Aufbau einer digitalen Datensammlung, zur Realisierung einer Möglichkeit zur Identifikation der Ort-Charakter Relation in Tolkiens Mittelerde Werken.

Toni Matzdorf (s14500)

Abgabe: 13.3.2020

Verfasser: Toni Matzdorf

ORCID: https://orcid.org/0000-0003-1519-7427

Potsdam, 13.3.2020

Titelbild: Mittelerde Karte: MMIII New Line Productions Inc. (tm) Tolkien Ent. Lic. to New Line Productions. Inc.

Inhaltsverzeichnis

1. Glossar

Digitale Geistes- und kulturwissenschaftliche Forschungsdaten

Unter digitalen geistes- und kulturwissenschaftlichen Forschungsdaten werden innerhalb von DARIAH-DE all jene Quellen / Materialien und Ergebnisse verstanden, die im Kontext einer geistes- und kulturwissenschaftlichen Forschungsfrage gesammelt, erzeugt, beschrieben und / oder ausgewertet werden und in maschinenlesbarer Form zum Zwecke der Archivierung, Zitierbarkeit und zur weiteren Verarbeitung aufbewahrt werden können.[1]

Forschungsdatenmanagement

Forschungsdatenmanagement bezeichnet den Prozess, der alle Methoden und Verfahren umfasst, die zur Sicherung der langfristigen Nutzbarkeit von Forschungsdaten angewendet werden: die Generierung, die Bearbeitung, die Anreicherung, die Archivierung und die Veröffentlichung. Im Ergebnis entstehen selbstbeschreibende Forschungsdaten. Zu Projektbeginn empfiehlt es sich, die Methoden und Verfahren in einem Datenmanagementplan zu beschreiben.[2]

Metadaten

„Metadaten sind strukturierte Daten, die Objekte (wie Daten, Dokumente, Personen usw.) beschreiben. Sie bewirken, dass den Beschreibungen unterschiedlicher Objekte eine einheitliche Struktur zugrunde liegt und erleichtern so das Suchen, Finden und Selektieren relevanter Objekte aus der Vielzahl möglicher Objekte.“ [3]

Text- und Data Mining

Unter Text Mining versteht man in der allgemeinen Fachliteratur die Datenanalyse natürlichsprachlichen Werken (Artikel, Bücher usw.), wobei Text als Datenform verwendet wird. Sie wird oft mit Data Mining, der numerischen Analyse von Datenwerken verbunden und als "Text- und Data Mining" oder einfach "TDM" bezeichnet. [4]

Open Access

Open Access meint, dass diese Literatur kostenfrei und öffentlich im Internet zugänglich sein sollte, sodass Interessierte die Volltexte lesen, herunterladen, kopieren, verteilen, drucken, in ihnen suchen, auf sie verweisen und sie auch sonst auf jede denkbare legale Weise benutzen können. In Fragen des Copyrights sollte die einzige Einschränkung darin bestehen, den jeweiligen Autorinnen und Autoren Kontrolle über ihre Arbeit zu belassen und deren Recht zu sichern, dass ihre Arbeit angemessen anerkannt und zitiert wird. [5]

---

[1] Forschungsdaten: Oltersdorf und Schmunk: Forschungsdaten, 2016, S.181.

[2] Forschungsdatenmanagement: https://www.forschungsdaten.org/index.php/Forschungsdatenmanagement (12.3.2020)

[3] Glossar der DINI AG KIM (Kompetenzzentrum Interoperable Metadaten): https://www.kimforum.org/Subsites/kim/DE/Materialien/Glossar/glossar_node.html (12.3.2020)

[4] Elsevier TDM Glossary: https://www.elsevier.com/__data/assets/pdf_file/0018/102906/TDM-Glossary.pdf (12.3.2020)

[5] Budapester Open Acess Erklärung: https://www.budapestopenaccessinitiative.org/translations/german-translation (12.3.2020)

Python NLK

Unter dem Begriff des „NLTK“, welche innerhalb einer Python Umgebung genutzt wird, versteht man Softwaresysteme oder -dienste, die die automatische Analyse von Text erleichtern, z.B. die Extraktion von benannten Entitäten. [6]

GitHub

GitHub ist ein Onlinedienst, der Software-Entwicklungsprojekte auf seinen Servern bereitstellt.

[6] Elsevier TDM Glossary: https://www.elsevier.com/__data/assets/pdf_file/0018/102906/TDM-Glossary.pdf (12.3.2020)

2. Einleitung

„One Data to Rule them All“

Man stelle sich vor, man könne durch die Bibliothek Bruchtals wandeln und hätte nur einen Tag Zeit all das Wissen der Elben in sich aufzunehmen. Ein Schatz, vermutlich kostbarer als der eine Ring, den Sauron in den Feuern des Schicksalberges heimlich schmiedete. Man stelle sich vor, man könnte das Wissen aus den Büchern und Schriftrollen extrahieren und hätte an diesem einen Tag die wichtigsten Erkenntnissen aller Zeitalter mit einem Mal vor sich zu liegen. Welch erstaunliche Erkenntnisse könnten dem geneigten Leser dort erwarten? Könnte man herausfinden ob es mehr Elben oder Zwerge gibt? Oder könnte man herausfinden wie lang die Strecke von Beren und Luthien war als sie die Silmarilsteine retteten im Vergleich zur Entfernung die Frodo zurücklegte, um den Ring in das Feuer zu werfen? Die vorliegende Arbeit kann diese Fragen zum derzeitigen Zeitpunkt nicht beantworten, jedoch die Frage: Wo in Mittelerde tauchte Gandalf, einer der fünf Istari die nach Norden kamen, überall auf und hat er schon jeden Ort von Mittelerde gesehen? Weiterhin könnte die hier vorliegende Arbeit die Frage beantworten an welchen Orten Frodo Beutlin war, als er auf dem Weg nach Mordor war. In Diskussionen der letzten Jahre wird der Zugang zu Texten über die intensive Lektüre einzelner Texte oft als close reading bezeichnet, der Zugang zu großen Textsammlungen über statistische Aussagen.[7]
Die vorliegende Arbeit soll einen kleinen Teil dazu beitragen die Möglichkeiten der Data Science und die Anwendung von Methoden am Beispiel Literarischer Werke aufzeigen und ein Bewusstsein dafür schaffen, auch Urheberrechtliche Werke für die Forschung bereitzustellen. Zudem soll die Arbeit aufweisen wie wertvoll es sein kann Werke und Daten nicht nur in PDF oder Epub bereit zu stellen, sondern eben jene Daten in Maschinen verarbeitende Formate auszugeben. Dadurch besteht die Möglichkeit anhand von Text- und Data Mining Fragestellungen zu beantworten und Forschung zu fördern.Sodass es möglich ist, quantitativ aufzeigen zu können, was in einem Text ungefähr drin sein kann. Es macht aufmerksam. Was man dort findet ist nicht eine Erklärung oder eine Interpretation, sondern es ist zunächst mal ein Indiz, darüber, dass da was ist. Zudem soll das Projekt einen groben Überblick über die Methoden und Inhalte der Digital Humanities geben, insbesondere dem Gegenstand der digitalen Sammlungen. Um eine gute wissenschaftliche Praxis sichenr zu können, soll in dieser Arbeit zudem versucht werden die FAIR-Prinzipien und die Verwendung des RDMO Tools auf kleine, private Datenprojekte anzuwenden .Zudem sollen hierbei Standards aus den Informationswissenschaften angewandt werden, wie zum Beispiel der Gegenstand der Open Access Veröffentlichung. Ziel des Projekts ist der Aufbau einer digitalen Datensammlung zu Realisierung einer Identifikationsmöglichkeit zur Ort-Charakter Relation in Tolkiens Mittelerde Werken. Um Ergebnisse erzielen zu können, wurden für die Bearbeitung des Forschungsgegenstandes die Werke „Das Silmarilion“[8], „Der Hobbit“[9], „Der Herr der Ringe“[10] von J.R.R Tolkien in der deutschen Übersetzung als Ausgangsbasis ausgewählt.

---

[7] A New Companion to Digital Humanities: Schreibman,Susan/Siemens,Ray/Unsworth,John(Hg.):A New Companion to Digital Humanities. Chichester 2016.

[8] T Das Silmarillion: Tolkien, C., & Krege, W. (2010). *Das Silmarillion*. Stuttgart, Deutschland: Klett-Cotta.

[9] Tolkien, J. R. R. (2010). *Der Hobbit: Oder Hin und zurück* (13., Aufl. Aufl.). Stuttgart, Deutschland: Klett-Cotta.

[10] Der Hobbit: Tolkien, J. R. R. (2010). *Der Hobbit: Oder Hin und zurück* (13., Aufl. Aufl.). Stuttgart, Deutschland: Klett-Cotta.

Doch warum gerade ebenjenes Fiktionales Epochale Werk? Für wen könnte die Analyse Tolkiens Werke interessant sein? Beobachtet man die letzten Jahren der Buchveröffentlichungen und Kinoverfilmungen, so fällt dem Betrachter auf, dass zwischen den Jahren 2012 und 2019 einige neue Veröffentlichungen zu Tolkiens Mittelerde Welt erfolgt sind. Neben der Verfilmung des Hobbits, wurden ebenso die Bücher „Beren und Luthien“ oder „Der Fall Gondolin“ verlegt und auch im Videospielbereich gab es zahlreiche neue Veröffentlichungen. Nicht zuletzt wurde die Welt rund um Mittelerde in das allgemeine Bewusstsein gerufen, als Amazone bekannt gab eine Serie zu produzieren. Neben dieser erfreulichen Nachricht gab es jedoch in jüngster Zeit auch traurige Mitteilungen, als der Tod Christopher Tolkien bekannt gegeben wurde. Viele Jahre lang verwaltete dieser die Unterlagen seines Vaters und verlegte in zeitlichen Abständen immer wieder neue, bisher unveröffentlichte, Geschichten aus dem Nachlass J.R.R. Tolkiens. Aus diesem Grund könnte die hier vorliegende Arbeit und die daraus sich ergebenen Ergebnisse für Literaturwissenschaftler, Informationswissenschaftler, Digital Humanities und der interessierten Öffentlichkeit von größeren Interesse sein. Einer noch nicht breiten Öffentlichkeit, ist das Webbasierte, englischsprachige, Lord oft he Ring Project[11] bekannt. Dieses nicht-kommerzielle, private Projekt zeigt welche erstaunlichen Erkenntnisse sich aus Daten und Textmining generieren lassen. Um der deutschsprachigen Community ebenso die Möglichkeit zu bieten Daten zu verarbeiten und genieren, wurde die Entscheidung getroffen die deutschen Übersetzungen als Datenbasis zu verwenden. Zu Beginn des Berichts wird der Strand der Forschung näher beleuchtet, daraufhin wird die Methodik vorgestellt und die Ergebnisse präsentiert. Abschließend folgt ein Fazit, in dem die Ergebnisse rekapituliert werden, und ein Ausblick auf mögliche nachfolgende Forschung angestellt wird.

[11] Lord oft he Ring Project: http://lotrproject.com/ (12.3.2020)

3. Stand der Forschung

3.1 Grundlagen digitaler Datensammlungen und Korpora

Unter Sammlungen wird im Allgemeinen eine Anhäufung von Materialien, zu definierten thematischen Schwerpunkten verstanden. Diese können durch bedeutsame Persönlichkeiten aus Kultur, Wissenschaft oder Historie bestimmt werden. Meist wird die Ausrichtung einer Sammlung aber nicht nur durch thematische, sondern auch durch materialspezifische (z.B. Handschriften, Nachlässe etc.) sowie sprachliche Aspekte (z.B. Asiatica, Orientalica) festgelegt. Außerdem sind Sammlungen oft Eigentum einer Institution und werden von dieser fachlich betreut und gepflegt. Im Rahmen der Pflege erfolgt auch die Weiterentwicklung, der Ausbau, die Strukturierung, die Langzeitarchivierung, die Bereitstellung und die formale sowie inhaltliche Erschließung. Für die wissenschaftliche Forschung sind Sammlungen wichtige Abbilder von Forschungsaktivitäten aus historischer Perspektive und verdeutlichen somit wichtige Kontexte und Zusammenhänge.[12]
Eine digitale Sammlung hat im Gegensatz zur analogen Sammlung noch weitere Möglichkeiten der Anreicherung. Dazu gehören Einbettung und Verlinkung verschiedener Medienformen (z.B. Bild, Film, Musik). Weitere Aspekte sind die besseren Interaktionsmöglichkeiten zwischen Nutzenden, z.B. durch Kommentier- und Annotationsfunktionen und eine bessere Nachnutzbarkeit durch die heutigen Möglichkeiten, die Dokumente maschinenlesbar zur Verfügung zu stellen.[13] Das alles trifft auch für digitale Editionen zu, jedoch werden hier die Dokumente nicht nur gesammelt, transkribiert, mit Metadaten beschrieben und verlinkt, sondern es erfolgen zusätzlich kritische Bearbeitungen. Hierbei werden beispielsweise historische Texte an die heutige Rechtschreibung angepasst und die Dokumente sind nicht nur maschinenlesbar, sondern durch Aufbereitung mittels Auszeichnungssprachen (meist XML) für andere nachnutzbar. Weitere Merkmale von digitalen Editionen sind die Kennzeichnung und Referenzieren von Orts- und Personennamen innerhalb der Dokumente und Vernetzung von diesen, z.B. innerhalb einer Chronologie sowie die möglichst genaue Abbildung des Textes mitsamt allen strukturellen Besonderheiten, beispielsweise Anmerkungen der Schreibenden in möglichst originalgetreuer und bearbeitbarer Form, sodass eine fortlaufende Erschließung und Nachnutzung möglich ist.[14]
Anhand dieser Ausführungen wird deutlich, dass digitale Editionen alle Merkmale von digitalen Sammlungen erfüllen, aber noch zusätzlich einige spezielle Charakteristika haben. Deshalb kann zusammenfassend gesagt werden, dass digitale Editionen ein Sonderfall von digitalen Sammlungen sind.
Ein weiterer wichtiger Aspekt in Bezug auf die Erhebung von Datensammlungen, ist das Erstellen von Korpora. Als solche werden digitale Sammlungen von Texten oder gesprochener Sprache bezeichnet. Innerhalb eines Korpus wiederum befinden sich unterschiedlichste Datensätze, welches digitale Repräsentation eines bestimmten, einzelnen Gegenstands wissenschaftlicher Forschung sein können.
Datensätze können im Prinzip alle relevanten Untersuchungsgegenstände repräsentieren: gesprochene Äußerungen Filme, historische Dokumente und vieles mehr.

