Fall, dass diese Metadaten vollständig und einheitlich ausgefüllt werden. Das erschwert den Prozess der Identifizierung und stellt ein großes Problem, welches gelöst werden soll. Bei der Veröffentlichung von Daten, wird nicht wirklich auf Vollständigkeit und Einheitlichkeit, bei ihrer Beschreibung, geachtet. Dies führt bei Datenportalen zu einer schlechten Qualität und in weiterer Folge zu einer Erschwernis die veröffentlichten Daten zu verwenden.

Im Zuge des Open-Data@WU[5] Projekts, wird die Qualität von offenen Datenportalen überprüft und versucht diese zu verbessern. Die Projektmitglieder haben Zugang zu 92 Datenportalen und können die gewonnen Daten dazu nutzen, Überlegungen anzustellen und Vorschläge zu machen, um dessen Qualität zu verbessern. Um dieses Vorhaben durchzuführen, wurden Ziele gesetzt, die hier beschrieben werden:

1. Diese große Menge an Metadaten systematisch zu bearbeiten, um am Ende eindeutig identifizierte Lizenzen daraus hervorzubringen
2. Das Beschreiben dieser Lizenzen mit den richtigen Metadaten
3. Das Aufstellen von Heuristiken, die es ermöglichen, durch die Angabe weniger Parameter, wahrscheinliche Lizenzmodelle bzw. sogar die richtige Lizenz zu bestimmen

3.2 Methodisches Vorgehen zur Generierung von Heuristiken

Zur Verbesserung der Qualität in diesem Bereich, muss eine Herangehensweise definiert werden. Die Fülle an Daten muss nach und nach aufgearbeitet werden, um am Ende stichhaltige Heuristiken formulieren zu können. Ziel dieser Überlegung ist das Aufstellen von Heuristiken, die evaluiert werden und zur Anwendung kommen sollen. Das Vorgehen sieht folgendermaßen aus:

[4] http://ckan.org/features-1/metadata/
[5] http://data.wu.ac.at/

1. Datensatz einführen: Die Beschreibung der Größe des Datensatzes, dessen Herkunft und Bereitstellung.
2. Lizenzen identifizieren: Beschreibung der grundsätzlichen Identifizierung von Lizenzen, durch die angegebenen Informationen in den Metadatensätzen. Zusätzlich werden erste Ideen für die Heuristiken herausgenommen.
3. Daten begutachten und filtern: Die Beschreibung der computergestützten Begutachtung und Vorbereitung der Lizenzinformationen. Außerdem wird die Filterung der Daten beschrieben, die das Arbeiten mit der Fülle an Daten erleichtert.
4. Zusammenhänge finden: Eine Methode wird beschrieben, um das Erkennen von Zusammenhängen zwischen den Informationen zu erreichen. Dies wird benötigt, um ein Gesamtbild über die Lizenzbeschreibungen zu schaffen.
5. Heuristiken aufstellen: Beschreibung der Heuristiken, die eine eindeutige Identifikation gewährleisten sollen.
6. Heuristiken evaluieren: Die quantitative Analyse der Daten und dessen Beschreibung, welche und wie viele Lizenzen identifiziert werden konnten.
7. Kritische Begutachtung: Als letztes werden die Heuristiken kritisch diskutiert und Verbesserungsvorschläge gemacht.

Geht man so Schritt für Schritt an die Daten heran, ist man in der Lage Heuristiken aufzustellen, die dabei helfen sollen, die eingesetzten Lizenzen eindeutig zu identifizieren.

3.3 Einführung des Datensatz

Ein Großteil der Untersuchungen in diesem Bereich besteht darin, vorhandene Lizenzinformationen, in den beschriebenen Metadaten, zu begutachten und daraus Ähnlichkeiten, Zusammenhänge und grundsätzliche Fehler zu erkennen. Die benötigten Daten wurden von Mitarbeitern des Open-Data@WU Projektes bereitgestellt. Diese haben den erlaubten Zugang zu offenen Datenportalen dazu genutzt, die benötigten Bereiche der Metadaten, in denen die Lizenzinformationen beschrieben sind, von veröffentlichten Datensätzen zu extrahieren und in geeigneter Form abzuspeichern. Letztendlich wurden Daten von 92 Portalen überlassen, die alle Lizenzinformationen, von darauf veröffentlichten Datensätzen, beinhalten. Insgesamt handelt es sich hierbei um 353.767 Datensätze. In den nächsten Schritten wurden diese computergestützt aufbereitet und begutachtet. Nachdem ein erster Eindruck darüber entstanden ist, wurden diese gefiltert, um leichter damit arbeiten zu können. Als letztes wurde versucht Zusammenhänge zwischen den Lizenzbeschreibungen zu erkennen, um das Gesamtbild ein wenig

zu verfeinern. Abschnitt 3.5 beschreibt die Aufbereitung der Daten und gibt erste Anhaltspunkte, wie man Lizenzen aus vorhandenen Metadaten identifizieren kann.

3.4 Lizenzen identifizieren

Um zu erkennen, welche Lizenzierung zur Veröffentlichung der Daten verwendet wurde, können vier Metadatenfelder eingesehen werden:

1. `license_id`
2. `license_title`
3. `license_url`
4. `license`

Das sind die üblichen Felder, die in den Metadaten der veröffentlichten Datensätzen beschrieben sind. Je nachdem wie vollständig diese Felder ausgefüllt sind, kann man auf die gewählte Lizenzierung schließen. Grundsätzlich sieht man diese Felder in der selben Reihenfolge ein. Aber schon nach wenigen Versuchen, die Lizenz eines öffentlichen Datensatzes zu identifizieren, ist zu erkennen, dass diese Reihenfolge nicht effizient ist. Außerdem sind Metadatensätze unvollständig oder falsch beschrieben. Da es keine standardisierte Methode zur Beschreibung dieser Daten gibt, sind in vielen Feldern, der Lizenzbeschreibungen, völlig sinnlose Informationen eingetragen. Diese gilt es zu identifizieren und zu versuchen sie richtig zu beschreiben.

3.5 Daten aufbereiten

Die große Menge an Daten ist auf den ersten Blick sehr unübersichtlich. Deshalb werden diese Datensätze in weiteren Schritten computergestützt aufbereitet und gefiltert. Nach der Filterung bleiben nur mehr wenige Lizenzbeschreibungen zur Untersuchung übrig. Diese gilt es in Verbindung zu bringen, um das Gesamtbild über die Lizenzbeschreibungen zu verfeinern.

3.5.1 Daten begutachten und filtern

Die Daten wurden im `JSON`[6] Format überlassen, denn dieses gewährleistet eine strukturierte Gliederung und ist übersichtlich. Die einzelnen `JSON` Dateien sind in eine `MONGO`[7]

[6] `http://json.org/`
[7] `http://www.mongodb.org/`

Datenbank importiert worden, da diese eine sehr gute Verwaltung, von Dateien in diesem Format, ermöglicht. Die einzelnen Dokumente haben immer den gleichen Aufbau. Jede Datei beinhaltet vier Felder. Die ersten drei Felder (`portal_api`, `portal_title`, `portal_url`) beinhalten Metadaten zu den Portalen und beschreiben sie, in dem der Portalname, die URL zur API Schnittstelle und die URL zur Website angegeben wird. Das folgende Beispiel zeigt den Aufbau solcher Metadaten und dessen Beschreibungen, für das britische Open Data Portal in `JSON`:

```
{
 'portal_api': 'http://data.gov.uk/api',
 'portal_title': 'data.gov.uk',
 'portal_url': 'http://data.gov.uk/',
 'licenses':[

   {
    "license_id": "null",
    "license_title": "null",
    "license_url": "missing",
    "license": "null"
   },
   {
    "license_id": "uk-ogl",
    "license_title": "UK Open Government Licence (OGL)",
    "license_url": "http://www.nationalarchives.gov.uk/doc/
  open-government-licence/version/2/",
    "license": "UK Open Government Licence (OGL)"
   },
```

Metadaten des britischen Open Data Portals in JSON

Diese ersten drei Felder geben Aufschluss darüber, auf welchem Portal die Datensätze zur Verfügung stehen. Anhand der `portal_url` kommt man direkt zu der gewünschten Website. Der Eintrag im Feld `portal_api` enthält einen Link, der zur API Schnittstelle des Portals führt, über diese die Daten bezogen werden können. Das vierte Feld `licenses` ist ein weiteres Objekt, das alle Lizenzinformationen, die zur Veröffentlichung von Datensätzen über das Portal verwendet werden, beinhaltet. Wie schon weiter oben in Abschnitt 3.4 beschrieben, werden Lizenzinformationen in vier Metadatenfelder aufgeteilt. Da auf einem Portal mehrere Datensätze mit gleichen Informationen für Lizenzen veröffentlicht werden, müssen diese gefiltert werden. Durchgeführt wird dieses Vorhaben mit dem `distinct`-Befehl in der Datenbankanwendung. Dieser Befehl führt einen Vergleich der vier Felder in `licenses` durch und filtert alle mehrfach verwendeten Varianten heraus. Eine Wiederholung dieses Vorganges für jedes Portal, liefert am

Ende eine komplette Liste aller Lizenzen, in verschiedensten Varianten durch Metadaten beschrieben. Das führt zu einem Ergebnis von 283 Lizenzinformationen, die in unterschiedlichster Art beschrieben sind.