---

[12] Vgl. Andreas Degkwitz, „Digitale Sammlungen – Vision eines Neubeginns“, Bibliothek Forschung und Praxis 38, Nr. 3 (19. Januar 2014): S. 413, https://doi.org/10.1515/bfp-2014-0064 (12.3.2020)

[13] Vgl. Degkwitz, Andreas. „Digitale Sammlungen – Vision eines Neubeginns“. Bibliothek Forschung und Praxis 38, Nr. 3 (19. Januar 2014). https://doi.org/10.1515/bfp-2014-0064. S. 414. (12.3.2020)

[14] Vgl. Patrick Sahle, „Digitale Editionen“, in Digital Humanities: eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein (Stuttgart: J.B. Metzler Verlag, 2017), S. 234-239. (12.3.2020)

Bei der Zusammenstellung einzelnen Datensammlungen sollten ebenfalls die unterschiedlichen Typen von Datensammlungen mitberücksichtigt werden. Projektbezogene Datensammlungen beschreiben Sammlungen, die von einer spezifischen Forschungsgruppe für einen spezifischen Forschungszweck erstellt worden sind. Gemeinschafts-Datensammlungen hingegen, werden von verschiedenen Forschenden genutzt und eignen sich für die Bearbeitung unterschiedlicher Fragestellungen aus verschiedenen Forschungsfeldern. Dabei sind diese Sammlungen vorwiegend auf eine mittelfristige Archivierung angelegt. Referenz-Datensammlungen sind für die Bearbeitung von Fragestellungen aus mehreren Disziplinen geeignet und oftmals sehr umfangreich und auf eine langfristige, institutionalisierte Archivierung der Daten, sowie einer technischen und inhaltliche Nachnutzbarkeit ausgelegt .Der Umfang der Datenerhebung richtet sich neben der Forschungsfrage und des Themengebietes sehr stark an den Typus der Datensammlung. So finden sich in der Fachliteratur Hinweise zur Handhabung verschiedenster Erhebungsarten. So wird Beispielsweise eine Vollständige Datensammlung nur bei sehr präzise umgrenzten und bereits gut erschlossenen Untersuchungsgegenständen empfohlen. Der Umfang einer Sammlung hängt hierbei stark von der spezifischen Fragestellung ab und kann daher nicht absolut festgelegt werden. Bei einer Repräsentativen Stichprobe wird eine Teilmenge von Datensätzen aus der Grundgesamtheit alle relevanten Datensätze durch zufällige und damit einer unvoreingenommenen Auswahl herausgebildet. Eine Alternative zu einer repräsentativen Sammlung stellt die Balancierte Sammlung dar. Diese beinhaltet die gezielte Konstruktion einer nach einer kleinen Anzahl von Kriterien balancierten Sammlung. Welche Kriterien als wesentlich erachtet werden, hängt von der jeweiligen Forschungsfrage ab. Das primäre Ziel ist es für alle Kombinationen wesentlicher Merkmale eine Mindestanzahl von Datensätzen (oder eine vergleichbare Textmenge) verfügbar zu machen. Unter einer Opportunistischen Auswahl im Gegensatz, versteht man einen pragmatischen Ansatz, der beispielsweise für wenig erschlossene Gegenstandsbereiche verwendet wird. Wesentlich ist in jedem Fall, die Auswahlstrategie klar zu dokumentieren. [15] Sodass Daten und Metadaten auch von anderen genutzt werden können, ist es von Bedeutung bei der Erstellung von Daten und ihrer Beschreibung durch Metadaten Standards zu verwenden. Dieses Verfahren wird häufig unter dem Begriff der Interoperabilität verstanden, also der Benutzbarkeit der Daten und Metadaten in unterschiedlichen Kontexten. Dabei geht es nicht nur um technische Interoperabilität, Nutzung von Daten auf unterschiedlichen Betriebssystemen und mit verschiedenen Werkzeugen, sondern auch um inhaltliche Interoperabilität, also die Verständlichkeit der Datenstrukturen und der Metadaten. Die Aspekte der Daten, die vereinheitlicht werden müssen, hängen stark von den Datentypen ab, die bearbeitet werden. Hierbei sollen im Folgenden einige Metadatentypen aufgezeigt werden. Unter dem Begriff der Deskriptive Metadaten werden eben jene Daten verstanden, die durch die Nennung des Urhebers, des Entstehungsdatums oder des Entstehungsorts beschrieben worden sind. Des Weiteren werden in ihnen Dokumente inhaltlich und formal beschrieben. Eine kleine Menge solch deskriptiver Metadaten war meist schon Grundlage für die Auswahl der relevanten Datensätze. Strukturelle Metadaten beschreiben, aus welchen kleineren Einheiten sich ein Datensatz zusammensetzt.

[15] Vgl. Patrick Sahle, „Digitale Editionen“, in Digital Humanities: eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein (Stuttgart: J.B. Metzler Verlag, 2017), S. 223-240

Administrative Metadaten dagegen beziehen sich unter anderem auf die Entstehungsgeschichte des digitalen Dokuments und welche Institution für die Erstellung verantwortlich war, sowie unter welcher Lizenz das Digitalisat verfügbar ist. [16]Technische Metadaten schließlich beziehen sich beispielsweise darauf, in welchem Dateiformat ein Digitalisat vorliegt oder welchen Umfang die entsprechende Datei hat. Für eine Langfristige Datensicherung sollten die Ergebnisse und alle Notwendigen Dokumente zur Nachvollziehbarkeit der Datenverarbeitung und Prozessierung im besten Fall Open Access veröffentlicht werden. Eine Langzeitarchivierung und Veröffentlichung zum Beispiel auf einem Repositorium dienen mehrerer Zwecken. Zum einen können so die Grundlagen der eigenen, auf der Datensammlung beruhenden Forschungsergebnisse offengelegt werden, so dass diese auch von anderen reproduziert und überprüft werden können. Zum anderen wird dadurch anderen Forschenden ermöglicht diese Daten für neuer Fragestellungen zu nutzen. Die Langzeitarchivierung kann jedoch nicht von einzelnen Forschenden verlangt werden, sondern nur von Institutionen adäquat gesichert werden. Öffentliche Einrichtungen wie Bibliotheken und Archive, an projektübergreifende Infrastrukturinitiativen (wie DARIAH und CLARIN) oder an langfristig finanzierte Repositorien (wie das europäische Zenodo) bieten daher eine gute Möglichkeit Forschung zugänglich und Nachhaltig zu gestallten. Hierbei werden die einzelnen Datensätze und Dokumente mit eindeutigen persistenten Identifikatoren ausgezeichnet, um so eine eindeutige Zuordnung und Nachnutzung zu gewährleisten. [17]

[16] Creative Commons (o.J.): Namensnennung 3.0 in Deutschland. URL: https://creativecommons.org/licenses/by/3.0/de/legalcode (12.3.2020)

[17] Vgl. Patrick Sahle, „Digitale Editionen“, in Digital Humanities: eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein (Stuttgart: J.B. Metzler Verlag, 2017), S. 223-240

3.2 FAIR Prinzipien

In einer Zeit in der aufgrund des zunehmenden Volumens, wachsender Komplexität und zunehmender Geschwindigkeit der Datenerstellung immer mehr auf Unterstützung Maschineller Techniken zurückgegriffen wird, um mit erhobenen Daten umzugehen. Im Jahr 2016 veröffentlichte die FORCE11 in „FAIR Guiding Principles for scientific data management and stewardship „[18]die "FAIR-Leitprinzipien für das wissenschaftliche Datenmanagement und die Verwaltung von Daten" Beabsichtigt war es, Richtlinien zur Verbesserung der Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendung von digitalen Forschungsdaten zu erstellen. Dabei betonen die Leitlinien, weiterführend als Prinzipien betitelt, die maschinelle Verarbeitbarkeit von Daten, welche innerhalb eines Forschungsprozesses erhoben und verarbeitet werden, auf diese zuzugreifen, sie miteinander zu verknüpfen und wiederzuverwenden können. Im Folgenden sollen die Inhalte der Prinzipien prägnant dargelegt werden, um anschließend ebenjene mit dem im Projekt erhobenen Daten verknüpft zu werden. Hierbei bedeutet FAIR jedoch nicht die uneingeschränkte Zugänglichkeit der Daten, eine Einschränkung der Zugänglichkeit, z.B. aufgrund von Datenschutz, widerspricht den FAIR-Prinzipien nicht. Die Prinzipien beziehen sich im Wesentlichen auf drei Arten von Daten: Daten (oder jedes digitale Objekt), Metadaten (Informationen über dieses digitale Objekt) und Infrastruktur. Das Wichtigste Prinzip der Fair-Prinzipien lautet: “machine-actionability”.

Findable

Um Daten Wieder- und verwenden zu können müssen diese auch „auffindbar“ sein. Sowohl Daten als auch Metadaten sollten für die professionelle Verarbeitung sowohl für Menschen als auch für Computer leicht zu finden sein. Maschinenlesbare Metadaten sind für das automatische Auffinden von Datensätzen und Diensten unerlässlich, daher ist dies ein wesentlicher Bestandteil.

F1. (Meta-)Daten werden mit einem global eindeutigen und dauerhaften Identifikator versehen, wie URN oder DOI.

F2. Die Daten werden mit umfangreichen Metadaten beschrieben.

F3. Metadaten enthalten klar und deutlich die Kennung der Daten, die sie beschreiben.

F4. (Meta-)Daten werden in durchsuchbaren Repositorien registriert oder indiziert.

Auf das Projekt bezogen lassen sich folgende Aussagen zur „Auffindbarkeit“ treffen. Zur Thematik des Unterelements **F1** kann angebracht werden, dass es nicht möglich ist als Privatperson ohne Institutionelle Hilfe einen die Daten mit einem dauerhaften Identifikator zu versehen. Jedoch wird angestrebt die Daten auf verschiedenen Wegen zu publizieren, um sicherstellen zu können das diese erreichbar bleiben. Die Anforderungen aus **F2** werden in diesem Datenprojekt durch das Erstellen einer RDF-Datei gewährleistet, in dieser werden Metadaten des Projekts dargelegt, zudem werden wichtige Metadaten in einem Forschungsdatenmanagementplan festgehalten.

---

[18] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016). (12.3.2020)

Wie in **F3** angebracht, so wird dies im Projekt wie folgend gehändelt, innerhalb der RDF-Datei, welche Metadaten zu den einzelnen Datensätzen enthält, wird jedes Datenset eindeutig benannt und ausgezeichnet. Die in **F4** angebrachte Forderung kann nur in einem kleinen Umfang erfüllt werden, da es derzeit nicht möglich ist die Daten auf ein institutionelles Repositorium sichern zu können. Jedoch werden die Daten auf ein GitHub-Repository ausgelagert, sowie auf der Plattform Zenodo[19] veröffentlicht und so der Öffentlichkeit zugänglich gemacht, hierbei wird es sich jedoch nur um die Urheberechtsfreien Daten handeln.

Accessible

Neben dem Auffinden der Daten sollte es dem Nutzer auch ermöglicht werden auf die die erforderlichen Daten zugegriffen zu können. Hierbei sollte der Nutzer wissen, wie auf die Daten zugegriffen werden kann, einschließlich Authentifizierung und Autorisierung.

A1. (Meta-)Daten sind durch die Verwendung eines standardisierten Kommunikationsprotokoll auffindbar.

A1.1 Das Protokoll ist offen, frei und universell einsetzbar.

A1.2 Das Protokoll erlaubt ein Authentifizierungs- und Autorisierungsverfahren, falls erforderlich.

A2. Metadaten sind auch dann zugänglich, wenn die Daten nicht mehr verfügbar sind.

Auf das Projekt bezogen lassen sich folgende Aussagen zur „Zugänglichkeit“ treffen. Zur Thematik des Unterelements **A1** kann angebracht werden das standardisierte Webprotokolle verwendet werden, um so die Daten Zugänglich zu gestallten. Des Weiteren wird neben den verwendeten Daten eine Ausführliche Datendokumentation mit Instruktionen der Zugänglichkeit auf dem Repositorium publiziert. Die Anforderungen aus **A1.1** werden in diesem Datenprojekt unter **A1** zusammengefasst. Wie in **A1.2** angebracht, so wird dies im Projekt wie folgend gehändelt, da GitHub eine freie Nutzungsfläche anbietet und nur Urheberrechtsfreie Daten veröffentlicht werden, ist es nicht Notwendig ein Authentifizierungs- oder Autorisierungsverfahren einzuleiten. Die in **A2** angebrachte Forderung kann nur in einem kleinen Umfang erfüllt werden.

[19] Zenodo: https://zenodo.org/

Interoperable

In der Regel müssen Daten mit anderen Daten zusammengeführt werden. Darüber hinaus müssen sie in der Lage sein mit Anwendungen zur Analyse, Speicherung und Verarbeitung zusammenarbeiten zu können.

I1. (Meta-)Daten verwenden eine formale, zugängliche, gemeinsame und allgemein anwendbare Sprache für die Wissensdarstellung.

I2. (Meta-)Daten verwenden Vokabulare, die den FAIR-Prinzipien folgen.

I3. (Meta-)Daten enthalten qualifizierte Verweise auf andere (Meta-)Daten.

Auf das Projekt bezogen lassen sich folgende Aussagen zur „Interoperabilität“ treffen. Zur Thematik des Unterelements **I1** kann angebracht werden, dass alle Daten in standardisierten und in der Öffentlichkeit anerkannten Formaten publiziert werden. Die Anforderungen aus **I2** werden in diesem Datenprojekt durch das RDF-Vokabular der DNB und der DBpedia abgedeckt. Das Unterelement **I3** wird in diesem Projekt mit Unterelement **I2** gleichermaßen gehandelt.

Reusable

Das übergeordnete Ziel der FAIR-Prinzipien ist die Optimierung der Wiederverwendung von Daten. Um dies zu erreichen, sollten Metadaten und Daten gut beschrieben werden, so dass sie in verschiedenen Umgebungen repliziert und/oder kombiniert werden können.

R1. Meta(daten) werden mit einer Vielzahl von genauen und relevanten Attribute ausgezeichnet.

R1.1. (Meta-)Daten werden mit einer klaren und zugänglichen Datennutzungslizenz freigegeben.

R1.2. (Meta-)Daten sind mit detaillierter Provenienz verbunden.

R1.3. (Meta-)Daten erfüllen domänenrelevante Gemeinschaftsstandards.