3.5.2 Zusammenhänge zwischen Datensätzen finden

Ein weiter Punkt in diesem Abschnitt ist, Zusammenhänge zwischen Lizenzbeschreibungen hervorzuheben. Dieser Schritt ist wichtig in Hinblick auf die eindeutige Identifizierung der verwendeten Lizenzen. So kann man sich zum Beispiel aus zwei oder mehreren unvollständigen Metadatensätzen, bei denen ein Zusammenhang besteht und man nicht 100-prozentig weiß um welche Lizenz es sich handelt, einen rekonstruieren, der die eindeutige Identifizierung gewährleistet. Hier ein einfaches Beispiel, um dieses Vorhaben aufzuzeigen. Gegeben ist ein unvollständiger Datensatz, der folgendermaßen aussieht:

Metadatenfeld	**Beschreibung**
`license_id`:	iodl2
`license_title`:	missing
`license_url`:	missing
`license`:	missing

Diese Metadaten lassen keine eindeutige Identifizierung zu. Nur anhand der angegebenen `license_id`, kann man nicht bestimmen, um welches Lizenzmodell es sich handelt. Das durchsehen weiterer Lizenzbeschreibungen lässt erkennen, dass diese `license_id` bei anderen Veröffentlichungen auch verwendet wurde. So findet man zum Beispiel Metadaten der Lizenz, die folgendes Aussehen haben:

Metadatenfeld	**Beschreibung**
`license_id`:	iodl2
`license_title`:	Italian Open Data License 2.0
`license_url`:	http://www.dati.gov.it/iodl/2.0/
`license`:	Italian Open Data License 2.0

Diese Art von Lizenzbeschreibung lässt erkennen, welche Lizenzierungen verwendet wurden. Deshalb ist es in dieser Situation wichtig, Zusammenhänge zwischen verschiedenen Lizenzbeschreibungen herauszufinden. Diese bewerkstelligen, dass man sofort erkennt welche Metadatensätze ähnlich sind und somit die gleiche Lizenz beschreiben. Wenn man das schon im Vorhinein weiß, hilft es dabei fehlerhafte Beschreibungen zu korrigieren bzw. unvollständige zu vervollständigen.

3.6 Heuristiken aufstellen

Eines der Ziele in dieser Arbeit ist das Aufstellen von Heuristiken. Diese Heuristiken sollen später dazu verwendet werden, eine Anwendung in der Programmiersprache `Phyton`[8] zu schreiben, die es ermöglicht, maschinell Lizenzbeschreibungen zu überprüfen und zu bestimmen, um welche Lizenz es sich handelt. In den letzten Abschnitten wurde die Aufarbeitung der bereitgestellten Daten beschrieben. Es wurde besonders auf die Vorgehensweise bei der Identifikation der Lizenzen eingegangen. Außerdem wurden dabei Erfahrungen gemacht, die den Identifikationsprozess effizienter gestalten lassen. Werden diese zwei Elemente verallgemeinert und beschrieben, entstehen dadurch Heuristiken, die in weiterer Folge in einen Algorithmus umgebaut werden können und zur Problemlösung beitragen sollen. Die in diesem Abschnitt aufgestellten Heuristiken, wurden von den Mitgliedern des OpenData@WU Projektes direkt, in dessen erstellten Anwendung `portalwatch`[9], verwendet.

3.6.1 Prozess zur Generierung der Heuristiken

Der Prozess zum Aufstellen der Heuristiken, lässt sich folgendermaßen skizzieren. Als erstes wird die Vorgehensweise herangezogen und verallgemeinert. Zu den einzelnen Schritten wird versucht, sämtliche Erfahrungen einfließen zu lassen und den Ablauf entsprechend anzupassen. Es kann sein, dass zusätzliche Schritte gebraucht werden bzw. weniger falls die Identifizierung, aufgrund von Erfahrungen, schon eindeutig ist. Außerdem ist es auch möglich, dass sich der Ablauf ändert. Dies ist der Fall, wenn zum Beispiel die `license_id` nicht ausgefüllt wurde und man deshalb nicht als nächstes die `license_url` einsieht, sondern zuerst versucht, Informationen aus dem `license_title` zu gewinnen und danach die URL untersucht. Man kann schon erkennen, dass Erfahrungen und Schlussfolgerungen besonders wichtig für die Heuristiken sind. Darum wird versucht diese ausführlich miteinzubeziehen. Hat man diese zwei Elemente berücksichtigt und einen ungefähren Ablauf entwickelt, um Lizenzen anhand ihrer Metadaten zu identifizieren, beginnt die Ausformulierung der Heuristiken.

3.6.2 Heuristik zum Ablauf der Identifizierung

Um von den gegebenen Metadaten auf die gewählte Lizenz zu kommen, wird folgender Ablauf vorgeschlagen:

[8] `https://www.python.org/`
[9] `http://data.wu.ac.at/portalwatch/`

1. Das `license_id`-Feld einsehen und dessen Inhalt analysieren. Ist eine eindeutige Identifikation möglich, müssen keine weiteren Schritte durchgeführt werden. Sollte aber keine bzw. nur eine ungefähre Bestimmung möglich sein, müssen weitere Schritte eingeleitet werden. Bei der ungefähren Bestimmung, sollten alle Möglichkeiten, die in Frage kommen, aufgezählt werden, um durch das Ausschlussverfahren und weitere Informationen, auf die richtige zu kommen.

2. Das `license_url`-Feld einsehen und die angegebene URL analysieren. Der Link kann schon Hinweise dazu liefern, welche Lizenzierung gewählt wurde. Nachdem Schritt eins durchgeführt wurde, hilft diese Information dabei, unter allen mögliche Lizenzen, Eine, eindeutig zu identifizieren. Kann die eingetragene URL keiner Lizenz zugeordnet werden, besteht die Möglichkeit, den Rechtstext einzusehen.

3. Das `license_title`-Feld einsehen und dessen Inhalt analysieren. Dieser Schritt ist ein weiterer Anhaltspunkt, um auf die gewünschte eindeutige Identifizierung zu kommen, wenn noch immer keine eindeutige Identifikation durchgeführt werden konnte. Durch den darin enthaltenen Titel der Lizenz, sollte nun endgültig die Identifizierung abgeschlossen werden können.

Das vierte Metadatenfeld `license` um Lizenzen zu beschreiben, wurde außer Acht gelassen, da es überwiegend redundante Informationen beinhaltet. Wenn das `license`-Feld nicht leer ist, wird überwiegend die LizenzID oder der Lizenztitel einfach übernommen und ein zweites mal neben dem ursprünglichen Feld, in dieses eingetragen. Das macht das Einsehen des Feldes überflüssig.

Diese Schritte können in Ihrer Reihenfolge abweichen. Es kann zum Beispiel Schritt drei vor Schritt zwei durchgeführt werden, oder sogar am Beginn des Vorganges. Es hat sich aber gezeigt, dass die Metadatenfelder `license_id` und `license_url` häufiger beschrieben werden, als der Lizenztitel. Ein weiter Punkt, der in die Reihenfolge der Schritte einfließt, ist die Effizienz. Häufig ist schon eine eindeutige Identifizierung der Lizenz möglich, nachdem die ID und die URL der Lizenzbeschreibungen begutachtet wurden.

3.6.3 Heuristik zur Identifizierung durch Lizenz ID

Zu Beginn des Identifizierungsprozesses, wird die `license_id` eingesehen. Dieses Feld beinhaltet die Abkürzung der gewählten Lizenz. Im Identifizierungsprozess wird versucht, durch die darin enthaltene Abkürzung, auf die verwendete Lizenz zu schließen. Um dieses Vorhaben durchführen zu können, ist es notwendig zu wissen, wie diese ID aufgebaut ist. Aus Erfahrung setzt sich die ID aus folgenden Punkten zusammen:

1. Abkürzung der Organisation, Initiative oder Regierung, die das Lizenzmodell bereitstellt (z.B. CC für Creative Commons)
2. Abkürzung der Art der Lizenz (z.B. BY für eine Attribution-Lizenz)
3. Abkürzung eventueller Restriktionen (z.B. NC für NonCommercial, ND für No-Derives, usw.)
4. Versionsnummer der Lizenz (z.B. CC-BY-4.0, UK-OGL-2.0, usw.)

Diese Elemente werden durch einen Bindestrich verbunden und in das Metadatenfeld `license_id` eingetragen. Lizenzen und dessen Rechtstexte, werden von Zeit zu Zeit überarbeitet und erweitert. Deshalb gibt es verschiedene Versionen. Manche Versionen sind schon veraltet und finden deshalb keinen Gebrauch mehr. Andererseits sind neuere Versionen noch nicht vollständig in allen Ländern einsetzbar. Deshalb ist die Versionsnummer bei der Angabe der Lizenzmetadaten wichtig. Es kommt vor, dass die Versionsnummer fehlt. Das hat meistens den Grund, dass es nur eine Version der Lizenz gibt. Außerdem gibt es auch Fälle, bei denen in den Metadaten, der zu Verfügung gestellten Datensätzen, die `license_id` nicht ausgefüllt bzw. falsche Lizenztexte eingetragen wurden. Dieses Szenario macht es fast unmöglich die Lizenz zu erkennen, ohne weitere Informationen zu begutachten. Betrachtet man nun diese einzelnen Abschnitte in der ID, ist es möglich, die wahrscheinlichen Lizenzen einzugrenzen und im Idealfall sogar eine davon eindeutig zu identifizieren. Dies kann durchgeführt werden, indem die LizenzID eingelesen wird und die Bindestriche als Begrenzer verwendet werden. Je nachdem wie viele Bindestriche vorkommen, ergeben sich verschiedene Varianten. Als nächstes werden die getrennten Elemente untersucht. Auf Groß- und Kleinschreibung darf nicht geachtet werden.

3.6.4 Heuristik zur Identifizierung durch Lizenz URL

Ist es nicht gelungen die Lizenz anhand der ID zu identifizieren, ist es notwendig die URL einzusehen. In diesem Metadatenfeld ist ein Hyperlink eingetragen, der zum offiziellen Rechtstext der Lizenz führt. Es kommt auch vor, dass der Link auf eine Zusammenfassung der Lizenzbedingungen verweist. Kann man schon anhand des Links erkennen, welche Lizenzierung gewählt wurde, sind keine weiteren Schritte nötig. Um die Lizenz schon am Link zum Rechtstext zu erkennen, wird versucht, den Link in offiziellen und standardisierten Metadatenbeschreibungen zu finden und die Lizenz, anhand dieser, zu identifizieren. Dieses Vorgehen ist aber nur möglich, wenn der Link direkt auf den richtigen Rechtstext verweist. Es hat sich aber gezeigt, dass Links, die ins `license_url`-Feld eingetragen wurden, auf die falschen Ressourcen verweisen. Man kommt nicht sofort auf den gewünschten Rechtstext. Solche Fälle ergeben sich, wenn

die angegebenen Links auf die Open Definition[10] oder Open Source[11] Website referenzieren, die die gewünschte Lizenzierung erwähnen. Da dies jedoch auch gewährleistet, dass man Lizenzen eindeutig identifizieren kann, müssen diese URLs auch in Betracht gezogen werden. Deswegen muss erst die richtige Ressource gesucht werden und diese wird dann begutachtet. In den meisten Rechtstexten kann man schon am Titel erkennen, wie der Name der Lizenz ist. Dies sollte dann, im Idealfall, zu einer eindeutigen Identifikation führen.

3.6.5 Heuristik zur Identifizierung durch Lizenztitel

Hat nach der Untersuchung der ersten zwei Metadatenfelder noch immer keine eindeutige Identifizierung stattgefunden, muss der Lizenztitel herangezogen werden. In dem `license_title`-Feld sollte immer die vollständige Bezeichnung der Lizenz stehen. Die Durchsicht der Metadaten, lässt dies jedoch als eher unwahrscheinlich erachten. In vielen Fällen wird die ID einfach übernommen und in das `license_title`-Feld eingetragen. Außerdem kam es auch vor, dass sogar ganze Absätze von den Rechtstexten, in das Lizenzfeld eingetragen wurden, oder eben nur Teile des Titels eingetragen wurden. Man kann zwar eventuell aus diesem Absatz Schlagwörter herausfiltern, die Aufschluss darüber geben, welche Lizenzierung von Open Data hier gewählt wurde, ist aber aufwändiger und nicht zielführend. Erfahrungen haben gezeigt, dass die Bezeichnung oft in die eigene Sprache übersetzt wird. Dies führt bei der Überprüfung oft zu Verwirrung. Außerdem ist dies, in Hinblick auf die maschinelle Durch- führung der Identifizierung, ein Problempunkt der beachtet werden sollte. Zur Identifikation einer Lizenzierung, wird der Text im Feld `license_title` eingelesen und auf Schlagworte überprüft. Je nachdem wie viele Schlagworte übereinstimmen, kann eine Identifizierung stattfinden.

3.7 Heuristiken evaluieren

Nachdem nun die Heuristiken aufgestellt und näher spezifiziert wurden, startet hier deren Evaluierung. Die nächsten zwei Abschnitte beschäftigen sich mit der Analyse der zugrunde liegenden Daten und der Qualität der formulierten Heuristiken. Vor allem wird hier auf die quantitative Analyse der Kombinationen von verschiedenen Lizenzbeschreibungen und allgemeinen Statistiken Wert gelegt. Am Ende wird noch untersucht, wie viele Lizenzen am Ende tatsächlich identifiziert werden konnten.

[10] `http://opendefinition.org/licenses/`
[11] `http://opensource.org/licenses`

3.7.1 Quantitative Analyse der Daten

Beginnen wir mit der quantitativen Analyse der Lizenzenmetadaten und Lizenzinformationskombinationen. Tabelle 1 zeigt eine Liste von allen vorkommenden Lizenzbeschreibungsvarianten und dessen Häufigkeiten. Die beschriebenen Metadatenfelder werden mit den Buchstaben I (`license_id`), T (`license_title`), U (`license_url`) und L (`license`) abgekürzt und in der Tabelle dargestellt. Ist ein Metadatenfeld in der jeweiligen Lizenzbeschreibung ausgefüllt, wird es mit einem "checkmark" versehen. Dies ermöglicht, die Tabelle nach der Häufigkeit der Lizenzbeschreibungsvariante zu durchsuchen. Die Durchsicht der Daten hat gezeigt, dass 7 Varianten von ausgefüllten Metadatenkombinationen vorkommen.

Tabelle 1: Häufigkeiten von Lizenzbeschreibungen

Lizenzbeschreibung	**beschriebene Metadaten**				**N**	**%**
	I	T	U	L		
Variante 1	-	-	-	-	75.411	21,32
Variante 2	✓	-	-	-	230	0,07
Variante 3	-	-	-	✓	1.777	0,50
Variante 4	✓	✓	-	-	35.056	9,91
Variante 5	✓	-	-	✓	42	0,01
Variante 6	✓	✓	-	✓	156.947	44,36
Variante 7	✓	✓	✓	✓	84.294	23,83
Gesamt					353.767	100,00

Insgesamt sind 353.767 Lizenzbeschreibungen vorhanden. Auf den ersten Blick sehen die Informationen sehr unvollständig aus. Bei 75.411 dieser Beschreibungen kommt es vor, dass alle vier Metadatenfelder, die zur Identifizierung einer Lizenz beitragen, nicht ausgefüllt wurden. Lizenzbeschreibungen bei denen alle vier Felder als "nicht angegeben" deklariert wurden, sind also keine Seltenheit, denn in ca. einem Fünftel aller Fälle, kommt dies vor. Eher selten ist es der Fall, dass nur ein Metadatenfeld angegeben ist und die restlichen Felder leer gelassen wurden. Dies kommt nur bei 2.007 (Variante 2 zuzüglich Variante 3) Datensätzen vor.

Betrachtet man die Beschreibungen vom Standpunkt der Vollständigkeit aus, so erkennt man, dass in ca. 44 Prozent aller Lizenzbeschreibungen, drei der vier Metadatenfelder ausgefüllt sind. Dazu kommen noch die Metadatensätze, in denen alle vier Felder ausgefüllt sind, welche ca. 24 Prozent ausmachen. Rechnet man diese beiden Werte zusammen, kann man sagen, dass in ca. 68 Prozent aller Fälle mindestens 3 der 4 vorhandenen Metadatenfelder ausgefüllt sind. Dies bedeutet, dass die erste Annahme, die Metadaten wären eher unvollständig, falsch zu sein scheint.

Auch wenn Lizenzbeschreibungen vollständig beschrieben wurden, bedeutet dies aber nicht, dass man die Lizenzen immer identifizieren kann. In der quantitativen Analyse wurden Metadaten nur als "nicht angegeben" betrachtet, wenn ihr Wert `"missing"` oder `null` ist. Es gibt jedoch auch andere Fälle, bei denen weitere Überlegungen gemacht werden müssen. Das größte Problem bei den Datensätzen ist, dass auch wenn Angaben gemacht werden, diese nichtssagend oder falsch sein können. Sieht man sich die Daten durch, erkennt man viele verschiedene Arten von fehlerhaften Lizenzbeschreibungen. Bei 15.535 Datensätzen ist es zum Beispiel der Fall, dass in das `license_id`-Feld Texte wie `"Not specified"`, `"notspecified"`, `"notspec"` usw. eingetragen wurden. Auch im `license_title`-Feld ist in diesen Fällen angegeben, dass die Lizenz "nicht spezifiziert" ist. Diese Art von Lizenzbeschreibungstexten wurden als "falsche Lizentexte" klassifiziert und sind nicht bei der Erstellung von Tabelle 1 berücksichtigt worden.

Als Letztes wurde noch eine quantitative Analyse der Lizenzen gemacht, die erkennen lässt, welche Lizenzen am häufigsten, zur Veröffentlichung von Datensätzen verwendet werden und eindeutig identifiziert werden konnten. Tabelle 2 zeigt die 10 meist verwendeten Lizenzmodelle.

Tabelle 2: Top 10 Lizenzmodelle

Platz	**Lizenzmodell**	**N**	**%**
1	Creative Commons Attribution (CC-BY)	32.707	9,25
2	Open Government Licence United Kingdom (UK-OGL)	21.647	6,12
3	Creative Commons Zero (CC0)	12.920	3,65
4	Data Licence Germany - Attribution (dl-de-by)	8.083	2,28
5	Italian Open Data License (iodl)	6.141	1,74
6	Other NonCommercial (other-nc)	5.270	1,49
7	Other Open (other-open)	3.995	1,13
8	Creative Commons Attribution-NonCommercial (CC-BY-NC)	3.236	0,91
9	Other Public Domain (other-pd)	2.265	0,64
10	ukcrown (ukcrown)	640	0,18

3.7.2 Identifizierte Lizenzen anhand der Heuristiken

In Kapitel 3.6 wurden Heuristiken hervorgebracht und näher spezifiziert, um die bereitgestellten Datensätze anzusehen und anhand der darin enthaltenen Felder, die Lizenzen eindeutig identifizieren zu können. Hier gehen wir nun der Frage nach, wie viele Lizenzen, unter Einhaltung der Heuristiken, auch tatsächlich identifiziert werden konnten.

Die 353.767 Datensätze wurden in vorigen Abschnitten erstmals analysiert und gefiltert. Aus dieser großen Menge an Datensätzen, wurde durch Vergleiche der einzelnen Metadatenfelder, mehrmals verwende Lizenzbeschreibungen herausgefiltert. Der daraus entstandene Datensatz enthält 283 Lizenzbeschreibungen. Dieser Datensatz ist die Grundlage für die Anwendung unser Heuristiken. Im Endeffekt, konnten 68 Lizenzmodelle aus den Daten der 283 Lizenzbeschreibungen, eindeutig identifiziert werden.