Auf das Projekt bezogen lassen sich folgende Aussagen zur „Nachnutzung“ treffen. Zur Thematik des Unterelements **R1** kann angebracht werden, dass innerhalb des Datenmanagementplans genau deklariert wurde unter welchen Umständen die Daten erhoben, verarbeitet und mit welchen Tools diese wiederverwendet werden können. Die Forderungen aus **R1.1** werden in diesem Datenprojekt durch die Verwendung einer Open Access Lizenz (CC BY-NC) erfüllt. Forderung **R1.2** wird in diesem Projekt nicht näher betrachtet. Jedoch die in **R1.3** angebrachte Forderung nach der Erfüllung domänenrelevanter Gemeinschaftsstandards, wird unter der zur Hilfenahme der FAIR-Prinzipien, der Verwendung eines Datenmanagementplans, einer Open Access Veröffentlichung und der ausführlichen Datendokumentation Folge geleistet. Um die Auffindbarkeit, die Zugänglichkeit, die Verarbeitung und das Wiederverwendung der im Projekt erhobenen Daten gewährleisten zu können, wurde im Laufe des Projekts ein Forschungsdatenmanagementplan unter der zur Hilfenahme des RDMO Tools erstellt. Dieser bietet die Möglichkeit ausführliche Erläuterung zum Umgang mit den erhobenen und erstellen Daten geben zu können.[20]

[20] Siehe Anhang: Forschungsdatenmanagementplan

3.3 Rechtslagen

Text- und Data Mining

Bevor Daten automatisiert analysiert werden können und Datensammlungen zu diesem Zweck angelegt werden, sollte geklärt werden, ob eine solche Nutzung zulässig ist und wie sich die derzeitige Rechtslage gestaltet. Nach §60 des Urheberechts[21] dürfen für die wissenschaftliche Forschung eine Vielzahl von Werken für die wissenschaftliche Forschung automatisiert ausgewertet werden. Das Ursprungsmaterial darf für nicht-kommerzielle Zwecke automatisiert und systematisch vervielfältigt werden, um daraus insbesondere durch Normalisierung, Strukturierung und Kategorisierung ein auszuwertendes Korpus zu erstellen. Dieser darf einem bestimmt abgegrenzten Kreis von Personen für die gemeinsame wissenschaftliche Forschung zugänglich gemacht werden. Das Korpus und die Vervielfältigungen des Ursprungsmaterials sind nach Abschluss der Forschungsarbeiten zu löschen. Zulässig ist es jedoch, das Korpus und die Vervielfältigungen des Ursprungsmaterials den in den §§ 60e und 60f genannten Institutionen zur dauerhaften Aufbewahrung zu übermitteln. § 60d UrhG gewährt jedoch kein Recht auf Zugang zu den zu analysierenden Daten, sondern setzt diesen vielmehr voraus. in der eigenen Einrichtung vorhandene analoge Bestände dürfen zum Zwecke des Text- und Data-Mining digitalisiert werden. Dies gilt sogar für per Fernleihe beschaffte Werke. [22]

Urheberrecht an den Werken J.R.R. Tolkiens

Die Rechte an den deutschen Übersetzungen der von J.R.R Tolkien verfassten Werken liegen, laut der Deutschen Tolkien Gesellschaft, bei den jeweiligen ÜbersetzerInnen und deren direkten Rechteinhabern, für den *Herrn der Ringe*, den *Hobbit* und diverse weitere Bücher bei der Hobbit Presse von Klett-Cotta. Jegliche weiteren Urheberrechtliche Belange der originalen Werke befinden sich im Besitz der Tolkien Estate. Tolkiens Familie und vor allem sein Sohn Christopher, der sich bis zu seinem Tod im Jahr 2020 um sein literarisches Erbe kümmerte. [23] Um eine konkrete Rechtliche Absicherung gewährleisten zu können, wurde in diesem speziellen Fall des Forschungsprojektes der Klett-Cotta Verlag explizit um Erlaubnis nach §60 des Urheberrechtsgesetztes gebeten, die deutschen Übersetzungen nach anhand wissenschaftlicher Standards bearbeiten zu dürfen. Der Verlag meldete sich mit folgender Antwort auf diese Anfrage zurück: „Da sich das Projekt ausschließlich auf Ihre Hausarbeit bezieht und nicht veröffentlicht oder sonst kommerziell genutzt werden wird, freuen wir uns, Ihnen zu bestätigen, dass wir mit einer Nutzung der E-Books im Rahmen des § 60d UrhG einverstanden sind.“[24]

---

[21] Gesetz über Urheberrecht und verwandte Schutzrechte §60 Text und Data Mining https://www.gesetze-im-internet.de/urhg/60d.html (12.3.2020)

[22] Forschungsdaten.info, Text-und Data Mining: https://www.forschungsdaten.info/themen/rechte-und-pflichten/text-und-data-mining/ (12.3.2020)

[23] Deutsche Tolkien Gesellschaft: https://www.tolkiengesellschaft.de/ueber-j-r-r-tolkien/rechtliche-infos-rund-um-tolkien/ (12.3.2020)

[24] Siehe Anhang: Schriftverkehr Klett-Cotta Verlag

4. Methodik

Im Folgenden werden Rahmenbedingungen zum Methodischen Vorgehen besprochen. Forschungsgegenstand des Projekts ist das Ziel, der Erstellung einer digitalen Datensammlung, zur Realisierung einer Identifikationsmöglichkeit der Ort-Charakter Relation in Tolkiens Mittelerde Werken. Um dieses Ziel erreichen zu können wird angestrebt, aus unterschiedlichen Quellen eine Projektbezogene Datensammlung zu erstellen. Dabei werden im Vorfeld Auswahlkriterien festgelegt, anhand welcher die einzelnen Daten erhoben werden sollen.[25] Dabei richtet sich die Erhebung auf eine Mischform der Opportunistischen Auswahl, sowie einer Balancierten Datensammlung. Die erhobenen Daten sollten sich in ihren Grundzügen auf das Prinzip der FAIR-Principles beziehen und anhand dieser angewendet werden. Die verwendeten Methoden und Tools in diesem Projekt sollten sich auf die Verwendung von Python, RDF sowie standardisierter Software aus dem Bereich der Digital Humanities beschränken. Aufgrund dessen, dass, dass Projekt im Bereich Data Science angesiedelt ist, sollten Methoden aus diesem Bereich Anwendung finden. Um die Literarischen Werke von J.R.R. Tolkien maschinenlesbar verarbeiten zu können, werden diese im Vorfeld als digitale Ausgabe in das Datenset aufgenommen, anhand verschiedener Tools exportiert/formatiert und anhand Datenwissenschaftlicher Methoden, beispielhaft mittels Natural Language Processing, normalisiert und Tokenisiert. Auf der Grundlage der einzelnen Auswahlkriterien, werden weiterhin Daten zum Gegenstand „Ortsnamen“ sowie „Charakterentäten“ unter der zur Hilfenahmen automatischer Prozesse erhoben und anhand einer analogen Begutachtung validiert. Des Weiteren werden, wenn notwendig, einzelne Entitäten digitalisiert, um so, eine automatisierte Bearbeitung möglich zu machen. Zur Sicherung der einzelnen Metadaten, wird angestrebt wichtige Daten zum Verständnis der Sammlung und der Ergebnisse, ein in RDF/XML Verfasstes Metadatensets zu erstellen. Da in diesem Projekt urheberrechtlich geschützte Literatur analysiert wird, sollte weiterführend auf die Rechtslage der einzelnen Dateien hingewiesen werden. Die nicht durch das Urheberrecht geschützten Daten sollten nach Projekt Abschluss auf ein öffentlich zugängliches Repository abgelegt werden und unter einer Open Access Lizenz veröffentlich werden.

[25] siehe Auswahlkriterien Katalog

5. Datendokumentation

5.1 Datenerhebung

Der Korpus der Projektbezogene Datensammlung, welche anhand vordefinierter Auswahlkriterien erstellt wurde, besteht hierbei ausfolgenden Elementen. Bei den den Datensätzen “*Das Silmarillion“*, und „Der Hobbit“ handelt es sich um das Literarische Werk von J.R.R Tolkien in der deutschen Übersetzung von Wolfang Krege (2010), erschienen im Klett-Cotta Verlag. Dieser Datensatz wurde im Vorfeld mittels Software vom Ausgangsformat „epub“ und einer „kfx-zip „Formatierung in das Datei Format ".txt" und „pdf“ formatiert. hierbei wurde darauf geachtet, dass die Zeichencodierung im Bereich UTF-8 liegt. Ebenso wurde dieses Verfahren beim Datensatz "Der Herr der Ringe" in der deutschen Übersetzung von Margaret Carroux (2010), erschienen im Klett-Cotta Verlag angewendet. Um die Ursprungsdatei in ein maschinenlesbares Format formatieren zu können, war ist notwendig diese mittels "Calibre" zu konvertieren. Daraufhin wurde für die weitere Verarbeitung die Anhänge, das Impressum und die Titelseiten gelöscht, da es für das Projekt wichtig ist den reinen Literarischen Text analysieren zu können. Bei dem Datensatz "*Orte aus Mittelerde“*, handelt es sich um eine in UTF-8 formatierte CSV Tabelle mit 916 individuellen Datensätzen zu allen historisch wichtigen Bezugsorten innerhalb der Bücher "Das Silmarillion", "Der Hobbit" und "Der Herr der Ringe“. Um die Daten in ein maschinenlesbares Format formatieren zu können, ist es notwendig diese mittels eines Webscrapping Tools (Table to Excel) aus dem Internet zu extrahieren. Zur Extraktion der Orte aus den Werken von J.R.R Tolkien wurde sich für die Website "Ardapedia" entschieden. Diese Forum basierte Website genießt in der deutschsprachigen "Mittelerde" Community ein hohes Ansehen und bietet vertrauensvolle und auf dem "Vier-Augen-Prinzip" basierte Daten an. Des Weiteren bezieht diese aktive Community ihre Daten aus mehreren vertrauensvollen Quellen, wie zum Beispiel aus den bisher veröffentlichen Büchern, der Deutschen Tolkien Gesellschaft und vielen weiteren Quellen. Anschließend an die Extraktion der Daten, wurden diese in eine "csv" formatierte Tabelle (Excel, Numbers) überführt und normalisiert. Zur Validierung und Erweiterung der Daten wurden diese händisch mit analogen Quellen abgeglichen (Lexikon, Atlas,). Ein ähnliches Verfahren wurde ebenso bei dem Datensatz der Charaktere aus Mittelerde angewandt. Bei dem Datensatz „ *Charaktere Mittelerde“*, handelt es sich um eine in UTF-8 formatierte CSV Tabelle mit 605 individuellen Datensätzen zu allen historisch wichtigen Charakteren innerhalb der Bücher "Das Silmarillion", "Der Hobbit" und "Der Herr der Ringe“*.:* Um die Daten in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels der Exportfunktion vom Browser "Safari" in PDF abzuspeichern. Im darauffolgenden Schritt wurde diese Datei unter zur Hilfenahme der Software "Calibre" in Word exportiert, und anhand von Funktionen innerhalb der Software gesäubert. Anschließend wurde diese Datei in eine "csv" Datei umgewandelt und normalisiert. Weitere Daten wurden aus den bereits im Vorfeld bearbeiteten literarischen Werken entnommen, da diese in einigen Fällen im Anhang eine Charakterliste beinhaltete. Beim Datensatz der digitalisierten Karte Mittelerdes, handelt es sich um ein Digitalisat des Close Up Poster Herr der Ringe - Karte von Mittelerde Riesenformat 135,5 x 98cm. Hergestellt im Digitalisierungslabor der Fachhochschule Potsdam, im Format "JPG" und „TIF".

5.2 Datenstruktur

Originaldaten

Literarische Werke

Die erworbene literarischen Werke Tolkiens „Das Silmarillion“, „Der Hobbit“ und „Der Herr der Ringe“ lagen am Anfang des Projekts in strukturierter Form in einem durch Digital Rights Management geschützten Format vor.

Datensatz: Orte Mittelerde

Die Original Daten zu Ortsbezogenen Einheiten aus den zu betrachtenden Werken, lagen am Beginn des Projekts in statischer, strukturierter Form vor. Die Hauptsächlich benutzen Daten befanden sich auf der im Netzt frei Verfügbaren Internetseite Ardapedia[26] und wurden in einer strukturierten Tabellenform angeboten. Weitere Ortsnamen konnten aus den Büchern „Das große Mittelerde-Lexikon”[27] und “Historischer Atlas von Mittelerde”[28]entnommen werden. An dieser Stelle sei anzumerken, dass sich die Übersetzerin des Historischen Atlas von Mittelerde stark an die Übersetzungen von Wolfgang Krege und Helmut Pesch halt.

Datensatz: Charaktere Mittelerde

Die Original Daten zu Charakterbezogenen Einheiten aus den zu betrachtenden Werken, lagen am Beginn des Projekts in statischer Form vor. Die Hauptsächlich benutzen Daten befanden sich auf der im Netzt frei Verfügbaren Internetseite Ardapedia[29] und wurden in einer strukturierten Auflistung angeboten. Um diese jedoch verarbeiten zu könne, mussten diese aus dem Text herausgelöst und in anderen Datenformate exportiert werden. Dazu wurde im ersten Schritt ein PDF erzeugt, welche wiederum im Anschluss in eine Worddatei überführt wurde. Nach der Bearbeitung dieser Daten, konnte eine CSV Tabelle geniert werden.

---

[26] Ardapedia (2015, Februar 18): Der Herr der Ringe (Namensübersetzungen)/Orte. Abgerufen 12. März 2020, von http://ardapedia.herr-der-ringe-film.de/index.php/Der_Herr_der_Ringe_(Namens%C3%BCbersetzungen)/Orte (12.3.2020)

[27] Foster, R. (2002). *Das große Mittelerde-Lexikon. Ein alphabetischer Führer zur Fantasy-Welt von J.R.R. Tolkien.* Stuttgart, Deutschland: Lübbe.

[28] Fonstad, K. W. (2014). *Historischer Atlas von Mittelerde* (16., völlig überarbeitete verb. Neuaufl. (REV). Aufl.). Stuttgart, Deutschland: Klett-Cotta.

[29] Ardapedia (2015, Februar 18): Der Herr der Ringe (Namensübersetzungen)/Orte. Abgerufen 12. März 2020, von http://ardapedia.herr-der-ringe-film.de/index.php/Der_Herr_der_Ringe_(Namens%C3%BCbersetzungen)/Orte (12.3.2020)

Übersetzungen

Die besondere Herausforderung bei der Erhebung und Strukturierung der im Projekt erhobenen Daten besteht in der Zusammenführung unterschiedlichster Namensräume. Die inkonsistenten Namenräume entstehen hierbei aus dem Umstand heraus, dass die literarischen Hauptwerke J.R. R Tolkiens im Laufe der Jahre von unterschiedlichen Übersetzern übersetzt und an die neue deutsche Rechtschreibung angepasst wurden. Die im Projekt verwendeten deutschen Übersetzungen wurden nach ihrer Verfügbarkeit und Qualität ausgewählt. In der allgemeinen Wahrnehmung werden die Arbeiten von Wolfgang Krege, Helmut Pesch und Margaret Carroux besonders hervorgehoben, da diese sich sehr ähneln und nur minimale Abweichungen zueinander aufweisen. Des Weiteren halten sich die Übersetzungen sehr stark an den literarischen Originaltext. Problematisch ist jedoch die inkonsequente Durchführung der Groß-und Kleinschreibung bei Ortsnamen und Charakternamen. Dieses inkonsequente Verhalten, lässt sich aus dem Versuch heraus erklären die Lebendigkeit der Sprache Tolkiens zu bewahren. Ein weiterer wichtiger Punkt in Bezug auf die Sprachlich-Semantische Ebene der einzelnen Übersetzungen liegt auf der Verwendung Elbischer Namen. Wie aus dem Werken Tolkiens entnommen werden kann, so ist die Sprache der Elben weit mehr als ein fiktionaler Schnörkel. Beispielsweise treten in der Geschichte des Herrn der Ringe Elben als handelnde Personen wenig hervor, jedoch ihre Sprache rückt anhand von Ortsnamen, Flüssen, Landschaften, Länder und Seen stark in den Vordergrund. Ebenso werden einige Figuren im Laufe der einzelnen Geschichten bei ihren Namenssynonymen (Zusatznamen, Elbennamen) genannt.