Diese eindeutig identifizierten Lizenzen, wurden in einer Liste dargestellt und mit Hilfe der richtigen Metadaten beschrieben. Die Liste besteht aus den maschinell lesbaren Metadaten, die die Lizenz beschreiben. Diese Tabelle wurde im `JSON` Dateiformat geschrieben. Es wurde darauf geachtet, dass die `license_url` direkt zum vollständigen Rechtstext der Lizenz führt. Bei vielen Lizenzbeschreibungen aus den Portalen führen die URL's ins Nichts. In diesen Szenarien muss man selbst den Weg zum richtigen Rechtstext finden. Deshalb ist es besonders wichtig, dass der richtige Link angegeben wird, um das eindeutige Identifizieren der Lizenzen zu erleichtern. Letztendlich entstand eine Liste, mit allen möglichen Lizenzen und dessen Metadaten, als `JSON` Datei. Sie baut auf der von der Open Knowledge Foundation bereitgestellten `JSON` Datei [12] auf, die offene Lizenzen und ihre Metadaten enthalten. Diese wird durch die identifizierten Lizenzen, aus den bereitgestellten Daten, ergänzt und dient als Grundlage, der zu entwickelnden Anwendung (Kapitel 5).

In Hinblick auf die Kompatibilität zweier Lizenzen, müssen diese aber noch genauer beschrieben werden. Wie man in Abbildung 1 sieht, werden Rechte, Pflichten und Verbote in den Lizenzbedingungen definiert, die zur Bestimmung der Kompatibilität benötigt werden. Um diese Kompatibilitätsprüfung auch maschinell durchführen lassen zu können, müssen diese Bestandteile in den Metadaten beschrieben werden. Für häufig verwendete Lizenzierungen, existiert bereits eine detaillierte Metadatenbeschreibung, bei eher Unbekannten müssen diese aber erst ausformuliert werden. In diesem Bereich gibt es bereits Forschungen, die zu späterem Zeitpunkt genauer erläutert werden (siehe Abschnitt 4.4).

3.8 Kritische Begutachtung

Der letzte Abschnitt des Kapitels, beschäftigt sich mit der Zufriedenheit der aufgestellten Heuristiken. Grundsätzlich lässt sich sagen, dass der Ablauf der eindeutigen Identifizierung absolut effizient ist. Das einzige was man daran in Frage stellen könnte, wäre, warum man nicht als Erstes die Lizenz URL einsieht. Dadurch wäre eine noch schnellere Identifikation gewährleistet. Dieser Gedanke ist völlig nachvollziehbar, jedoch schnell zu widerlegen. Grund dafür ist, dass die URL häufiger nicht ausgefüllt wird als

[12] `http://licenses.opendefinition.org/licenses/groups/all.json`

die ID. Außerdem führen die angegebenen URLs oft nicht direkt zu den Rechtstexten, weshalb es von Vorteil wäre, schon erste Ideen zu haben, welche Lizenz gemeint sein könnte. Darum wird vorgeschlagen mit der Einsicht der LizenzID zu beginnen.

Bei der Analyse des Lizenztitels, sollten noch einige Überlegungen angestrebt werden. Dieser wird häufig in verschiedenen Sprachen geschrieben. Als Beispiel wäre hier zu nennen, dass die "Creative Commons Attribution" Lizenz in Deutsch oft mit "Creative Commons Namensnennung" beschrieben wird. Auf italienischen Datenportalen wird diese Lizenz auch mit "Creative Commons Attribuzione" betitelt. Zu großen Problemen kommt es bei asiatischen Datenportalen. Da diese andere Zeichensätze zur Beschreibung von Lizenzinformationen verwenden. Die eben genannten Aspekte wären ein Grund für weitere Überlegungen, bezüglich der Analyse des Lizenztitels und in einigen Fällen auch der LizenzID.

Trotzdem herrscht große Zufriedenheit mit den aufgestellten Heuristiken. Durch die definierte Vorgangsweise konnten mehrere Lizenzmodelle, ohne viel Aufwand, identifiziert werden. Es wurden alle möglichen Aspekte miteinbezogen und wichtige Erkenntnisse erläutert. Außerdem hat sich bei der Erstellung der Anwendung in Kapitel 5 gezeigt, dass durch die Implementierung dieser Heuristiken, auch tatsächlich die Identifizierung maschinell durchgeführt werden kann.

4 Kompatibilität von Lizenzen

Der zweite zentrale Punkt in dieser Arbeit ist die Überprüfung der Kompatibilität von Lizenzen. Die theoretischen Grundlagen zum Thema wurden schon in Abschnitt 2.2.5 geklärt. Nun wird im Zuge der Qualitätsverbesserung in offenen Datenportalen versucht zu zeigen, wie die Kompatibilität von Lizenzen überprüft werden kann, um das Arbeiten mit ihnen zu erleichtern. In diesem Bereich wurde schon größtenteils Forschungsarbeit betrieben, die in dem Kapitel erläutert wird und der eigene Erkenntnisse beigefügt werden.

4.1 Problemaufriss und Zielsetzung

Größtenteils wurde schon geklärt, warum das Wissen über die Kompatibilität von Lizenzen wichtig ist. Es wird benötigt, um mehrere Datensätze in einer Anwendung zu kombinieren. Hier muss festgestellt werden, ob die Daten es erlauben, verändert, kombiniert oder erweitert zu werden. Wie man mit bestimmten Datensätzen umgehen darf, wird in den Rechtstexten beschrieben.

Das größte Problem in diesem Bereich ist, dass Lizenzen zur Veröffentlichung von Daten verwendet werden, die nicht auf Anhieb erkennen lassen, wie offen sie sind. Da jede Organisation, Regierung und Institution eigene Lizenzmodelle erstellen und bereitstellen kann, gibt es eine Vielzahl von Lizenzierungen. Hauptsächlich wurden diese schon auf Kompatibilität mit anderen Lizenzen überprüft, trotzdem kommt es immer wieder vor, dass Modelle verwendet werden, die nicht auf Anhieb erkennen lassen, wie kompatibel sie mit bekannten Lizenzen sind. In solchen Fällen, müssen die Lizenzbedingungen eingesehen werden, um herauszufinden, wie mit den jeweiligen Daten umzugehen ist. Nachdem alle Bedingungen geklärt wurden, müssen diese in weiterer Folge mit den Bedingungen anderer Lizenzierungen verglichen werden. Diese zwei Schritte wurden schon in vorhergehenden Forschungsarbeiten ([1], [10], [11]) thematisiert und dienen als Grundlage dieses Kapitels, die helfen soll die Kompatibilität zwischen verschiedenen Lizenzen zu bestimmen. Um dieses Vorhaben zu erklären, werden diese Arbeiten beschrieben und gezeigt, wie die darin beschriebenen Überlegungen auf unsere Daten angewendet werden können.

4.2 Methodisches Vorgehen

Nachdem in Abschnitt 3, dem größten Qualitätsproblem, die unvollständigen Lizenzinformationen in offenen Datenportalen, Abhilfe geschaffen wurde, wird in diesem Teil der Arbeit versucht ein Grundwissen zu schaffen, wie die Kompatibilität, der, in dem Kapitel eindeutig identifizierten Lizenzen, bestimmt werden kann. Folgende Ziele sollen dabei helfen:

1. Ein Verständnis schaffen, wie das Definieren von Lizenzbedingungen aus Rechtstexten erfolgen kann, um sie in den Metadaten zu beschreiben
2. Ein Verständnis zu schaffen, wie diese Lizenzbedingungen verglichen werden, um zu erkennen ob Lizenzen kompatibel sind
3. Eine Verbindung zwischen den theoretischen Arbeiten und den verfügbaren Daten herstellen
4. Heuristiken aufstellen, die dazu verwendet werden sollen, die identifizierten Lizenzen auf deren Kompatibilität zu überprüfen

Nun wird eine Vorgehensweise definiert, um die aufgestellten Ziele nacheinander zu erreichen und am Ende ein Wissen bereitzustellen, wie die Kompatibilität von unbekannten Lizenzmodellen erkannt werden kann. Folgende Schritte müssen durchgeführt werden:

1. Das Beschreiben von grundsätzlichen Ansätzen, die dazu verwendet werden, um die Kompatibilität von Lizenzmodellen zu überprüfen. Es soll ein Grundwissen entstehen, das später benötigt wird, wenn auf spezielle Forschungen eingegangen werden soll.

2. Das Beschreiben einer Forschungsarbeit, die dabei helfen soll, Lizenzbedingungen von unbekannten Lizenzierungen herauszufinden. Dieses Wissen wird vor allem benötigt, wenn Lizenzen verwendet werden, die noch nicht durch Metadaten beschrieben sind. Hierbei müssen die Rechtstexte begutachtet und Restriktionen in den Metadaten beschrieben werden.

3. Das Beschreiben von Forschungsarbeiten, die es sich zum Ziel gesetzt haben, diese in den Metadaten beschriebenen Lizenzbedingungen zu vergleichen und daraus dessen Kompatibilität zu erkennen.

4. Das Beschreiben von Heuristiken, die dabei helfen soll, die Kompatibilität mehrerer Datensätze zu bestimmen.

Nachdem diese Schritte durchgeführt wurden, ist es möglich die Metadaten der Lizenzbeschreibungen auf dessen Kompatibilität untereinander zu überprüfen.