Datensäuberung

Nachdem die, in den zu betrachtenden Werken, Orts und Charakternamen identifiziert wurden und in ein maschinenlesbares Format übertragen wurden, mussten diese Daten gesäubert werden, sodass sie im Anschluss anhand von Data Science Methodiken prozessiert werden konnten.

Literarische Werke

Im ersten Schritt wurde der Kopierschutz der literarischen Werke Tolkiens „Das Silmarillion“, „Der Hobbit“ und „Der Herr der Ringe“ aufgehoben, um sie so in ein verarbeitendes Format überführen zu können. Nachdem sie in das maschinenlesbare Format „txt“ überführt wurden, mussten diese händisch bearbeitet werden. Dabei mussten Anhänge, Titelblätter entfernt werden, umso einen reinen Textkorpus erhalten zu können.

Datensatz: Orte Mittelerde

Nachdem die Ortsnamen aus den gänzlich unterschiedlichen Quellen in ein strukturiertes Datenformat überführt wurden und in das maschinenlesbare Format „csv“ überführt wurden, konnten diese gesäubert werden. Im ersten Schritt wurden dafür die strukturierten Daten in einzelnen Tabellenspalten aufgesplittet. Ergänzt und validiert wurden diese Daten durch die Sichtung der analogen Quellen“ Das große Mittelerde Lexikon“ und „Historischer Atlas Mittelerde“. So konnten Zusatznamen und Elbennamen extrahiert und strukturiert wiedergegeben werden, anschließend erfolgte eine Formatierung in UTF-8, Besonders schwierig gestaltete sich die Formatierung von Zusatzzeichen wie â, á, ô.ó, so wie Umlaute wie Â; ß; Ä.Ü. Aufgrund von unterschiedlichen Betriebssystemen, die während er Bearbeitung der Daten benutzt wurden, konnten ca. 1% der Daten nicht formatiert werden und müssen dabei in der vorläufigen Auswertung der Daten berücksichtigt werden.

Datensatz: Charaktere Mittelerde

Nachdem die Charakternamen in ein bearbeitbares Format (MS Word) überführt wurden, konnten unrelevanten Daten wie Geburtsjahr-Sterbejahr, Zeitalter, Verwandtschaftsverhältnisse und viele weitere extrahiert werden, sodass am Ende eine durch Kommata getrennte Auflistung von Namen /Zusatznamen entstehen konnte. Anschließen wurde diese Auflistung in maschinenlesbares Format „csv“ überführt und weiterbearbeitet werden. Im ersten Schritt wurden dafür die strukturierten Daten in einzelnen Tabellenspalten aufgesplittet. Ergänzt und validiert wurden diese Daten durch die Sichtung der analogen Quellen“ Das große Mittelerde Lexikon“ und „Historischer Atlas Mittelerde“. So konnten Zusatznamen und Elbennamen extrahiert und strukturiert wiedergegeben werden, anschließend erfolgte eine Formatierung in UTF-8, Besonders schwierig gestaltete sich die Formatierung von Zusatzzeichen wie â, á, ô, ó, so wie Umlaute wie Â; ß; Ä.Ü. Aufgrund von unterschiedlichen Betriebssystemen, die während er Bearbeitung der Daten benutzt wurden, konnten ca. 1% der Daten nicht formatiert werden und müssen dabei in der vorläufigen Auswertung der Daten berücksichtigt werden.

5.3 Dateiengröße

Die Gesamtgröße aller Dateien beläuft sich auf ca. 4 GB.

Das Silmarillion	758 KB
Der Hobbit	594 KB
Der Herr der Ringe	2,8 GB
Orte aus Mittelerde	29 KB
Projekt Mittelerde (RDF)	8 KB
Projekt Mittelerde.py	Ca.1,1 GB
Ort-Charakter Visualisierung	ca. 1 GB

5.4 Versionierung

Bei den zu digitalisierenden Textdateien werden verschiedene Versionen des Datensatzes erzeugt. Bestimmte Änderungen am Datensatz machen eine Versionierung nötig, dazu zählen Veränderungen der textuellen Struktur und die Weiterleitung an Mitarbeitende.

Das Silmarillion	Artikel_Name_Zustand (komplett, oder <<leer>>)
Der Hobbit	Artikel_Name_Zustand (komplett, oder <<leer>>)
Der Herr der Ringe	Artikel_Name_Zustand (komplett, oder <<leer>>)
Orte aus Mittelerde	Orte_Mittelerde_Zustand (Formatiert, EndResult, unvollständig, <<Leer>>)
Projekt Mittelerde (RDF)	Charaktere_Mittelerde_Zustand (Formatiert, EndResult, unvollständig, <<Leer>>)
Projekt Mittelerde.py	Name_der_Datei_Versionsnummer
Ort-Charakter Visualisierung	Einfaches Kopieren

5.5 Datennutzung

Alle Datensätze werden zur Aufbereitung des Forschungsgegenstandes bis zum jeweiligen Abschluss an jedem Arbeitstag genutzt. Die Dateien durchlaufen die Prozesse Erhebung/Erstellung, Bereinigung, Aufbereitung und Analyse. Für alle erwähnten Datensätze reichen die üblichen Infrastrukturressourcen (CPU-Stunden, Bandbreite, Speicherplatz, etc.) aus.

5.6 Formate

Das Projekt enthält verschiedene Arten von Dateitypen, sowie werden innerhalb des Bearbeitungszeitraumes Dateien in Formaten wie: PDF, EPUB, CSV, TXT, Python, TIF, JPG und docx verarbeitet. Textdateien liegen teilweise noch in analoger, teilweise schon in digitaler Form vor. Um eine gute Interoperabilität zu gewährleisten, muss bei allen Datensätzen im Vorfeld eine Datenmigration durchgeführt werden. In nachfolgender Tabelle sind die standardisierten Formate, in denen die Datensätze vorliegen und die Datenformate, in denen sie nach der Migration ausgegeben werden, sowie die Software, durch die sie verwendet werden können, aufgelistet. Digitale Quellen vs. Analoge Quellen

Datensatz	**Liegt vor in (Format)**	**Wird ausgegeben in (Format)**	**Software**
Textdateien born digital	docx PDF EPUB (kfx-zip) pages	txt, PDF Excell docx	Texteditor, Adobe Reader, Calibre, Microsoft Excel Microsoft Word Pages
Bilddateien	Jpg Tif	/	Photoshop Picture Browser

5.7 Speicherung

Alle Dateien werden während des Projekts an mehreren Orten gespeichert, um sicherzugehen, dass keine Daten verlorengehen. Es empfiehlt sich, Daten stets an mindestens zwei verschiedenen Orten auf zwei verschiedenartigen Trägern zu speichern. Daher fungiert die eigene Infrastruktur des eigenen Rechners und die Infrastruktur der Arbeitsstelle als primärer Speicherort, als sekundärer Speicherort ist fungierte eine externe Festplatte. Da die Gesamtgröße aller Dateien nicht allzu hoch ist, ist der Speicherlevel einer externen 1 TB Festplatte ausreichend. Dabei werden zwei Festplatten zu einem Laufwerk gruppiert. Daten werden sofort auf beide Festplatten übertragen. Fällt eine Festplatte aus, hat diese keine Auswirkungen auf die zweite Festplatte und die Daten können von dort „gerettet“ werden.

5.8 Langzeitarchivierung

Die Daten zur Langezeitarchivierung werden ausgewählt nach Kriterien, die in den nestor-Materialien „Vertrauenswürdige und abgesicherte Langzeitarchivierung multimedialer Inhalte“[30] behandelt werden. Verantwortlicher hierfür ist Michael Zaiss. Alle Datensätze müssen längerfristig aufbewahrt werden, damit sie in eventuellen Folgeprojekten nachgenutzt werden können. Außerdem sollen sie aufgrund gesellschaftlicher Relevanz dokumentiert werden. Wie lang die Daten aufbewahrt werden sollen, muss noch besprochen werden. Sie werden nach Projektende im GitHub- Repository und auf Zenodo veröffentlicht, gespeichert und gesichert. Für die Daten wird es keine Sperrfrist geben, sie sollen so schnell wie möglich zugänglich gemacht werden. Es wird eine Nutzerstatistik geben, um die Anzahl der Zugriffe zu ermitteln.

5.9 Metadaten

Um eine abgesicherte Langzeitarchivierung zu gewährleisten zu können, müssen Metadaten zu jeder Ebene (physische, logische, konzeptuelle) generiert werden. Nur so können Objekte langfristig identifiziert, lokalisiert, verarbeitet und dargestellt werden. Metadaten sind von hoher Wichtigkeit, da sie eine digitale Langzeitarchivierung überhaupt erst ermöglichen.[31] Zur Beschreibung der Daten und Kontextinformationen wurde ein XML/RDF Datenset angelegt, beschrieben wurden die einzelnen Datenpakete anhand der DNB sowie der DBpedia Ontologie.

5.10 Programmierumgebung und Code

Im folgenden Abschnitt soll nun der, zur Datenauswertung, erstellte Code näher betrachtet und erläutert werden. Zu Beginn sollen hier die Settings des „PyCharm“ Programm genauer erläutert werden, die für die Verarbeitung des Pythonskriptes notwendig sind. Neben der Verfügbarkeit des NLTK-Packages, der Installation der Panda Umgebung sowie der Installation einer virtuellen Umgebung, sollte ebenso auf dem Betriebssystem mindestens Python 3.7 installiert sein. Folgende Einstellungen sollten nach der Verfügbarkeit der einzelnen Komponenten vorgenommen werden. Die neue Umgebung sollte unter der Maßgabe einer „Virtualenv“ angegeben werden. Ebenso sollte der zugrunde liegende Interpreter als „Python 3.7“ eingestellt sein. Weitere Interpreter sind nicht notwendig. Ebenso ist es Notwendig nach dem Anlegen eines neuen Python Files im Projekt als ersten Schritt den „import nltk“ und „import pandas“ durchzuführen. Um ungefähr eine valides 80 Prozentigen Ergebnis erzählen zu können, wurde sich für die Durchführung zwei Varianten zur Ermittlung der Ort-Charakterrelation entschieden. Zum Erreichen einer möglichst breiten Datenbasis wurde der Code so gestaltet, sodass es dem Nutzer Möglich ist selbständig Orte und Namen eingeben zu können und trotzdem eine automatisierte Ausgabe zu erhalten. Ebenso wurde angestrebt dem Endnutzer viele Informationen über die Daten zu liefern, so ist es möglich neben der Anzahl an Worthäufigkeiten ebenfalls Sätze, in denen die Ort-Charakterrelation bemerkbar ist, sich anzeigen zu lassen. Im Folgenden sollen einige Auszüge des Codes präsentiert werden.

---

[30] Vertrauenswürdige und abgesicherte Langzeitarchivierung multimedialer Inhalte: https://d-nb.info/1000084205/34, S. 25. (12.3.2020)

[31] Vgl. Vertrauenswürdige und abgesicherte Langzeitarchivierung multimedialer Inhalte: https://d-nb.info/1000084205/34, S. 25. (12.3.2020)

```python
# Funktion um den gesamten Datensatz in der Python Konsole anziegen zu können
pd.set_option('display.max_rows', 100000)
pd.set_option('display.max_columns', 10)
pd.set_option('display.max_colwidth', 10000)
pd.set_option('display.width', None)
```

Abbildung 1: Funktionen, um den gesamten Datensatz in der Konsole anziehen zu können

```python
##Pfad der einzenen Dateien auf dem Betriebssystem
filepath_Ringe = "/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Der Herr der Ringe.txt"
filepath_Hobbit = "/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Der Hobbit.txt"<<<<<<<<<<<<<<<
filepath_Silmarillion = "/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Silmarillion.txt"
```

Abbildung 2: Pfad der einzelnen Daten auf dem Betriebssystem

```python
###################
# Alle 3 Bücher werden in einer Liste gestellt.
books = []
books = Ringe_toknz + Hobbit_toknz_sent + Silmarillion_toknz_sent
print(len(books)) # 37272 Lists
print(type(books)) # <class 'list'>
print(books)

########################
## Laden, Lesen der Charakter und Orte Dateien aus dem Speicher des Betriebssystems ###
All_Charaktere_Mitttelerde = '/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Charaktere_Mitttelerde_utf8.csv'
All_Orte_Mittelerde = '/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Orte_Mittelerde_utf8.csv'

# Diese Funktion liest die Datei der Charaktere
Charaktere = pd.read_csv(All_Charaktere_Mitttelerde, encoding="utf-8", sep=';')
#Ausgabe type(Charaktere))
# Funktion um die Namen der Charaktere in die Kleinschreibung zu formatieren.
Charaktere_low = Charaktere.apply(lambda x: x.astype(str).str.lower())
Charaktere_low_n = Charaktere_low.replace('\n', '')
Charaktere_low_n_t = Charaktere_low_n.replace('\t', '')
#Ausgabe (Charaktere_low_n_t.head(5))

Orte = pd.read_csv(All_Orte_Mittelerde, encoding="utf-8", sep=';')
Orte_low = Orte.apply(lambda x: x.astype(str).str.lower())
Orte_low_n = Orte_low.replace('\n', '')
Orte_low_n_t = Orte_low_n.replace('\t', '')
#Ausgabe (Orte_low_n_t.head(5))
```

Abbildung 3: Laden und Verarbeitung der literarischen Werke

```python
# Um das Ergebnis speichern zu können, wird durch diese Funktion eine "csv" Datei erzeugt.
file_name = str(person) + ".csv"
path = "/Users/tonimatzdorf/Desktop/neu"
file_path = path + file_name
df_result.to_csv(file_path, index = False)
```

Abbildung 4: Ergebnissicherung

```
##########################
####### 2. Lösung ######
##########################

# Gebe einen Charakter ein
c = input('Please enter a character A-Z:\n')
# Gebe einen Ort ein
o = input('Please enter an Ort A-Z:\n')

# i to iterate as index
i = 0
```

Abbildung 5: Angebot einer zweiten Lösungsstrategie

```
/Users/tonimatzdorf/PycharmProjects/ToniMittelerde/venv/bin/python "/Users/tonimatzdorf/Desktop/Python Projekt Mittelerde/Projekt Mittelerde_v1.py"
Please enter a character a-z:
frodo
Please enter an Ort a-z:
auenland
```

Abbildung 6: Eingabemaske

```
sie stellten keine wache auf; selbst frodo befürchtete keine gefahr, denn noch immer waren sie im herzen vom auenland.

sie sprechen den namen elbereth aus!«, sagte frodo erstaunt »wenige von diesem edelsten volk sind je im auenland zu sehen.

sie sprachen von vielen dingen, alten und neuen, und frodo stellte gildor viele fragen über die ereignisse in der weiten welt außerhalb des auenlands.
```

Abbildung 7: Satzausgabe

```
############ Der Hobbit ###############
frodo is present in Der Hobbit 0 mal
frodo im Bezug mit auenland 0 mal
############ Silmarillion ###############
frodo is present in Silmarillion 1 mal
frodo im Bezug mit auenland 0 mal
frodo war in auenland 0mal
```

Abbildung 8: Worthäufigkeiten

5.11 Datenauswertung

Im nächsten Schritt sollen nun die einzelnen Daten ausgewertet und näher betrachtet werden. Im ersten Analyse Schritt, soll die Worthäufigkeit des Charakters „Gandalf „vorangestellt werden. In den zu untersuchenden Büchern „Das Silmarillion“, „Der Hobbit“ und „Der Herr der Ringe“ taucht der Charakter Gandalf insgesamt 1236 Mal auf. Aufgeschlüsselt auf die einzelnen Bücher ergibt sich folgendes Bild. Im literarischen Werk „Das Silmarillion“ taucht Gandalf 22 Mal auf, in „Der Hobbit“ 166 Mal und in „Herr der Ringe“ insgesamt 1048 Mal auf.

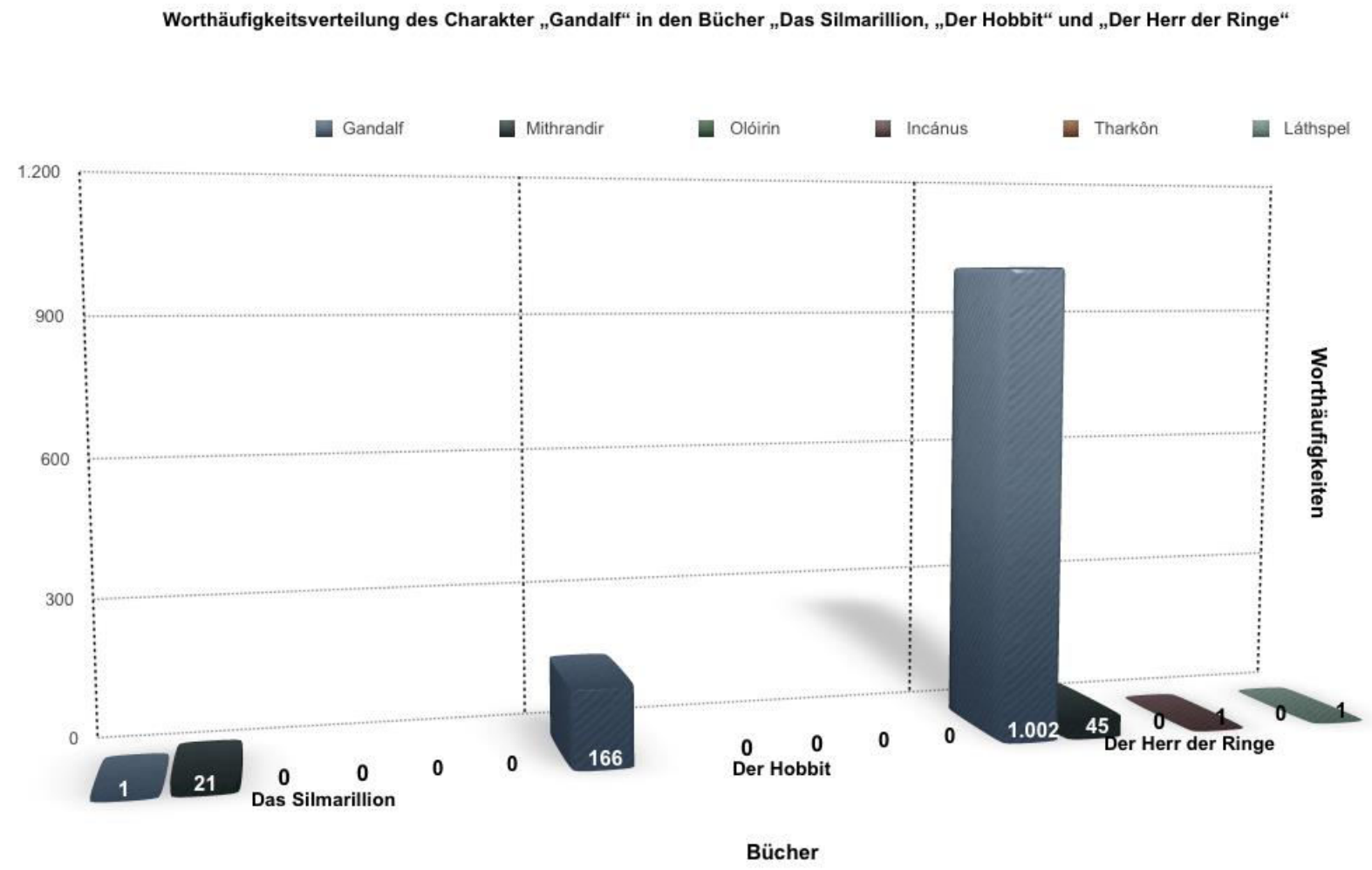


Abbildung 9: Worthäufigkeit „Gandalf“ in seinen verschiedenen Verkörperungen

Eine ähnliche Analyse soll ebenso am Beispiel „Frodo“ durchgeführt werden. In den zu untersuchenden Büchern „Das Silmarillion“, „Der Hobbit“ und „Der Herr der Ringe“ taucht der Charakter Frodo insgesamt 1.952 Mal auf. Aufgeschlüsselt auf die einzelnen Bücher ergibt sich folgendes Bild. Im literarischen Werk „Das Silmarillion“ taucht Frodo1 Mal auf, in „Der Hobbit“ 0 Mal und in „Herr der Ringe“ insgesamt 1951 Mal auf.

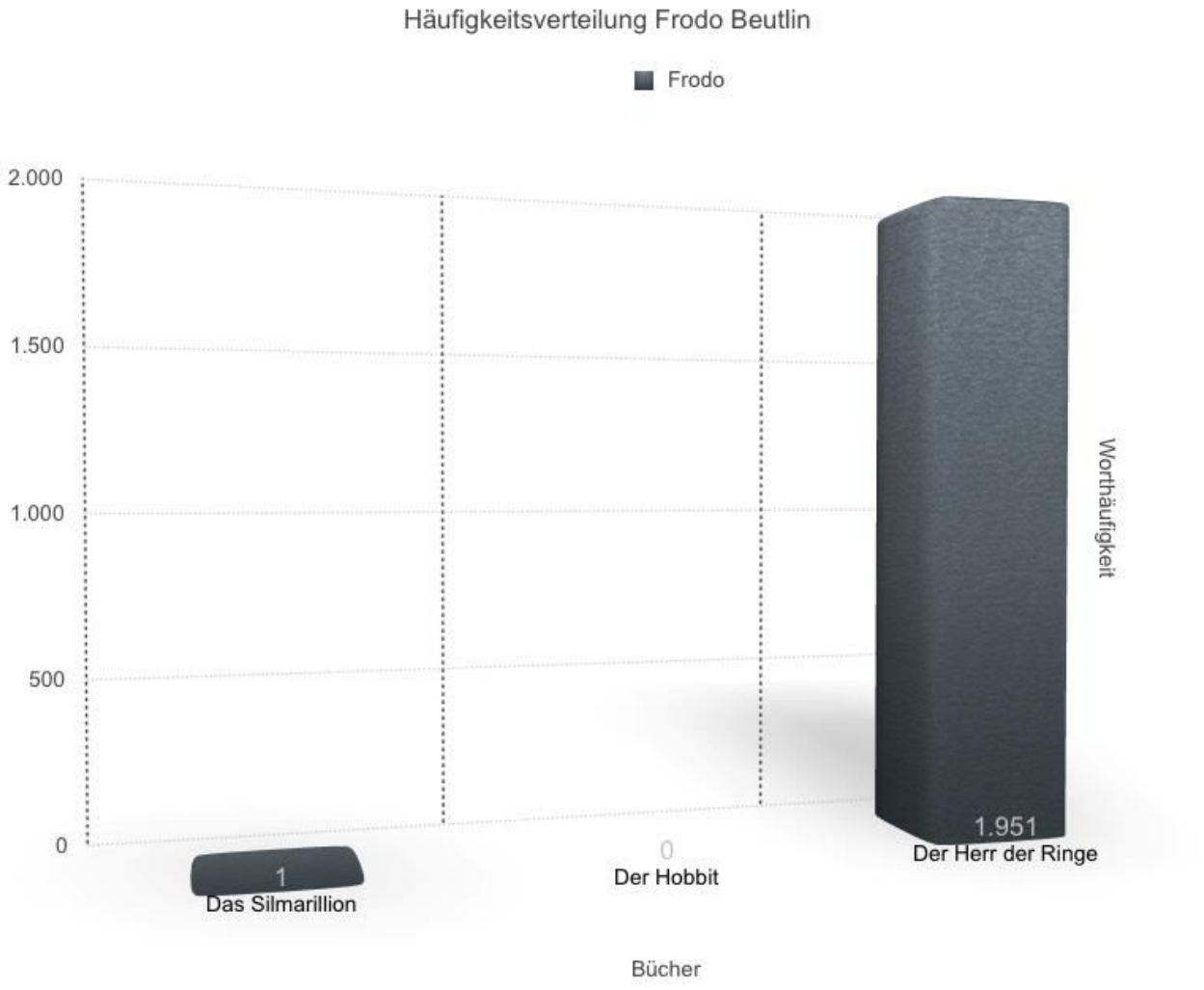


Abbildung 10: Worthäufigkeit „Frodo“

Die weiterführende Analyse, soll an dieser Stelle die Ort-Charakterrelationen für den Charakter Gandalf und Gandalf im Bezug zu allen Büchern betrachtet werden. Gandalf taucht dabei an folgenden Orten auf: in Anach, in Anduin, im Auenland, in Beutelsend, in Bree, in Bruch, in Bruchtal, am Caradhras, im Düsterwald, am Edoras,am Erech, im Fangorn, in Gondor, in den Häusern der Heilung, an Helms Tor, in Isengart ,in Ithilien, in Klamm, in Luch, in Minas Tirith, in Mindolluin, in den Minen von Moria, in Mordor, im Nebelgebirge, an der Oststraße, in Rohan, am Rothorn, am Rothornpass, am Umbar, im Wilderland, am Winkel. Frodo hingegen kann mit folgenden Orten in Verbindung gebracht werden. Der Hobbit, Frodo taucht an folgenden Orten auf: am Amon Hen, in Anach, im Auenland, am Bühl, in Beutelsend, im Bockland, am Brandywein, in Bree, am Bruch, in Bruchtal, am Cirith Ungol, im Düsterwald, am Ethel Dúath, am Erech, in Gondor, auf den Höhen, in Hobbingen, in Ithilien, in Helmsklamm, am Krickloch, in Lothlórien, in Luch, in Michelbingen, in Minas Tirith, in Mordor, im Nebelgebirge, im Ostviertel, in Rohan, in Sammath Saur, am Schicksalsberg, an den Schicksalsklüften, in der Schlucht, im Thal, in Tol Brandir, in Wasserau, auf der Wetterspitze, im Winkel, in Orodruin, in den Minen von Moria.