4.3 Allgemeine Ansätze zur Überprüfung der Kompatibilität

Um die Kompatibilität zweier Lizenzen zu bestimmen, muss darauf geachtet werden was für Rechte, Pflichten und Verbote verschiedene Lizenzmodelle haben und ob man diese mit anderen in Einklang bringen kann. Abbildung 1 zeigt einen Gesamtüberblick über diese drei Kriterien und ihre Attribute. In der Grafik werden Lizenzen beleuchtet und kontrolliert welche Attribute darin beschrieben werden. Je nachdem enthält die Lizenz gewisse Rechte, Pflichten und Verbote, oder nicht. Außerdem gibt es Klärungsbedarf wenn die Frage aufkommt, welche Lizenzen zur Veröffentlichung verwendet werden dürfen, wenn ein einzelner Datensatz verändert, bearbeitet oder adaptiert wird und daraus neue Datensätze entstehen. Abbildung 2 zeigt eine Matrix, die zur Lösung jener Fragestellung ihren Beitrag leistet. Diese Matrix überprüft jede Lizenz auf deren Kompatibilität mit anderen Lizenzen, unter dem Aspekt der Weiterverwendung. Sind Daten unter der CC0-Lizenz öffentlich zugänglich, dann können daraus entstandene neue Datensätze, unter jeder Lizenz veröffentlicht werden, die man präferiert. Ist ein Datensatz jedoch unter der CC-BY-SA-Lizenz frei zugänglich gemacht worden, kann, der veränderte oder bearbeitete Datensatz, nur unter derselben Lizenz öffentlich zur Verfügung gestellt werden.

	Permissions			Requirements				Prohibitions
License	Reproduction [1]	Distribution [2]	Derivative Works [3]	Notice [4]	Attribution [5]	Share Alike [6]	Copyleft [7]	Lesser Copyleft [8]	Non-Commercial [9]
CC0	X	X	X						
CC-PDM	X	X	X						
CC-BY-ND	X	X		X	X				
CC-BY-NC-ND	X	X		X	X				X
CC-BY	X	X	X	X	X				
CC-BY-SA	X	X	X	X	X	X			
CC-BY-NC	X	X	X	X	X				X
CC-BY-NC-SA	X	X	X	X	X	X			X
ODC-PDDL	X	X	X						
ODC-BY	X	X	X	X	X				
ODC-ODbL	X	X	X	X	X	X			
OGL 2.0	X	X	X	X	X				
OS OpenData	X	X	X	X	X	?			

Abbildung 1: Lizenzattributsmatrix

	Permissable License for derivative												
Original License	CC0	CC-PDM	CC-BY-ND	CC-BY-NC-ND	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-PDDL	ODC-BY	ODC-ODbL	OGL 2.0	OS OpenData
CC0	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
CC-PDM	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
CC-BY-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY-NC-ND	-	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY	N	N	Y	Y	Y	Y	Y	Y	N	Y?	Y	Y	Y
CC-BY-SA	N	N	N	N	N	Y	N	N	N	N	N	N	N
CC-BY-NC	N	N	N	Y	N	N	Y	Y	N	N	N	N	N
CC-BY-NC-SA	N	N	N	N	N	N	N	Y	N	N	N	N	N
ODC-PDDL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ODC-BY	N	N	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
ODC-ODbL	N	N	N	N	N	N	N	N	N	N	Y	N	N
OGL 2.0	N	N	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y
OS OpenData	N	N	N?	N?	N?	Y	Y?	Y	N	Y?	Y	N	Y

Abbildung 2: Kompatibilitätsmatrix zur Weiterverwendung von Datensätzen

First License	CC0	CC-PDM	CC-BY-ND	CC-BY-NC-ND	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-PDDL	ODC-BY	ODC-ODbL	OGL 2.0
CC0	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0
CC-PDM	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0
CC-BY-ND	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY-NC-ND	-	-	-	-	-	-	-	-	-	-	-	-
CC-BY	CC-BY	CC-BY	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	CC-BY	CC-BY	ODC-ODbL	CC-BY
CC-BY-SA	CC-BY-SA	CC-BY-SA	-	-	CC-BY-SA	CC-BY-SA	-	-	CC-BY-SA	CC-BY-SA	ODC-ODbL	CC-BY-SA
CC-BY-NC	CC-BY-NC	CC-BY-NC	-	-	CC-BY-NC	-	CC-BY-NC	CC-BY-NC-SA	CC-BY-NC	CC-BY-NC	-	CC-BY-NC
CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	-	-	CC-BY-NC-SA	-	CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	CC-BY-NC-SA	-	CC-BY-NC-SA
ODC-PDDL	No restrictions	No restrictions	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	No restrictions	ODC-BY	ODC-ODbL	OGL 2.0
ODC-BY	ODC-BY	ODC-BY	-	-	ODC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	ODC-BY	ODC-BY	ODC-ODbL	ODC-ODbL
ODC-ODbL	ODC-ODbL	ODC-ODbL	-	-	ODC-ODbL	ODC-ODbL	-	ODC-ODbL	ODC-ODbL	ODC-ODbL	ODC-ODbL	ODC-ODbL
OGL 2.0	OGL 2.0	OGL 2.0	-	-	CC-BY	CC-BY-SA	CC-BY-NC	CC-BY-NC-SA	OGL 2.0	ODC-BY	ODC-ODbL	ODC-ODbL
OS OpenData	OS Open Data	OS Open Data	-	-	OS OpenData	OS OpenData	?	?	OS OpenData	OS OpenData	?	OS OpenData

Abbildung 3: Kompatibilitätsmatrix um Datensätze zusammenzuführen

Weiteres sollte man wissen, welche Lizenzen verwendet werden können, wenn man zwei Datensätze zusammenführt und daraus ein neuer entsteht. Hier hilft die Kompatibilitätsmatrix um Datensätze zusammenzuführen (Abbildung 3). In dieser Matrix werden zwei Lizenzen gegenüber gestellt und bestimmt, welche verwendet werden soll um die neu entstandenen Daten zu lizenzieren. Außerdem kann man auch sehen, welche Lizenzen nicht miteinander vereinbar sind. Alle drei Abbildungen wurden dem `GitHub` Beitrag "License Compatibility" [2] entnommen und wiederverwendet. Man kann erkennen, dass es zum Thema "Kompatibilität von offenen Lizenzen" schon Forschungsbeiträge gibt. Weiter Arbeiten werden in den folgenden Kapiteln beschrieben.

4.4 Lizenzbedingungen aus Rechtstexte definieren

Die Kompatibilität zwischen zwei Lizenzen lässt sich bestimmen, indem die Lizenzbedingungen dieser beiden verglichen werden. Bei bekannten und häufig genutzten Lizenzierungen, wurden deren Bedingungen schon aus den Rechtstexten herausgefunden und mit Metadaten beschrieben. Diesen Aspekt kann man sehr gut bei dem "CIPPIC Licensing Information Project for Open Licences" sehen. Dieses Projekt hat eine Website[13] erstellt auf der verschieden Lizenzen eingesehen werden können und außerdem ihre Kompatibilität untereinander überprüft werden kann. Die Metadaten der Lizenzen enthalten Rechte, Pflichten und Verbote. Folgendes Beispiel zeigt die Rechte der Creative Commons Attribution 4.0 Lizenz im `JSON` Format, beschrieben in den Metadaten[14]:

```
"rights":{"covers_circumventions": true,
          "covers_copyright": true,
          "covers_moral_rights": false,
          "covers_neighbouring_rights": true,
          "covers_patents_explicitly": false,
          "covers_sgdrs": true,
          "covers_trademarks": false,
          "prohibits_commercial_use": false,
          "prohibits_patent_actions": false,
          "prohibits_tpms": true,
          "prohibits_tpms_unless_parallel": false,
          "right_to_distribute": true,
          "right_to_modify": true,
          "right_to_use_and_reproduce": true},
```

Ausschnitt aus den Metadaten der CC-BY-4.0

[13] `http://clipol.org/`
[14] `http://clipol.org/licences/95?tab=licence_metadata`

Hier kann man sehen, welche Rechte ein Anwender des Datensatzes, der unter der CC-BY-4.0 Lizenz veröffentlicht wurde, hat. Zum Beispiel hat er das Recht die Daten zu verwenden, zu vervielfältigen und zu verändern. Es ist zu erkennen, dass man aus diesem Metadatensatz eine Lizenzattributsmatrix, wie in Abbildung 1 gezeigt, vervollständigen kann, um in weiterer Folge dessen Kompatibilität herauszufinden. Es werden jedoch immer wieder Lizenzen verwendet, die eher unbekannt sind und nicht durch Metadaten beschrieben wurden bzw. dessen Rechte, Pflichten und Verbote nicht in den Metadaten definiert sind. In diesen Fällen müssen die Rechtstexte durchgesehen und nach den Elementen gesucht werden, um diese dann in die Lizenzbeschreibungen zu integrieren.

Dieses Vorhaben ist natürlich auch maschinell durchführbar und wird in dem Forschungsbeitrag "These Are Your Rights" von Elena Cabrio und Alessio Aprosio [1] aufgezeigt. In dieser Arbeit haben es sich die Autoren zum Ziel gesetzt, die Erstellung von maschinell lesbaren Lizenzinformationen zu unterstützen.

Startpunkt dieses Vorhabens sind die in natürlicher Sprache geschriebenen Rechtstexte. Diese sollen maschinell überprüft werden und automatisch in maschinenlesbare Beschreibungen übertragen werden. Um die Metadaten einer bestimmten Lizenz zu beschreiben, wurden zwei Sprachen verwendet. Die Creative Commons Rights Expression Language[15] (`CC REL`) und die Open Digital Rights Language[16] (`ODRL`). Dargestellt werden die maschinell lesbaren Lizenzinformationen in `RDF`[17] (Resource Description Framework), genauer gesagt in der `Turtle` Syntax.

Nun wurde versucht einen Algorithmus zu erstellen, der selbstständig lernt wie solche Rechtstexte aufgebaut sind und aus diesen, Schlagworte herausfiltert, die dann automatisch durch `CC REL` oder `ODRL` beschrieben werden, um am Ende eine vollständige maschinell lesbare Lizenzbeschreibungen `RDF` hervorzubringen. Natürlich kann diese Methode auch manuell durchgeführt werden, indem die Rechtstexte durchgelesen und Lizenzbeschreibungen erstellt werden. Für mehr Informationen lesen sie [1].