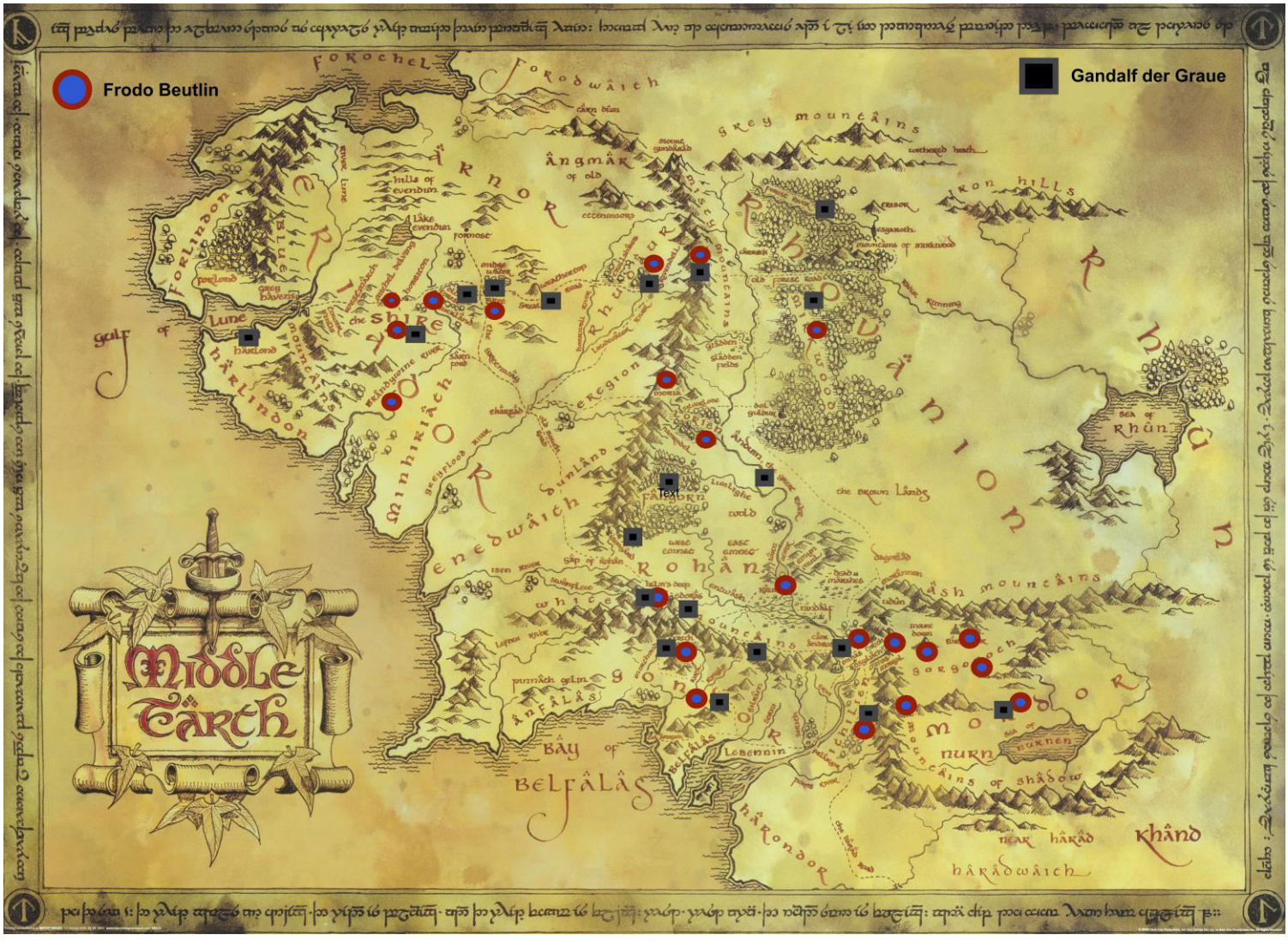


Abbildung 12: Ort-Charakterrelation Gandalf der Graue und Frodo Beutlin

6. Fazit

In der abschließenden Betrachtung des gesamten Projekts lassen sich folgende Themenschwerpunkte identifizieren: Zugänglichkeit der Daten, Anwendung der FAIR-Prinzipien und des Webbasierten RDMO Tools für private Forschungsprojekte, Probleme bei der Datenstruktur, Herausforderungen bei der Python Programmierung sowie die Datenauswertung. Wie im Kapitel 3 beschrieben, so berechtigt der §60 des Urhebergesetzes, nicht die Zugänglichkeit des Forschungsgegenstandes, sondern setzt das Vorliegen der Ressourcen voraus. Dies hat dementsprechende Auswirkungen auf den Forschungsprozess und kann zu Behinderungen in der Analyse der Daten führen. Im vorliegenden Projekt wurden aus privaten Finanzmitteln alle Materialien käuflich erworben, sowie einzelne Tools zur Analyse und Aufbereitung der Daten. Ebenso ist anzumerken, dass einige wichtigen Ressourcen für den deutschsprachigen Raum nicht in elektronischer Form, jedoch in englischer Sprache, vorliegen. Dementsprechend erschwert dieser Umstand die automatische Datenerhebung und Analyse. Das Formatieren und Umwandeln einzelner elektronischer Ressourcen stellt, durch bereits Open Source Verfügbaren Tools, keine Schwierigkeit dar, wodurch Datensätze automatisiert bzw. semi-automatisiert zugänglich und verarbeitbar gemacht werden können. Einzig und allein die Erhebung aus analogen Daten, die Bearbeitung und Säuberung von Daten nimmt einen Großteil des Arbeitsprozess in Anspruch und ist im Fall eines alleinigen Forschungsprozesses fehleranfällig. Ebenso die Verwendung unterschiedlicher Betriebssysteme zur Datenverarbeitung und Datendarstellung ist als fehleranfälligen Schwerpunkt identifizierbar. Um Fehler zu Vermeiden und einen organisierten, strukturierten Übersicht über einen großen Datenkorpus behalten zu können, bieten die FAIR Prinzipien in Verbindung mit der Nutzung des RDMO Tools einen großen Mehrwert und kann auch in privaten Forschungsprojekten angewandt werden. Jedoch können nicht alle Thematischen Schwerpunkte erfüllt werden, da hierbei die Infrastruktur eines Datenzentrums nicht vorhanden ist. Dementsprechend müssen Abstriche bei der Rechenpower, aufgrund des immensen Datenvolumens und der Erzeugung persistenter Identifier zur Langzeitarchivierung gemacht werden. Ebenso traten während der Programmierung des Python Skript diverse Herausforderungen auf, welche im Vorfeld nicht abzusehen waren. Einige der Herausforderungen sollen im Folgenden näher betrachtet werden. Einige der Herausforderungen sollen im Folgenden näher betrachtet werden. Bei einem alternativen Tokenisierungsverfahren bilden Folgen von Buchstaben ein Token, ebenso alle Folgen von Ziffern. Alle anderen Zeichen bilden für sich genommen ein Token. Dabei kommt es zu einen Herausforderungen, wenn Texte nach mehreren Keys und Values durchsucht werden sollen. Ist der Text Tokenisiert und alle Absätze und Leerzeichen konnten entfernt werden, befindet sich jedes Wort als einzelne Entität. Möchte man nun jedoch anhand der initiierten Namensliste im literarischen Text Charaktere finden, ist dies nur schwer möglich, da die eine Vielzahl der Charaktere Zusatznamen und diese aus mehreren Wörtern bestehen, wie zum Beispiel: „Eruin der I“ oder „der schwarze Fürst“. Das System splittet die einzelnen Wörter und zieht sie damit aus dem eigentlichen, aber für die Suche relevanten, Kontext. Ebenso gestaltet sich die Zusammenführung der einzelnen Zusatznahmen der Charaktere als Schwierig, da diese bis zu 25 Zusatznamen enthalten können. Ebenso konnten erhebliche Herausforderungen bei der Generierung des Skripts festgestellt werden. Aufgrund der geringen Rechenleistung stellte die immense Datenmenge von über 550.00ß0 Textzeilen, über 1000 Datensätzen, über 37.000 Sätzen und über 35.000 Charakter -Ort Kombinationen eine Herausforderung dar. Des Weiteren konnte keine Lemmatisieren des Textes vorgenommen, aufgrund des sprachlichen Problems der Elbischen Namen und Orte. Schlussendlich lässt sich Zusammenfassen, dass innerhalb einer Kooperation mit einer größeren Institution und mehr Zeit viele weitere Forschungsfragen beantwortet werden könnten und der derzeitige Code ein vorläufiges Ergebnis darstellt von dem was möglich sein kann.

7. Ausblick

Im nun folgenden Abschnitt soll ein Ausblick über weitere Möglichkeiten zur Erforschung der literarischen Werke von J.R.R. Tolkien gegeben werden. Weitere Forschungsprojekte und auf zeitlich unbegrenzte Vorhaben, könnten in erster Instanz die bisher noch unsauberen Daten in diesem Projekt bereinigen umso, sollte es darum gehen eine Gesamtanalyse der Populationsrate zu ermitteln, valide Daten zu erhalten. Zudem könnte auf längere Zeit der Aufbau eines vollständigen Charakters- und Ortsbezogenen Korpus angestrebt werden, um weiterführende Forschungsprozesse unterstützen zu können. Da der, für die Analyse verwendete Code, die Datensammlung aus ihrer Ursprungsdatei heraus lädt und diese in den Code indexiert, ist es möglich die Sammlung stetig zu erweitern und neue Orte und Personen hinzufügen zu können. Ebenso könnten weitere Daten gesammelt werden wie Gedichte, Lieder, Verwandtschaft Beziehungen, Geburtsjahre und vieles mehr. Ist ein solcher Korpus existent könnten in weiteren Projekten andere Werke analysiert werden, wie zum Beispiel die Versionsgeschichte der Geschichte rund um Beren und Luthien. Dabei sollten stehts Rechtliche Aspekte im Blick behalten werden und Forschungsergebnisse und /oder Forschungsdaten frei zugänglich veröffentlicht werden.

8.Literaturverzeichnis

A New Companion to Digital Humanities: Schreibman,Susan/ Siemens,Ray/Unsworth,John (Hg.):A New Companion to Digital Humanities. Chichester 2016.

Andreas Degkwitz, „Digitale Sammlungen – Vision eines Neubeginns“, Bibliothek Forschung und Praxis 38, Nr. 3 (19. Januar 2014): S. 413, https://doi.org/10.1515/bfp-2014-0064

Ardapedia (2015, Februar 18): Der Herr der Ringe (Namensübersetzungen)/Orte. Abgerufen 12. März 2020, von http://ardapedia.herr-der-ringe-film.de/index.php/Der_Herr_der_Ringe_(Namens%C3%BCbersetzungen)/Orte

Ardapedia (2015, Februar 18): Der Herr der Ringe (Namensübersetzungen)/Orte. Abgerufen 12. März 2020, von http://ardapedia.herr-der-ringe-film.de/index.php/Der_Herr_der_Ringe_(Namens%C3%BCbersetzungen)/Orte

Budapester Open Acess Erklärung: https://www.budapestopenaccessinitiative.org/translations/german-translation

Creative Commons (o.J.): Namensnennung 3.0 in Deutschland. URL: https://creativecommons.org/licenses/by/3.0/de/legalcode (16.06.2019)

Das Silmarillion: Tolkien, C., & Krege, W. (2010). *Das Silmarillion*. Stuttgart, Deutschland: Klett-Cotta.

Der Hobbit: Tolkien, J. R. R. (2010). *Der Hobbit: Oder Hin und zurück* (13., Aufl. Aufl.). Stuttgart, Deutschland: Klett-Cotta.

Der Herr der Ringe: Tolkien, J. R. R. (2009). *Der Herr der Ringe: Erster Teil: Die Gefährten. Zweiter Teil: Die zwei Türme. Dritter Teil: Die Rückkehr des Königs*. Stuttgart, Deutschland: Klett-Cotta.

Deutsche Tolkien Gesellschaft: https://www.tolkiengesellschaft.de/ueber-j-r-r-tolkien/rechtliche-infos-rund-um-tolkien/

Elsevier TDM Glossary: https://www.elsevier.com/__data/assets/pdf_file/0018/102906/TDM-Glossary.pdf

Forschungsdaten: Oltersdorf und Schmunk: Forschungsdaten, 2016, S.181.

Forschungsdatenmanagement: https://www.forschungsdaten.org/index.php/Forschungsdatenmanagement

Forschungsdatenmanagementplan: Aufbau einer Datensammlung, zur Realisierung einer Möglichkeit zur Identifikation der Ort-Charakter Relation in Tolkiens Mittelerde Werken.

Foster, R. (2002). *Das große Mittelerde-Lexikon. Ein alphabetischer Führer zur Fantasy-Welt von J.R.R. Tolkien.* Stuttgart, Deutschland: Lübbe.

Fonstad, K. W. (2014). *Historischer Atlas von Mittelerde* (16., völlig überarbeitete verb. Neuaufl. (REV). Aufl.). Stuttgart, Deutschland: Klett-Cotta.

Forschungsdaten.info, Text- und Data Mining: https://www.forschungsdaten.info/themen/rechte-und-pflichten/text-und-data-mining/

Gesetz über Urheberrecht und verwandte Schutzrechte §60 Text und Data Mining: https://www.gesetze-im-internet.de/urhg/60d.html

Glossar der DINI AG KIM (Kompetenzzentrum Interoperable Metadaten): https://www.kimforum.org/Subsites/kim/DE/Materialien/Glossar/glossar_node.html

Lord oft he Ring Project: http://lotrproject.com/

Patrick Sahle, „Digitale Editionen“, in Digital Humanities: eine Einführung, hg. von Fotis Jannidis, Hubertus Kohle, und Malte Rehbein (Stuttgart: J.B. Metzler Verlag, 2017), S. 234-239.

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata

Schriftverkehr Klett-Cotta Verlag

Vertrauenswürdige und abgesicherte Langzeitarchivierung multimedialer Inhalte: https://d-nb.info/1000084205/34, S. 25.

Zenodo: https://zenodo.org/

9.Abbbildungsverzeichnis

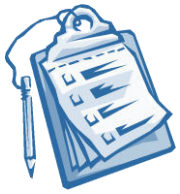
10. Eigenständigkeitserklärung

Hiermit versichere ich, dass ich Toni Matzdorf, die vorliegende Arbeit ohne Fremder Hilfe und nur die angegeben Hilfsmittel zur Erstellung der Hausarbeit verwendet habe.

Potsdam, 12.3.2020

11.Anhang

Kriterienkatalog zur Datenerhebung und Auswahl



	Thematischer Schwerpunkt	Akzeptanzkriterium
	Datenbasis	
	Literarische Werke	.
		Das Literarische Werk ist in deutscher Sprache Verfügbar
		Das Literarische Werk gehört zum Hauptkanon der Reihe.
		Das Literarische Werk beinhaltet wichtige Aspekte der Reihe.
		Das Literarische Werk ist eine Aufzählung von wichtigen Ereignissen der Hauptreihe.
		Das Literarische Werk wurde von anerkannten Übersetzern anhand von Linguistischen Standards übersetzt.
		Das Literarische Werk ist in der deutschsprachigen Community anerkannt.
		Das Literarische Werk ist im Handel frei verfügbar.
		Das Literarische Werk ist in einem elektronischen Format verfügbar.
		Das Literarische Werk ist maschinell verarbeitbar.
		Das Literarische Werk bietet einen größeren Mehrwert, auch wenn es nicht digital verfügbar ist.
	Ort- und Charakterbezogene Daten	
		Die Daten müssen Ortsbezogene Entitäten beinhalten.
		Die Daten müssen Charakterbezogene Entitäten beinhalten.
		Die Daten müssen frei zugänglich sein.
		Die Daten müssen in einer strukturellen Form bereits erhoben sein.
		Die Daten müssen maschinell verarbeitbar ein.

Forschungsdatenmanagementplan

Aufbau einer Datensammlung, zur Realisierung einer Möglichkeit zur Identifikation der Ort-Charakter Relation in Tolkiens Mittelerde Werken.

Wie lautet die primäre Forschungsfrage des Projektes?

Erstellung einer digitalen Sammlung, zur Realisierung einer Identifikationsmöglichkeit der Ort-Charakter Relation in Tolkiens Mittelerde Werken, zur Ermittlung der Ortsbezogenen Populationrate.

Welcher Disziplin / welchen Disziplinen ist das Projekt zuzuordnen?

Geistes- und Sozialwissenschaften / Literaturwissenschaft

Wann beginnt die Projektlaufzeit?

31. Januar 2020

Wann endet die Projektlaufzeit?

13. März 2020

Gibt es von weiteren Seiten (z. B. von der Fachcommunity) Anforderungen an das Datenmanagement, die beachtet werden müssen?

Ja

Welche Anforderungen an das Datenmanagement sind dies?

1. Die erhobenen Daten sollten sich in ihren Grundzügen auf das Prinzip der FAIR-Principles beziehen und anhand dieser angewendet werden.

2. Die verwendeten Methoden und Tools in diesem Projekt sollten sich auf die Verwendung von Python, RDF sowie standardisierter Software aus dem Bereich der Digital Humanities beschränken

3. Da in diesem Projekt urheberrechtlich geschützte Literatur analysiert wird, sollte im Folgenden auf die Rechtslage hingewiesen werden.

4. Die nicht durch das Urheberrecht geschützten Daten sollten auf ein öffentlich zugängliches Repository abgelegt werden und unter einer Open Access Lizenz gestellt werden.

Um was für einen Datensatz handelt es sich?