Wenn man nun diese Vorgehensweise auf unsere Daten anwenden will, kann man genau so handeln. Man müsste sich Gedanken über eine geeignete Notation machen und überlegen, wie Rechte, Pflichten und Verbote im `JSON` Format beschrieben werden können. In weiteren Schritten muss ein Algorithmus erstellt werden, der die vorhandenen Lizenztexte einliest und einzelne Wörter überprüft. Werden bestimmte Schlagwörter gefunden, die in unserem vorher definierten "Wortschatz" vorkommen, dann werden diese herausgefiltert und daraus Lizenzbeschreibungen im `JSON` Format beschrieben.

[15] `http://creativecommons.org/ns`
[16] `http://www.w3.org/ns/odrl/2/`
[17] `http://www.w3.org/RDF/`

4.5 Lizenzbedingungen auf Kompatibilität prüfen

Gehen wir nun einen Schritt weiter und versuchen die Lizenzbeschreibungen auf dessen Kompatibilität zu überprüfen. Im vorherigen Kapitel wurde erklärt, dass diese Lizenzbeschreibungen in den Metadaten eines Datensatzes maschinell beschrieben werden können, in dem man aus dem Rechtstext Schlagworte herausfiltert und in einer bestimmten Notation und Syntax festhält. Diese Metadaten können in weiterer Folge zur Bestimmung der Kompatibilität mit anderen offenen Datensätzen verwendet werden. Wie man in solchen Fällen vorgehen kann, beschreiben die Autoren Serena Villata und Fabien Gandon in ihrer wissenschaftlichen Arbeit "Licenses Compatibility and Composition in the Web of Data" [10]. Diese Arbeit beschäftigt sich mit den Unklarheiten der Lizenzen unter denen verschiedene Datensätze veröffentlicht wurden. Diese Unklarheiten ergeben sich aus den nicht einheitlich beschrieben Metadaten und dem nicht vollständigen Wissen über die Kompatibilität von Lizenzen. Durch Erklärungen, wie Lizenzbeschreibungen einheitlich beschrieben werden und welche Regeln bei der Bestimmung der Kompatibilität verwendet werden können, wird in diesem Forschungsbeitrag versucht Unklarheiten zu beseitigen.

Die Autoren erwähnen in ihrer Arbeit, dass Lizenzen grundsätzlich nach folgendem Schema aufgebaut sind. Jede Lizenz besteht aus einem Set an Modellen. Genauer gesagt sind diese Modelle `Rechte, Pflichten und Verbote`. Jedes Modell enthält ein Set an Elementen, die in Verbindung zu den Modellen stehen. Man könnte auch sagen, Elemente sind die Ausprägungen der Modelle. Um die Kompatibilität zweier Lizenzen zu überprüfen, werden nun die einzelnen Elemente verglichen und bewertet. Diese Bewertung findet im Rahmen eines bestimmten Regelwerks statt. Die Autoren haben Kompatibilitätsregeln definiert und einzelne Elemente untereinander verglichen bzw. bewertet. Durch diese Regeln ist es nun möglich, die einzelnen Ausprägungen von Lizenzen zu vergleichen und am Ende ihre Kompatibilität zu bestimmen.

Ein Teilbereich dieser Regeln sind jene, die bestimmen welche Elemente andere Elemente subsumieren. Zum besseren Verständnis wird folgendes Beispiel genannt. Nehmen wir an, es gibt zwei Lizenzen, die auf deren Kompatibilität überprüft werden sollen. Die erste Lizenz enthält das Element `Reproduction` in dessen Rechten. Dieses Recht erlaubt es, den Benutzern Kopien von jenem Datensatz zu machen. In der zweiten Lizenz ist das Recht `Sharing` verankert, dass den Benutzern das Weiterverbreiten und Weiterverwenden des Datensatzes, auf kommerzielle Weise, erlaubt. Vergleicht man die zwei Elemente, erkennt man, das `Sharing` einem Benutzer mehr Rechte einräumt als `Reproduction` und somit und somit eine Subsumtionsbeziehung zwischen den zwei Elementen besteht. Dies bedeutet, dass die zwei Elemente kompatibel sind. Die Regeln der Subsumtion entnehmen sie der Tabelle 1 in [10, Seite 6].

Eine mögliche Situation wäre, dass in einer Lizenz Rechte verankert sind, die in der zweiter Lizenz nicht spezifiziert sind. Beispielhaft dafür wäre, wenn in der ersten Lizenz Verbote spezifiziert werden, die zweite Lizenz solche Elemente nicht enthält. Die Forscher haben nun alle Elemente die in Rechten, Pflichten und Verbote vorkommen können, auf Kompatibilität mit "nicht spezifizierten" Elementen untersucht. Sie sind zu folgenden Erkenntnissen gekommen:

1. Pflichten sind kompatibel mit nicht spezifizierten Elementen. Folgende Elemente wie `Attribution, Notice, SourceCode` und `CopyLeft` beeinflussen in keinster Weise die Kompatibilität zwischen Elementen zweier Lizenzen. Einzig und allein das `ShareAlike` Element beeinflusst die Kompatibilität zwischen zwei Lizenzen, indem die zweite Lizenz, die selbe sein muss wie die erste.

2. Verbote sind nicht kompatibel mit nicht spezifizierten Elementen. Der Hauptgrund dafür ist, dass, zum Beispiel, die kommerzielle Nutzung als Standardeinstellung angenommen wird und man in weiterer Folge keine Kompatibilität bestimmen kann, wenn in der anderen Lizenz die kommerzielle Nutzung verboten wird.

3. Rechte sind nicht kompatibel mit unspezifizierten Elementen. Hier wird angenommen, dass nicht spezifizierte Elemente einer Lizenz, gleichbedeutend mit einer Ablehnung der Kompatibilität ist.

Nachdem nun die einzelnen Teilbereiche betrachtet wurden, kann die Kompatibilität bestimmt werden. Zwei betrachtete Lizenzen sind kompatibel, wenn

1. Die Modelle der Lizenzen kompatibel sind

2. Die Modelle der Lizenzen gleich sind

3. Die Modelle aus Elementen bestehen, die die Kompatibilitätsregeln befolgen

Punkt drei kann noch weiter aufgeteilt werden. Die Elemente eines Modells einer Lizenz sind kompatibel, wenn

1. Die Elemente eines Modells kompatibel sind

2. Die Elemente eines Modells gleich sind

3. Die Elemente eines Modells den Subsumationsregeln folgen

4. Die Elemente eines Modells kompatibel sind mit nicht spezifizierten Elementen

Diese Regeln werden nun in eine Reihenfolge gebracht und als Algorithmus dargestellt. Die Autoren erwähnen in ihrer Arbeit [10] einen erstellten Algorithmus, der die oben besprochenen Regeln beinhaltet und dem Problem, der Überprüfung der Kompatibilität zweier Lizenzen, entgegenwirken soll. Ein weiterer Punkt der in diesem Forschungsbeitrag besprochen wird, ist die Zusammenführung von zwei kompatiblen Lizenzen, zu Einer, wenn ein Datensatz weiterverwendet werden soll. Auf diesen Aspekt wird aber nicht näher eingegangen.

Die in diesem Kapitel beschriebene Vorgehensweise, kann auch für unsere Metadaten angewendet werden. Es ist möglich, von den beschriebenen Lizenzbeschreibungen, auf die Metadaten der Lizenz zu schließen, in denen die Rechte, Pflichten und Verbote der Lizenz aufgezeigt werden. Aufgrund dieser Informationen, ist es nun möglich, so vorzugehen, wie weiter oben im Kapitel beschrieben. Die einzelnen Bestandteile der Lizenz, werden nach und nach auf dessen Kompatibilität mit den Bestandteilen einer anderen Lizenz überprüft, um am Ende bestimmen zu können, ob zwei Lizenzen im Ganzen kompatibel sind.

4.6 Heuristiken aufstellen

In diesem Abschnitt werden wir uns nun damit beschäftigen, einen Ablauf zur Bestimmung der Kompatibilität, unserer eindeutig identifizierten Lizenzen, zu skizzieren. Nachdem schon Forschungen in diesem Bereich vorhanden sind und diese schon erläutert wurden, wird versucht, die in den Arbeiten beschriebenen Vorgehensweisen auf die identifizierten Lizenzen anzupassen und anzuwenden. Grundsätzlich bleibt die Vorgehensweise gleich, jedoch müssen einige Aspekte beachtet werden. Zum Beispiel, das verwendete Vokabular und die Syntax der Lizenzbeschreibungen. Wie schon erwähnt, wurden die Lizenzbeschreibungen der `JSON` Syntax überlassen. Diese ermöglicht, einen strukturierten Aufbau der Metadaten einer Lizenz und erlaubt einen einfachen Vergleich von zwei unterschiedlichen Lizenzen.