Das Silmarillion:

Bei diesem Datensatz handelt es sich um das Literarische Werk von J.R.R Tolkien "Das Silmarillion" in der deutschen Übersetzung von Wolfang Krege (2010), erschienen im Klett-Cotta Verlag. Dieser Datensatz wurde im Vorfeld mittels Software vom Ausgangsformat PDF in das Datei Format ".txt" formatiert. hierbei wurde darauf geachtet, dass die Zeichencodierung im Bereich UTF-8 liegt.

Der Hobbit:

Bei diesem Datensatz handelt es sich um das Literarische Werk von J.R.R Tolkien "Der Hobbit" in der deutschen Übersetzung von Wolfang Krege (2012), erschienen im Klett-Cotta Verlag. Dieser Datensatz wurde im Vorfeld mittels Software vom Ausgangsformat PDF in das Datei Format ".txt" formatiert. hierbei wurde darauf geachtet, dass die Zeichencodierung im Bereich UTF-8 liegt.

Der Herr der Ringe:

Bei diesem Datensatz handelt es sich um das Literarische Werk von J.R.R Tolkien "Der Herr der Ringe" in der deutschen Übersetzung von Margaret Carroux (2010), erschienen im Klett-Cotta Verlag. Dieser Datensatz wurde im Vorfeld mittels Software vom Ausgangsformat PDF in das Datei Format ".txt" formatiert. hierbei wurde darauf geachtet, dass die Zeichencodierung im Bereich UTF-8 liegt.

Ortsbezogene Einheiten aus Mittelerde:

Bei diesem Datensatz handelt es sich um eine in UTF-8 formatierte CSV Tabelle mit 916 individuellen Datensätzen zu allen historisch wichtigen Bezugsorten innerhalb der Bücher "Das Silmarillion", "Der Hobbit" und "Der Herr der Ringe".

Charakterbezogene Einheiten aus Mittelerde:

Bei diesem Datensatz handelt es sich um eine in UTF-8 formatierte CSV Tabelle mit 605 individuellen Datensätzen zu allen historisch wichtigen Charakteren innerhalb der Bücher "Das Silmarillion", "Der Hobbit" und "Der Herr der Ringe".

Projekt Mittelerde (RDF):

Bei diesem Datensatz handelt es sich um die Beschreibung der Metadaten der einzelnen Datensätze innerhalb eines RDF Schemata. Zudem wurde hier exemplarisch am Beispiel des Charakter "Gandalf" Metadaten zu seiner Figur und den Orten an deren er zugegen war festgehalten.

Digitalisierte Landkarte Mittelerde:

Bei diesem Datensatz handelt es sich um ein Digitalisat des Close Up Poster Herr der Ringe - Karte von Mittelerde Riesenformat 135,5 x 98cm. Hergestellt im Digitalisierungslabor der Fachhochschule Potsdam, im Format "JPG" und "TIF".

Pythoncode für die Datenauswertung:

Bei diesem Datensatz handelt es sich um das ausführende Pythonscript für die Auswertung der Fragestellung. Zusätzlich zum Code wurden hierbei Kommentare verfasst um den Code nachvollziehen zu können.

Ort-Charakter Visualisierung:

Bei diesem Datensatz handelt es sich um die Visualisierung der Ort-und Charakterkonstellation für einen ausgewählten Charakter aus dem gesamten Datenset.

Wird der Datensatz selbst erzeugt oder nachgenutzt?

Das Silmarillion*:* Erzeugt

Der Hobbit*:* Erzeugt

Der Herr der Ringe: Erzeugt

Ortsbezogene Einheiten aus Mittelerde: Erzeugt

Charakterbezogene Einheiten aus Mittelerde: Erzeugt

Projekt Mittelerde (RDF): Erzeugt

Digitalisierte Landkarte Mittelerde*:* Erzeugt

Pythoncode für die Datenauswertung*:* Erzeugt

Ort-Charakter Visualisierung*:* Erzeugt

Was ist die tatsächliche oder erwartete Größe des Datensatzes?

Das Silmarillion*:* genau: 758 KB

Der Hobbit*:* genau: 594 KB

Der Herr der Ringe: genau: 2,9 MB

Ortsbezogene Einheiten aus Mittelerde: genau: 16 KB

Charakterbezogene Einheiten aus Mittelerde: genau: 29 KB

Projekt Mittelerde (RDF): genau: 8 KB

Digitalisierte Landkarte Mittelerde*:* genau: 1,09 GB

Pythoncode für die Datenauswertung*:* 1 GB bis 1 TB

Ort-Charakter Visualisierung*:* weniger als ein GB

In welchen Formaten liegen die Daten vor?

Das Silmarillion:

Die Originale Textdatei liegt als kfx-zip (Amazone) Datei vor. Diese Datei wurde mit digitalen Softwarekomponenten in PDF sowie in eine UTF-8 formatierte Datei umgewandelt.

Der Hobbit:

Die Originale Textdatei liegt als kfx-zip (Amazone) Datei vor. Diese Datei wurde mit digitalen Softwarekomponenten in PDF sowie in eine UTF-8 formatierte Datei umgewandelt.

Der Herr der Ringe:

Die Originale Textdatei liegt als kfx-zip (Amazone) Datei vor. Diese Datei wurde mit digitalen Softwarekomponenten in PDF sowie in eine UTF-8 formatierte Datei umgewandelt.

Ortsbezogene Einheiten aus Mittelerde:

Die Originaldateiliegt als eine UTF-8 CSV Tabelle vor.

Charakterbezogene Einheiten aus Mittelerde

Die Originaldateiliegt als eine UTF-8 CSV Tabelle vor.

Projekt Mittelerde (RDF):

Die Originaldateiliegt als eine UTF-8 RDF Datei vor.

Digitalisierte Landkarte Mittelerde*:*

Das Digitalisat des Close Up Poster Herr der Ringe - Karte von Mittelerde Riesenformat 135,5 x 98cm. liegt, im Format "JPG" und "TIF" vor.

Pythoncode für die Datenauswertung*:*

Die Code-Datei für die Realisierung der Identifikationsmöglichkeit der Ort-und Charakterkonstellationen liegt als Python Dokument vor.

Ort-Charakter Visualisierung:

Die Visualisierung der Ort-Charakterkonstellation eines Charakters aus dem Datenset liegt als Bilddatei im "JPG" Format vor.

Welche Instrumente, Software, Technologien oder Verfahren werden zur Erzeugung oder Erfassung der Daten genutzt?

Das Silmarillion*:*

Um die Ursprungsdatei in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels "Calibre" zu konvertieren. Dabei wurde der Text von PDF zu ".TXT" umgewandelt und in UTF-8 abgespeichert. Daraufhin wurde für die weitere Verarbeitung die Anhänge, das Impressum und die Titelseiten gelöscht, da es für das Projekt wichtig ist den reinen Literarischen Text analysieren zu können.

Der Hobbit:

Um die Ursprungsdatei in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels "Calibre" zu konvertieren. Dabei wurde der Text von PDF zu ".TXT" umgewandelt und in UTF-8 abgespeichert. Daraufhin wurde für die weitere Verarbeitung die Anhänge, das Impressum und die Titelseiten gelöscht, da es für das Projekt wichtig ist den reinen Literarischen Text analysieren zu können.

Der Herr der Ringe:

Um die Ursprungsdatei in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels "Calibre" zu konvertieren. Dabei wurde der Text von PDF zu ".TXT" umgewandelt und in UTF-8 abgespeichert. Daraufhin wurde für die weitere Verarbeitung die Anhänge, das Impressum und die Titelseiten gelöscht, da es für das Projekt wichtig ist den reinen Literarischen Text analysieren zu können.

Ortsbezogene Einheiten aus Mittelerde:

Um die Daten in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels eines Webscrapping Tools (Table to Excel) aus dem Internet zu extrahieren. Zur Extraktion der Orte aus den Werken von J.R.R Tolkien wurde sich für die Website "Ardapedia" entschieden. Diese Forums basierte Website genießt unter in der deutschsprachigen "Mittelerde" Community ein hohes Ansehen und bietet vertrauensvolle und auf dem "Vier-Augen-Prinzip" basierte Daten an. Des Weiteren bezieht diese aktive Community ihre Daten aus mehren vertrauensvollen Quellen wie zum Beispiel aus den bisher veröffentlichen Büchern, der Deutschen Tolkien Gesellschaft und vielen weiteren Quellen. Anschließend an die Extraktion der Daten, wurden diese in eine "csv" formatierte Tabelle (Excel, Numbers) überführt und normalisiert. Zur Validierung und Erweiterung der Daten wurden diese händisch mit analogen Quellen abgeglichen (Lexikon, Atlas,).

Charakterbezogene Einheiten aus Mittelerde

Um die Daten in ein maschinenlesbares Format formatieren zu können, war es notwendig diese mittels der Exportfunktion vom Browser "Safari" in PDF abzuspeichern. Im darauf folgenden Schritt wurde diese Datei unter zur Hilfenahme der Software "Calibre" in Word exportiert, und anhand von Funktionen innerhalb der Software gesäubert. Anschließend wurde diese Datei in eine "csv" Datei umgewandelt und normalisiert. Weitere Daten wurden aus den bereits im Vorfeld bearbeiteten literarischen Werken entnommen, da diese in eiigen Fällen im Anhang eine Charakterliste beinhaltete. Zur Extraktion der Orte aus den Werken von J.R.R Tolkien wurde sich für die Website "Ardapedia" entschieden. Diese Forums basierte Website genießt unter in der deutschsprachigen "Mittelerde" Community ein hohes Ansehen und bietet vertrauensvolle und auf dem "Vier-Augen-Prinzip" basierte Daten an. Des Weiteren bezieht diese aktive Community ihre Daten aus mehren vertrauensvollen Quellen wie zum Beispiel aus den bisher veröffentlichen Büchern, der Deutschen Tolkien Gesellschaft und vielen weiteren Quellen.Zur Validierung und Erweiterung der Daten wurden diese händisch mit analogen Quellen abgeglichen (Lexikon, Atlas,).

Projekt Mittelerde (RDF):

Um die notwendigen Metadaten in ein maschinenlesbares Format formatieren zu können, mussten diese im ersten Schritt erfasst und auf der Grundlage eines standardisierten Vokabulars erfasst werden. Um alle notwendigen Daten zugänglich zu machen, wurden diese mit Hilfe der Software "Oxygen" erfasst und anhand der DNB und der DBpedia Ontologie sowie der nach der W3C standardisierten FOAF Syntax beschreiben.

Digitalisierte Landkarte Mittelerde*:*

Das Digitalisat des Close Up Poster Herr der Ringe - Karte von Mittelerde Riesenformat 135,5 x 98cm. liegt, im Format "JPG" und "TIF" vor und wurde im Digitalsierungslabor der FH Potsdam angefertigt. Bearbeitet wurde das Digitalisierte Objekt in Photoshop.

Pythoncode für die Datenauswertung*:*

Zur Erzeugung dieses Datensatzes wurde das Programm "PyCharm" verwendet unter der Beachtung einer Virtuellen Umgebung und unter der Verwendung der Python Version 3.7.

Ort-Charakter Visualisierung:

Die Visualisierung der Ort-Charakterkonstellation eines Charakters aus dem Datenpool wurde mit Hilfe eines Web basierten Datenvisualisierungstools angefertigt.

Welche Software, Verfahren oder Technologien sind notwendig, um die Daten zu nutzen?

Das Silmarillion:

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor, einen standardisierten Browser, die Software PyCharm in der Community Edition, mindestens Python 3.7 sowie die Komponenten: Python NLTK, Pandas, Anaconda, Jupiter Notebook und Virtualenv.

Der Hobbit:

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor, einen standardisierten Browser, die Software PyCharm in der Community Edition, mindestens Python 3.7 sowie die Komponenten: Python NLTK, Pandas, Anaconda, Jupiter Notebook und Virtualenv.

Der Herr der Ringe:

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor, einen standardisierten Browser, die Software PyCharm in der Community Edition, mindestens Python 3.7 sowie die Komponenten: Python NLTK, Pandas, Anaconda, Jupiter Notebook und Virtualenv.

Ortsbezogene Einheiten aus Mittelerde:

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor, einen standardisierten Browser oder Excel.

Charakterbezogene Einheiten aus Mittelerde

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor, einen standardisierten Browser oder Excel.

Projekt Mittelerde (RDF):

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: Texteditor oder ein standardisierter Browser.

Digitalisierte Landkarte Mittelerde*:*

Um die Dateien Nachnutzen zu können muss auf dem zu verwendenden Betriebssystem ein Rasterbildverarbeitendes Programm zur Verfügung stehen.

Pythoncode für die Datenauswertung*:*

Um die Dateien Nachnutzen zu können müssen folgende Softwarekomponenten zur Verfügung stehen: die Software PyCharm in der Community Edition, mindestens Python 3.7 sowie die Komponenten: Python NLTK, Pandas, Anaconda, Jupiter Notebook und Virtualenv.

Ort-Charakter Visualisierung:

Um die Dateien Nachnutzen zu können muss auf dem zu verwendenden Betriebssystem ein Rasterbildverarbeitendes Programm zur Verfügung stehen.

Welche Technologie bzw. welches Tool wird zur Versionierung verwendet?

Das Silmarillion:

Versionskontrollsystem: Artikel_Name_Zustand (komplett, oder <<leer>> für das schnellere Ansteuern)

Der Hobbit:

Versionskontrollsystem: Artikel_Name_Zustand (komplett, oder <<leer>> für das schnellere Ansteuern)

Der Herr der Ringe:

Versionskontrollsystem: Artikel_Name_Zustand (komplett, oder <<leer>> für das schnellere Ansteuern)

Ortsbezogene Einheiten aus Mittelerde:

Versionskontrollsystem: Orte_Mittelerde_Zustand (Formatiert, EndResult, unvollständig, <<Leer>> für eine schnellere Ansteuerung)

Charakterbezogene Einheiten aus Mittelerde:

Charaktere_Mittelerde_Zustand (Formatiert, EndResult, unvollständig, <<Leer>> für eine schnellere Ansteuerung)

Projekt Mittelerde (RDF):

Name_der_Datei_yy_mm_dd

Digitalisierte Landkarte Mittelerde:

Versionskontrollsystem: Name_der_Datei_yy_mm_dd

Pythoncode für die Datenauswertung:

Name_der_Datei_Versionsnummer

Ort-Charakter:

Einfaches Kopieren

In welchem Umfang werden Infrastrukturressourcen benötigt (CPU-Stunden, Bandbreite, Speicherplatz etc.)?

Das Silmarillion*:*

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Der Hobbit:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Der Herr der Ringe:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Ortsbezogene Einheiten aus Mittelerde:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Charakterbezogene Einheiten aus Mittelerde:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Projekt Mittelerde (RDF):

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Digitalisierte Landkarte Mittelerde:

Es werden folgende Infrastrukturressourcen benötigt: Digitalsierungslabor der FH Potsdam

Pythoncode für die Datenauswertung:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Ort-Charakter Visualisierung:

Die üblichen Infrastrukturressourcen des Arbeitsplatzes reichen aus.