4.6.1 Heuristik zum Ablauf der Kompatibilitätsprüfung

Am Anfang steht der Ablauf der Bestimmung der Kompatibilität zwischen zwei Lizenzen. Die identifizierten Lizenzen, die zur Veröffentlichung von Datensätzen, auf offenen Datenportalen, verwendet wurden, sind sehr unterschiedlich. Eine große Rolle spielt die Bekanntheit der gewählten Lizenzierung. Die schon mehrmals erwähnten Lizenzmodelle der Creative Commons Initiative, sind sehr bekannt und wurden deshalb schon durch standardisierte Metadaten beschrieben. Dieser Sachverhalt beeinflusst den Ablauf der Kompatibilitätsprüfung auf diese Weise, dass die Lizenzbedingungen nicht aus

den Rechtstexten neu definiert (Kapitel 4.4) und durch maschinell lesbare Metadaten festgehalten werden müssen. Andererseits werden auch Lizenzen zur Veröffentlichung von offenen Datensätzen verwendet, die noch nicht durch Metadaten beschrieben worden sind. Dies ist auch bei Creative Commons Lizenzmodellen der Fall. Es kommt vor, dass einzelne Lizenzen für spezifische Länder angepasst und individualisiert werden. Als Beispiel wäre hier die österreichische Version der Attribution-Lizenz (CC-BY-AT) zu nennen. Diese weicht zwar nicht stark von der ursprünglichen Attribution Lizenz ab, aber man müsste sich trotzdem den offiziellen Rechtstext der zugrundeliegenden Lizenzierung ansehen und die Lizenzbedingungen rekonstruieren, um die Kompatibilität mit anderen Lizenzen zu überprüfen. Noch schwieriger wird es bei Lizenzierungen, die nur für einen bestimmten Zweck hervorgebracht wurden. Diese sind meist völlig unbekannt und müssen ebenfalls durch den Rekonstruierungsprozess mit Metadaten beschrieben werden. Zur Bestimmung der Kompatibilität wird nun folgender Ablauf vorgeschlagen:

1. Die Suche nach den Metadaten, der zu vergleichenden Lizenzen, auf diversen Portalen (z.B. CLIPol). Sind die Lizenzmodelle schon ausführlich beschrieben worden, dann kann zu Schritt vier übergegangen werden. Ist das nicht der Fall, so müssen vorher noch die Lizenzbedingungen aus den Rechtstexten definiert werden.

2. Die Definition der Lizenzbedingungen aus den vorhandenen Rechtstexten. Nachdem Schlagwörter herausgefiltert wurden, werden diese in kompakten Lizenzbeschreibungen festgehalten, die kurz und prägnant die Rechte, Pflichten und Verbote der Lizenz aufzeigen.

3. Das Beschreiben der Lizenzen durch standardisierte Metadaten. Im Grunde besteht dieser Schritt darin, die Lizenzbedingungen in einen maschinell lesbaren Code zu übersetzen. Dies erleichtert es uns auch, die Kompatibilitätsprüfung manuell durchzuführen. Außerdem ist das die Grundlage für eine computergestützte Durchführung.

4. Die Prüfung der Kompatibilität der zu vergleichenden Lizenzen. Wie in Kapitel 4.5, werden hier die einzelnen Bestandteile der Lizenz nach bestimmten Regeln, überprüft. Die Elemente der Metadaten sind in den Abschnitten Rechte, Pflichten und Verbote unterteilt. Nachdem der Abschnitt mit den enthaltenen Rechten abgeglichen wurde und Kompatibilität aufweist, wird der nächste Abschnitt überprüft und so weiter. Sind alle Abschnitte der Lizenzen kompatibel, ist auch eine Kompatibilität im Ganzen gewährleistet.

4.6.2 Heuristik zur Definition der Lizenzbedingungen

Die beschriebene Vorgehensweise in Kapitel 4.4 ist nur von Nöten, wenn die gewählte Lizenz, unter der ein Datensatz frei zur Verfügung gestellt wurde, unbekannt ist und selten verwendet wird. In diesen Fällen kann man davon ausgehen, dass keine beschriebenen Metadaten vorhanden sind. Dies macht eine Überprüfung der Kompatibilität fast unmöglich, ohne weitere Informationen einzuholen. Deshalb wurde eine Forschungsarbeit aufgegriffen, die es sich zum Ziel gesetzt hat, Lizenzen und dessen Rechtstexte durchzugehen und die wichtigsten Aspekte daraus, oder anders gesagt deren Lizenzbedingungen, zu beschreiben und in weiterer Folge daraus maschinell lesbare Metadaten zu konstruieren. Genau wie in dieser Arbeit [10] beschrieben, kann auch mit unseren Lizenzen manuell vorgegangen werden. Startpunkt ist das Lesen des jeweiligen Rechtstextes. Es müssen die wichtigsten Informationen über die Rechte, Pflichten und Restriktionen beachtet werden. Dazu ist es hilfreich Schlagworte zu beachten. Diese sollten vorher definiert werden und dann im Rechtstext gesucht werden. Am einfachsten ist es, den Rechtstext in Englisch zu lesen, denn das macht es auch leichter, die herausgefilterten Lizenzbedingungen später in maschinell lesbare Codes zu übersetzen.

4.6.3 Heuristik zur Erstellung der Metadaten

Der nächste Teil beschäftigt sich mit dem Übersetzen der Lizenzbedingungen in maschinell lesbare Metadaten. In den Rechtstexten befinden sich alle wichtigen Bedingungen, die eine Lizenz mit sich bringt. Im vorigen Abschnitt wurde die Filterung der wichtigsten Rechte, Pflichten und Verbote besprochen. Diese werden in gekürzter und kompakter Form festgehalten. Nun ist es an der Zeit, diese Lizenzbedingungen mit maschinell lesbaren Metadaten zu beschreiben. Hierzu sollte man sich erst andere Metadaten ansehen und begutachten wie diese beschrieben worden sind. Will man später die Lizenzen auf dessen Kompatibilität überprüfen, ist es notwendig, dass diese vom Aufbau her gleich sind. Es sollte auch das gleiche Vokabular verwendet werden. Hierfür könnte man die CLIPol Website oder die `JSON` Datei der Open Knowledge Foundation verwenden, denn dies sind gute Quellen für beschriebene Metadaten von Lizenzen.

4.6.4 Heuristik zur Kompatibilitätsprüfung

Der Letzte Abschnitt dieses Kapitels beschäftigt sich mit der Prüfung der Kompatibilität von Lizenzen. Dabei werden die Lizenzbedingungen der Lizenzen verglichen und versucht zu erkennen, inwiefern sich diese ähnlich sind und ob sie das gleiche Grundprinzip haben. Um die Kompatibilität von Lizenzen zu bestimmen, ist ein stufenweiser

Vorgang nötig. Begonnen wird mit den einzelnen Elementen der Rechte, Pflichten und Verbote. Es wird jedes Element der einen Lizenz, mit dem jeweiligen Element der anderen Lizenz, verglichen. Durch vorher festgelegte Kompatibilitätsregeln, kann dessen Stimmigkeit bestimmt werden. Sind zum Beispiel alle Elemente der Rechte kompatibel mit den Elementen der Rechte in der anderen Lizenz, sind in weiterer Folge die Rechte der Lizenzen kompatibel. Diese Überprüfung wird auch für die Pflichten und Verbote durchgeführt. Als Nächstes wird untersucht, ob die einzelnen Teilbereiche (Rechte, Pflichten und Verbote) der Lizenzen kompatibel sind. Erst wenn alle Bedingungen erfüllt sind und Ähnlichkeiten aufweisen, sind die Lizenzen im Ganzen kompatibel.

5 Anwendung programmieren

Ein weiteres Ziel dieser Arbeit war es, einen Programmcode zu erstellen, der die einzelnen Metadatenfelder einliest und durch die in dieser Arbeit formulierten Heuristiken (siehe Kapitel 3.6) eindeutig zu identifizieren. Ist dies, aufgrund fehlerhafter Beschreibungen, nicht möglich, sollten Lizenzen vorgeschlagen werden, die eventuell in Frage kommen würden. Der Code wurde in der Programmiersprache `Python` geschrieben. Diese Scriptsprache hat sich im Umgang mit Datenbanken und `JSON` Schnittstellen als sehr effizient erwiesen. In diesem Kapitel wird nun die Erstellung der Anwendung beschrieben und der fertige Code Schritt für Schritt erklärt.

5.1 Beschreibung des erstellten Codes

Zu Beginn des Programms wird eine vorhandene Lizenzliste in `JSON` Dateiformat geladen, um dann dessen Lizenzinformationen auslesen zu können (`Zeile 1-2`). In diesem Fall besteht die Liste aus den eindeutig identifizierten Lizenzen in Kapitel 3. Außerdem werden Datentypen deklariert, die für das Arbeiten mit dem Programm notwendig sind (`Zeile 7`). Beim Start des Programms wird eine grafische Oberfläche gestartet, in der eine LizenzID, eine Lizenz URL und ein Lizenztitel eingeben werden können. Die übergebenen Daten werden in Strings abgespeichert (`Zeile 4-6`).

```
1 with open('licenses.json') as f:
2     json_data = json.load(f)
3
4 lid=e1.get()
5 title=e2.get()
6 url=e3.get()
7 a = [], b = [], c = []
8 textinput=''
```

Nun beginnt die eigentliche Identifizierung der Lizenz. In der `Zeile 10` wird die über die GUI eingegebene LizenzID aufgerufen, der String aufgesplittet und jeder Teil in einem Liste Datentyp abgespeichert. Als Begrenzer wurde hier ein Bindestrich gewählt. Der nächste Schritt wäre, dass für jeden einzelnen Teil der aufgesplitteten LizenzID überprüft wird, ob er im Metadatenfeld `license_id` der ersten Lizenz aus der geladenen Datei, vorkommt. Ist dies der Fall, dann werden die Metadaten der Lizenz abgespeichert, falls diese noch nicht abgespeichert wurden (`Zeile 11-13`). Wurden alle Teile überprüft, wird zur nächsten Lizenz der Datei gesprungen und derselbe Ablauf neu gestartet. Es wird so lange durch die `JSON` Datei iteriert, bis keine Metadaten mehr vorhanden sind.

```
9  for row in json_data['license']:
10     d=(lid.split('-'))
11     for x in d:
12         if (re.search(x, row['license_id'], re.I) and (row not in a)):
13             a.append(row)
```