Gibt es beabsichtigte (ggf. auch potentielle) Nutzungsszenarien, für die die Unterstützung durch Datenmanagement- oder IT-ExpertInnen sinnvoll oder notwendig ist?

Das Silmarillion*:* Nein

Der Hobbit*:* Nein

Der Herr der Ringe: Nein

Ortsbezogene Einheiten aus Mittelerde: Nein

Charakterbezogene Einheiten aus Mittelerde: Nein

Projekt Mittelerde (RDF): Nein

Digitalisierte Landkarte Mittelerde: Ja, Digitalisierung der originalen Ressource

Pythoncode für die Datenauswertung: Nein

Ort-Charakter Visualisierung: Nein

Wo wird der Datensatz während des Projektes gespeichert?

Das Silmarillion:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Der Hobbit:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Der Herr der Ringe:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Ortsbezogene Einheiten aus Mittelerde:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Charakterbezogene Einheiten aus Mittelerde:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Projekt Mittelerde (RDF):

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Digitalisierte Landkarte Mittelerde:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Pythoncode für die Datenauswertung:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Ort-Charakter Visualisierung:

Der Datensatz wird neben der Speicherung auf dem eigenen Arbeitsrechner zusätzlich auf einem Google Drive Server und einer externen Festplatte gesichert.

Wie und wie oft werden Backups der Daten erstellt?

Das Silmarillion:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird. Die Backup Daten werden nicht in das Repository übernommen.

Der Hobbit:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird. Die Backup Daten werden nicht in das Repository übernommen.

Der Herr der Ringe:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird. Die Backup Daten werden nicht in das Repository übernommen.

Ortsbezogene Einheiten aus Mittelerde:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird.

Charakterbezogene Einheiten aus Mittelerde:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird.

Projekt Mittelerde (RDF):

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird.

Digitalisierte Landkarte Mittelerde:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird. Die Backup Daten werden nicht in das Repository übernommen.

Pythoncode für die Datenauswertung:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird.

Ort-Charakter Visualisierung:

Ein Backup der Daten wird jeweils nach einer Änderung der Daten vorgenommen, dabei werden die alten Daten in ihrer jeweiligen Version in einen neuen Ordner verschoben um die neueren Versionen im Arbeitsordner speichern zu können. So ist gewährleistet das Änderungen Rückgängig gemacht werden können, wenn dies als Notwendig erachtet wird.

Wer ist verantwortlich für die Erstellung der Backups?

Das Silmarillion:

- Toni Matzdorf

Der Hobbit:

- Toni Matzdorf

Der Herr der Ringe:

- Toni Matzdorf

Ortsbezogene Einheiten aus Mittelerde:

- Toni Matzdorf

Charakterbezogene Einheiten aus Mittelerde:

- Toni Matzdorf

Projekt Mittelerde (RDF):

- Toni Matzdorf

Digitalisierte Landkarte Mittelerde:

- Toni Matzdorf

Digitalisierte Landkarte Mittelerde:

- Toni Matzdorf

Ort-Charakter Visualisierung:

- Toni Mazdorf

Welche Maßnahmen zur Gewährleistung der Datensicherheit werden getroffen (z. B. Schutz vor unbefugtem Zugriff, Datenwiederherstellung, Übertragung sensibler Daten)?

Das Silmarillion*:*

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Der Hobbit:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Der Herr der Ringe:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Ortsbezogene Einheiten aus Mittelerde:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Charakterbezogene Einheiten aus Mittelerde:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Projekt Mittelerde (RDF):

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Digitalisierte Landkarte Mittelerde:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Pythoncode für die Datenauswertung:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Ort-Charakter Visualisierung:

Durch das Sichern der Backup Daten in einzelnen Ordnern währen der Bearbeitungszeit ist es möglich Daten wiederherstellen zu können. Diese, sowie weitere Daten werden durch einen Passwortgeschützten Zugriff vor unbefugten Zugriff geschützt.

Soll dieser Datensatz veröffentlicht oder geteilt werden?

Das Silmarillion: Nein

Der Hobbit*:* Nein

Der Herr der Ringe: Nein

Ortsbezogene Einheiten aus Mittelerde: Ja, extern für alle

Charakterbezogene Einheiten aus Mittelerde: Ja, extern für alle

Projekt Mittelerde (RDF): Ja, extern für alle

Digitalisierte Landkarte Mittelerde: Nein

Pythoncode für die Datenauswertung: Ja, extern für alle

Ort-Charakter Visualisierung: Ja, extern für alle

Wenn nicht, begründen Sie dies bitte und unterscheiden Sie dabei zwischen rechtlichen und/oder vertraglichen Gründen und freiwilligen Einschränkungen.

Das Silmarillion:

Durch Urheberrechtliche Einschränkungen ist es nicht Möglich diese Datei der Öffentlichkeit zur Verfügung zu stellen.

Der Hobbit:

Durch Urheberrechtliche Einschränkungen ist es nicht Möglich diese Datei der Öffentlichkeit zur Verfügung zu stellen.

Der Herr der Ringe:

Durch Urheberrechtliche Einschränkungen ist es nicht Möglich diese Datei der Öffentlichkeit zur Verfügung zu stellen.

Digitalisierte Landkarte Mittelerde:

Durch Urheberrechtliche Einschränkungen ist es nicht Möglich diese Datei der Öffentlichkeit zur Verfügung zu stellen.

Welche Metadaten werden automatisch erhoben?

Das Silmarillion:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Der Hobbit:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Der Herr der Ringe:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Ortsbezogene Einheiten aus Mittelerde:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Charakterbezogene Einheiten aus Mittelerde:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Projekt Mittelerde (RDF):

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Digitalisierte Landkarte Mittelerde:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Pythoncode für die Datenauswertung:

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Ort-Charakter Visualisierung*:*

Folgende Metadaten werden automatisch erhoben: Art des Textdokuments, die Größe der Datei, Speicherort, Erstellungsdatum sowie Änderungsdaten.

Welche Metadaten werden semi-automatisch erhoben?

Das Silmarillion:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Der Hobbit:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Der Herr der Ringe:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Ortsbezogene Einheiten aus Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Charakterbezogene Einheiten aus Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Projekt Mittelerde (RDF):

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Digitalisierte Landkarte Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Pythoncode für die Datenauswertung:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Ort-Charakter Visualisierung:

Es beseht keine Notwendigkeit für den weiteren Prozess Metadaten semi-automatisch zu erheben.

Welche Metadaten werden manuell erhoben?

Das Silmarillion:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Der Hobbit:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Der Herr der Ringe:

Es beseht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Ortsbezogene Einheiten aus Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Charakterbezogene Einheiten aus Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Projekt Mittelerde (RDF):

Es beseht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Digitalisierte Landkarte Mittelerde:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Pythoncode für die Datenauswertung:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Ort-Charakter Visualisierung:

Es besteht keine Notwendigkeit für den weiteren Prozess Metadaten manuell zu erheben.

Wie sind die Daten strukturiert? In welchem Verhältnis stehen die einzelnen Komponenten zueinander? In welchem Verhältnis steht der Datensatz zu anderen im Projekt erhobenen oder genutzten Datensätzen?

Das Silmarillion:

Die Daten in dieser Datei befinden sich in einem Fließtext, getrennt durch Leerzeichen, Absätzen und Interpunktionen. Dargestellt werden diese in UTF-8. Die Metadaten des Literarischen Inhaltes werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Der Hobbit:

Die Daten in dieser Datei befinden sich in einem Fließtext, getrennt durch Leerzeichen, Absätzen und Interpunktionen. Dargestellt werden diese in UTF-8. Die Metadaten des Literarischen Inhaltes werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Der Herr der Ringe: Die Daten in dieser Datei befinden sich in einem Fließtext, getrennt durch Leerzeichen, Absätzen und Interpunktionen. Dargestellt werden diese in UTF-8. Die Metadaten des Literarischen Inhaltes werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Ortsbezogene Einheiten aus Mittelerde:

Die Daten in dieser Datei befinden sich in einer strukturierten Tabellenform, getrennt durch Zeilen und Spalten. Dargestellt werden diese in UTF-8. Die Metadaten des dieser Datei werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Charakterbezogene Einheiten aus Mittelerde:

Die Daten in dieser Datei befinden sich in einer strukturierten Tabellenform, getrennt durch Zeilen und Spalten. Dargestellt werden diese in UTF-8. Die Metadaten des dieser Datei werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Projekt Mittelerde (RDF):

Die Daten in dieser Datei befinden sich in einer strukturierten RDF/XML Form. Dargestellt werden diese in UTF-8. Die Metadaten dieser Datei werden in der RDF Datei "Projekt Mittelerde" repräsentiert.

Digitalisierte Landkarte Mittelerde:

Die digitalisierte Karte von Mittelerde, wird hierbei zum einen als". jpg", als auch in ".tif" Format dargestellt.

Pythoncode für die Datenauswertung:

Die Daten in dieser Datei befinden sich in einer strukturierten Python kodierten Form.

Ort-Charakter Visualisierung:

Die Visualisierung der Orts- Charakterkonstellation, wird hierbei zum einen als". jpg", als auch in ".tif" Format dargestellt.

Welches System von persistenten Identifikatoren soll genutzt werden?

Das Silmarillion: Nein

Der Hobbit*:* Nein

Der Herr der Ringe: Nein

Ortsbezogene Einheiten aus Mittelerde: URL und DOI

Charakterbezogene Einheiten aus Mittelerde: URL und DOI

Projekt Mittelerde (RDF): URL und DOI

Digitalisierte Landkarte Mittelerde: *Nein*

Pythoncode für die Datenauswertung: URL und DOI

Ort-Charakter Visualisierung*:* URL und DOI

Enthält dieser Datensatz personenbezogene Daten?

Das Silmarillion: Nein

Der Hobbit: Nein

Der Herr der Ringe*:* Nein

Ortsbezogene Einheiten aus Mittelerde: Nein

Charakterbezogene Einheiten aus Mittelerde*:* Nein

Projekt Mittelerde (RDF): Ja

Digitalisierte Landkarte Mittelerde*:* Nein

Pythoncode für die Datenauswertung: Nein

Ort-Charakter Visualisierung*:* Nein

Enthält dieser Datensatz nicht-personenbezogene sensible Daten?

Das Silmarillion*:* Nein

Der Hobbit: Nein

Der Herr der Ringe: Nein

Ortsbezogene Einheiten aus Mittelerde: Nein

Charakterbezogene Einheiten aus Mittelerde: Nein

Projekt Mittelerde (RDF): Nein

Digitalisierte Landkarte Mittelerde: Nein

Pythoncode für die Datenauswertung*:* Nein

Ort-Charakter Visualisierung: Nein

Werden Daten genutzt und/oder erstellt, die durch Urheber- oder verwandte Schutzrechte geschützt sind?

Ja

Muss dieser Datensatz langfristig aufbewahrt werden?

***Das Silmarillion**:* Nein

Der Hobbit: Nein

Der Herr der Ringe*:* Nein

Ortsbezogene Einheiten aus Mittelerde: Ja

Charakterbezogene Einheiten aus Mittelerde*:* Ja

Projekt Mittelerde (RDF): Ja

Digitalisierte Landkarte Mittelerde: Nein

Pythoncode für die Datenauswertung*:* Ja

Ort-Charakter Visualisierung: Ja

Wie lange müssen die Daten aufbewahrt werden?

Ortsbezogene Einheiten aus Mittelerde: Für immer.

Charakterbezogene Einheiten aus Mittelerde*:* Für immer.

Projekt Mittelerde (RDF): Für immer.

Pythoncode für die Datenauswertung: Für immer.

Ort-Charakter Visualisierung: Für immer.

Wie lang sollen die Daten nach Projektende (nach)nutzbar sein?

Ortsbezogene Einheiten aus Mittelerde: Für immer.

Charakterbezogene Einheiten aus Mittelerde*:* Für immer.

Projekt Mittelerde (RDF): Für immer.

Pythoncode für die Datenauswertung: Für immer.

Ort-Charakter Visualisierung: Für immer.

Schriftverkehr Klettt-Cotta/ Hobbit-Press

Absender: Toni Matzdorf

Empfänger: info@klett-cotta.de

Datum: 28.11.2019

Sehr geehrte Damen und Herren des Klett-Verlages,

mein Name ist Toni Matzdorf und ich befinde mich derzeit im 2. Mastersemesters der Informationswissenschaften. Im Rahmen des Kurses „Digitale Sammlungen“ haben wir als Semesteraufgabe eine selbsterstellte digitale Sammlung auszuwerten und anhand von Text und Data Mining zu Analysen (Bspw. Wie oft kommt das Wort x in einem Werk vor? oder Wie oft kommt Person x in einem Werk vor). Gern würde ich meine Analyse am Beispiel von Tolkiens Werken (Silmarillion, Hobbit, Kinder Hurins, Beren und Luthien, Der Fall von Gondolin, Herr der Ringe 1-3) durchführen. Mir liegen alle Werke als E-Book in digitaler Form vor. Da das Text und Data Mining in Deutschland noch nicht ausschließlich gesetzlich geregelt ist und man im Zweifelsfall immer den Verlag/ Urheber um Erlaubnis fragen muss (in UK ist das nicht-kommerzielle TDM bereits kostenlos erlaubt) schreibe ich Ihnen diese E-Mail. Das Projekt beschränkt sich ausschließlich auf eine Hausarbeit in einem Universitären Rahmen und wird nicht veröffentlicht und fällt somit unter die nicht-kommerzielle Nutzung. Zudem würden die Daten nur temporär, für die Zeit der Bearbeitung, gespeichert werden und danach gelöscht werden. Daher würde ich sie gerne um eine Erlaubnis bitten mit Digitalen Tools Daten, Personen, Charakter und Ortsdaten aus dem Werk, zu extrahieren.

Für weitere Informationen des Rechtlichen Rahmens siehe folgenden Link sowie UrhG § 60d: https://datenbank.nwb.de/Dokument/Anzeigen/146210_60d/

Ich würde mich sehr über eine positive Antwort freuen.

Mit freundlichen Grüßen
Toni Matzdorf

Absender: Winter Brigitte

Empfänger: Toni Matzdorf

Datum: 2.12.2019

Sehr geehrter Herr Matzdorf,

vielen Dank für Ihre Anfrage zur Nutzung der Tolkien-E-books für Ihre Hausarbeit.

Da sich das Projekt ausschließlich auf Ihre Hausarbeit bezieht und nicht veröffentlicht oder sonst kommerziell genutzt werden wird, freuen wir uns, Ihnen zu bestätigen, dass wir mit einer Nutzung der E-books im Rahmen des § 60d UrhG einverstanden sind.

Für ein Exemplar Ihrer Arbeit als PDF, wenn sie dann fertig ist, sind wir Ihnen dankbar.

Mit freundlichen Grüßen
Brigitte Winter