Als nächstes starten wir den selben Vergleich mit der übergebenen Lizenz URL. Diese wird auf eindeutige Übereinstimmungen mit den URLs, in den Metadaten der vorhin abgespeicherten Lizenzen geprüft. Es wird durch die abgespeicherten Metadaten iteriert und das Metadatenfeld `license_url` zu überprüfung geladen. Wenn die übergebene URL mit dem `license_url`-Feld übereinstimmt, werden die Metadaten wiederum in einem neuen Listen Datentyp abgespeichert (`Zeile 17-20`). Außerdem wird zu Beginn des Programmteils überprüft, ob eine URL eingegeben wurde. Ist das nicht der Fall, werden alle abgespeicherten Lizenzmetadaten übernommen und zum nächsten Schritt übergegangen (`Zeile 14-16`).

```
14 if url=='':
15         b=a
16     else:
17         for val in a:
18             for x in val['license_url']:
19                 if (re.search(url, str(x)) and (val not in b)):
20                     b.append(val)
```

Als letztes wird der übergebene Lizenztitel überprüft. Dieser wird ebenfalls gesplittet und die einzelnen Worte in einer Liste gespeichert (`Zeile 25`). Als Begrenzer wurden hier die Leerzeichen in dem Titel gewählt. Für jedes Wort wird überprüft, ob es in den `license_title`, einer der abgespeicherten Lizenzen, vorkommt. Wenn das der Fall sein sollte, werden die entsprechenden Metadaten abgespeichert, wenn nicht, dann wird zu der nächsten abgespeicherten Lizenz übergegangen (`Zeile 26-28`).

```
21 if title=='':
22         c=b
23     else:
24         for val in b:
25             e=(title.split())
26             for x in e:
27                 if (re.search(x, str(val['license_title']), re.I) and (
   val not in c)):
28                     c.append(val)
```

Die Idee hinter dem Algorithmus ist, dass aus einem "Pool" an Metadaten jene eingegrenzt werden, die aufgrund der eingegebenen Informationen passend sind. Durch die Implementierung der Heuristiken, werden nach und nach diese Eingrenzungen vorgenommen, bis am Ende eine, mehrere oder keine Lizenz vorgeschlagen werden kann. Abschließend, werden in `Zeile 29-33` die abgespeicherten Lizenzmetadaten ausgegeben, um zu erkennen, welche Lizenzen in Frage kämen. Sind keine Metadaten nach der letzten Überprüfung abgespeichert worden, wird als Ausgabe "Keine passende Lizenz gefunden!" angezeigt.

```
29 if not c:
30         textinput='No appropriate License found!'
31     else:
32         for x in c:
33             textinput=textinput+"\n"+str("Title: "+str(x['license_title'
   ])+", ID: "+str(x['license_id']))
```

5.2 Verbesserungen durch Projektmitglieder

Nach der Fertigstellung der Anwendung, wurde diese an die Projektmitglieder des OpenData@WU-Projekts übermittelt. Der Programmcode wurde in ihr Tool implementiert und getestet. Dabei haben sie festgestellt, dass Probleme mit der LizenzID und dem Lizenztitel auftreten können. Als Beispiel wurde hier folgendes genannt: Gibt man bei dem Lizenztitel `Other (Open)` ein, sollte man im Grunde nur eine Lizenz als Ergebnis bekommen. In Wirklichkeit, gibt das Programm aber mehrere Lizenzen aus. Grund dafür ist die Überprüfung, ob die einzelnen Worte in dem Metadatenfeld `license_id` vorkommen. Dabei wird, wie vorher schon erwähnt, der Titel aufgesplittet und jedes darin enthaltene Wort in dem Feld gesucht. In diesem Fall würden zwei Wörter aus dem übergebenen Titel resultieren - Other und Open. Da diese Wörter auch Bestandteil in anderen Lizenztiteln sind, kommt es zu fehlerhaften Vorschlägen. Deshalb wurde der Code von Mitgliedern des Projektes dementsprechend verändert.

Ihr Vorschlag war es, zuerst den ganzen Ausdruck zu "matchen" . Erst wenn dies kein Ergebnis liefert, wird der Titel bzw. die ID aufgesplittet und der letzte Teil entfernt. Danach wird die Überprüfung mit dem "übriggebliebenen" Ausdruck fortgesetzt, solange, bis eine Übereinstimmung gefunden wird. Das ermöglicht es, dem anfangs erwähnten Problem aus dem Weg zu gehen.

6 Schlussfolgerung und zukünftige Arbeiten

In dieser Arbeit wurde es sich zum Ziel gesetzt, die Qualität von offenen Datenportalen zu verbessern. In Hinblick auf die Lizenzinformationen der Datensätze, die auf diesen Portalen zu freien Verfügung gestellt werden, herrscht Verbesserungspotential, bezüglich der eindeutigen Identifizierung der gewählten Lizenz. Deshalb wurden Heuristiken aufgestellt, die eine eindeutige Identifizierung gewährleisten soll. Diese entstanden aus der Analyse von 353.767 Metadatensätzen, die zu Beginn der Arbeit, für Forschungszwecke, zur Verfügung gestellt wurden. Aus der Begutachtung dieser Daten, haben sich schnell einige Erkenntnisse herauskristallisiert, die später bei der Formulierung der Heuristiken halfen. In weiterer Folge wurden diese evaluiert und zur Anwendung gebracht. Der daraus entstandene Algorithmus (Kapitel 3.6.2) zur Problemlösung, wurde im Zuge dessen, als Programmcode, implementiert. Am Ende der Analyse, konnten aus den überlassenen Lizenzbeschreibungen im Endeffekt 68 Lizenzen eindeutig identifiziert werden. Dem Anfangs erwähnten Problem, der unvollständigen oder falsch beschriebenen Lizenzinformationen, wurde durch die definierten Heuristiken entgegengewirkt.

Im Zuge der Forschungsfrage wurde auch die Kompatibilität zwischen Lizenzen angesprochen und Ideen erarbeitet, um diese zwischen unbekannten Lizenzmodellen überprüfen zu können. Hierbei wurde auch ein Ablauf skizziert und Heuristiken entwickelt, die dabei helfen sollen, die Kompatibilität zwischen zwei Lizenzen zu bestimmen. Vor allem bei der Weiterverwendung von Datensätzen, spielt das Wissen darüber eine große Rolle. Dabei ist es der Fall, dass Lizenzierungen gewählt werden, die eher unbekannt sind und deswegen noch nicht durch standardisierte Metadaten beschrieben sind. Deshalb wurden in dieser Arbeit Ideen erläutert, die diesem Problem Abhilfe schaffen können.

Aufgrund der Breite dieses Forschungsbereiches, konnten nicht alle Aspekte detailliert untersucht werden. Das Thema der Kompatibilität wäre ein guter Startpunkt für weitere Untersuchungen und Überlegungen. Zukünftige Arbeiten zu den Lizenzinformationen von offenen Datensätzen wären unter anderem:

1. Die Definition eines Rahmens, in dem eine Kompatibilitätsprüfung stattfinden könnte
2. Die Erstellung einer Methode, die es ermöglichen soll Lizenzbedingungen in maschinell lesbare Metadaten, geschrieben im `JSON` Dateiformat, umzuwandeln
3. Die Erstellung eines standardisierten Vokabulars, um die Lizenzbedingungen mit Metadaten beschreiben zu können

Die Forschung in diesem Bereich ist bereits sehr ausgereift und lockt immer mehr Interessenten an. Das größte Problemfeld bei den Lizenzen von offenen Datensätzen bestehen darin, dass es keine zentralen Anlaufstellen gibt, die den Lizenzierungsprozess überwachen und somit Einheitlichkeit gewährleisten können. Im Zuge dessen, haben die Urheber von Datensätzen keinen Anreiz ihre Daten zu veröffentlichen, wenn sie in den Metadaten nicht ordnungsgemäß festhalten können, wie mit den Daten umgegangen werden soll. Dies führt wiederum zu einem Rückgang der veröffentlichten Daten, was die erhoffte Transparenz in unserer Wirklichkeit einschränkt. Um die Forschung in diesem Bereich schmackhafter zu machen, lesen Sie den Artikel Defining Expressive Access Policies for Linked Data using the ODRL Ontology 2.0 [9]. Dieser zeigt auf, wie weit fortgeschritten die Forschung im Moment ist und welche weiteren Untersuchungen notwendig sind.

Literaturverzeichnis

[1] Cabrio, Elena ; Aprosio, Alessio P. ; Villata, Serena: These Are Your Rights. In: *The Semantic Web: Trends and Challenges*. Springer, 2014, S. 255–269

[2] Open Data Institute: *Licence Compatibility*. `https://github.com/theodi/open-data-licensing/blob/master/guides/licence-compatibility.md`. Version: 2014

[3] Open Data Institute: *Publisher's Guide to Open Data Licensing*. `http://theodi.org/guides/publishers-guide-open-data-licensing/`. Version: 2014

[4] Open Knowledge Foundation: *Open Data Handbook*. `http://opendatahandbook.org/`. Version: 2014

[5] Open Knowledge Foundation: *The Open Definition*. `http://www.opendefinition.org`. Version: 2014

[6] Open Knowledge Foundation: *What kinds of open data?* `http://www.okfn.org/opendata`. Version: 2014

[7] Open Knowledge Foundation: *Why Open Data?* `http://www.okfn.org/opendata`. Version: 2014

[8] Pohl, Adrian: Open Data im hbz-Verbund. Erschienen in Pro-Libris 3/2010. In: *Preprint einsehbar unter http://www.hbz-nrw.de/dokumentencenter/produkte/lod/aktuell/pohl_2010_open-data.pdf* (2010)

[9] Steyskal, Simon ; Polleres, Axel: Defining expressive access policies for linked data using the ODRL ontology 2.0. In: *Proceedings of the 10th International Conference on Semantic Systems* ACM, 2014, S. 20–23

[10] Villata, Serena ; Gandon, Fabien: Licenses Compatibility and Composition in the Web of Data. In: *COLD*, 2012

[11] Villata, Serena ; Gandon, Fabien: Towards licenses compatibility and composition in the web of data. In: *11th International Semantic Web Conference ISWC 2012* Citeseer, 2012, S. 121