

Lernportfolio MLM

Maja Hoffmann

Wintersemester 21/22

Inhaltsverzeichnis

Woche 1 - Einführung und Organisatorisches	3
Woche 2 - Web Of Data	4
Woche 3 - Resource Description Framework	6
Woche 4 - Gastvortrag und praktische Übung	7
Woche 5 - DBpedia	8
Woche 6 - OpenRefine	10
Woche 7 - Gastvortrag	11
Woche 8 - Ontologies as Common Language	12
Woche 9 - Linked Open Data	13
Woche 10 - Klassifikation/Anreicherung, Datenqualität, Weiterentwicklung/Reparatur/Collaboration	15
Woche 11 - Gastvortrag	16
Woche 12 - Zusammenfassung	17
Glossar	19
Mindmap	21

Woche 1 - Einführung und Organisatorisches

13.10.2021

Organisatorisches

Der erste Satz im Lernportfolio.

Ich bin gespannt auf diese Prüfungsform. Ich glaube, dass es sehr helfen kann sich regelmäßig und schriftlich mit den Vorlesungsthemen auseinanderzusetzen und dabei Inhalte zu wiederholen und zu ergänzen. Ein bisschen Sorgen macht mir natürlich mein Zeitmanagement, vor allem weil ich den Aufwand der Vorlesungen noch nicht so ganz einschätzen kann.

Zur Struktur vom Lernportfolio: Ich habe mir jede Woche Notizen gemacht und ein paar Sachen aufgeschrieben, die mich interessiert haben. Daher habe ich einfach das Datum der Vorlesung aufgeschrieben. Jetzt ergänze ich und überarbeite sprachlich. Neue Ergänzungen/Gedanken sortiere ich thematisch passend ein. Die Quizfragen habe ich auch immer in der Vorlesung selbst aufgeschrieben, deswegen werde ich die auch einfach den entsprechenden Themen anfügen.

Weitere Anmerkungen: Die Quizfragen haben wir in den Vorlesungen auch beantwortet und meistens kollaborativ in den *shared notes* beantwortet. Wenn ich die Antwort ausreichend und gut fand, habe ich die auch so in mein Lernportfolio übernommen und nicht nochmal extra ausformuliert. Ich hoffe das ist ok.

Vorlesung - Notizen

Nach der ersten Vorlesung bin ich wirklich gespannt auf das Modul. Ich hatte mir die Inhalte nicht wirklich vorher angeschaut, also war ich ein bisschen überrascht, dass es hier um das Semantic Web geht, aber angenehm überrascht. Ich freue mich drauf, das alles mal aus einer anderen Perspektive kennen zu lernen.

Meine bisherige Erfahrung mit dem Semantic Web kommt nur aus einer relativ technischen Ebene. In meinem Linked Data Modul im Bachelor habe ich SPARQL gelernt und wie man die Daten, die es da gibt sinnvoll nutzt. Ich hoffe hier auch noch ein bisschen was über die „organisatorische“ und auch eher theoretische Ebene zu lernen. Noch nicht ganz klar ist mir, was die Themen Media Lifecycle Management und Linked Data bzw. Semantic Web miteinander zu tun haben, aber das wird sich sicherlich noch klären.

Besonders freue ich mich auf die Exkursionen und Gastvorträge. Einen Bezug zur „echten“ Welt herzustellen finde ich immer sehr wichtig in Vorlesungen.

Quiz

Was ist die Prüfungsleistung für das Modul Media Lifecycle Management?

Ein Lernportfolio mit wöchentlichen Einträgen zur Reflexion der Vorlesungsthemen.

Worin besteht der Unterschied in kontinuierlichen und diskreten Medien?

Kontinuierliche Medien bezeichnen die, deren Informationsgehalt in ihrer zeitlichen Veränderung liegt (z.B. Video, Audio). Diskrete Medien haben zu jedem Zeitpunkt den gleichen Informationsgehalt (z.B. ein Bild)

Welche Exkursionen und Gastvorträge sind im Modul Media Lifecycle Management geplant?

Exkursionen: Deutsche Nationalbibliothek, Deutsches Musikarchiv

Gastvorträge: Deutsche Digitale Bibliothek, Sächsisches Staatsarchiv, eccenca GmbH

Diese Woche bei der Gruppe Film:

Im Seminar haben wir uns erst mit den allgemeinen Eigenschaften von Medien auseinandergesetzt. Dabei sind sehr interessante Gespräche entstanden, die auch gezeigt haben, was das für ein großes Feld ist. Da wir als Seminargruppe ja ziemlich bunt gemischt sind, also alle möglichen verschiedenen Bachelorabschlüsse und Hintergründe haben glaube ich, dass die Gruppenarbeiten definitiv spannend werden.

Ich bin am Ende für das restliche Semester in der Gruppe Film gelandet. Darüber weiß ich bisher fast nichts, also bin ich gespannt was das Semester so bringt.

Glossareinträge der Woche:

Medien.

Media Lifecycle Management.

Woche 2 - Web Of Data

20.10.2021

Vorlesung - Notizen

Besonders interessant fand ich heute den historischen Abriss zum Thema Bibliotheken. Der letzte Kurs, den ich zum Thema Semantic Web/Linked Data gehört habe, fing damals direkt mit Internet an. Die Idee, dass das Internet Informationen mehr Menschen zur Verfügung stellt als früher (als die Bibliotheken nur den Gelehrten zur Verfügung standen), finde ich sehr interessant, aber ich bin nicht sicher, ob das so ganz der Wirklichkeit entspricht. Es sind halt einfach andere Informationen, die an anderen Stellen verfügbar sind. Also ich kann zwar in Zeiten des Internets auch als nicht-Akademiker*in auf Researchgate gehen, aber Zugang bekomme ich da trotzdem nicht zu allem. Also ich glaube zu einem gewissen Teil ist es immer noch so, dass bestimmte Informationen den „Gelehrten“ vorbehalten bleiben. Wie viel das Semantic Web da ändern kann weiß ich auch nicht.

The long tail. In der Vorlesung wurde das Konzept des long tail von Chris Anderson auf die Informationsdomäne angewendet. Da ich davon noch nie etwas gehört habe, habe ich es mal nachgeschlagen:

Die Idee dahinter ist, dass es bei vielen Produkten (v.a. digitalen Medien) ein paar bekannte und beliebte gibt, aber auch einen *long tail* der viel größer aber nischiger ist. Chris Anderson prognostiziert in seinem Artikel und Buch, dass Firmen mehr Erfolg haben, wenn sie diesen long tail erschließen, z.B. Netflix (haben mehr Erfolg, wenn sie auch kleinere Produktionen zur Verfügung stellen, weil dann mehr Leute den Service nutzen wollen, als wenn es nur die drei großen Blockbuster gibt).

Dieses Konzept kann man auch auf Informationen im Netz übertragen: es gibt ein paar große Seiten auf denen viel los ist, aber das eigentlich wertvolle und interessante ist im long tail, aber da muss man halt rankommen - Web 3.0 soll das also möglich machen.

Quiz

Was ist das Semantic Web?

Das Semantic Web oder Web 3.0 ist eine Erweiterung des World Wide Web nach den Standards des W3C, mit dem Ziel die Daten im Web zu formalisieren und maschinenlesbar zu machen.

Wie unterscheiden sich Web 1.0, 2.0 und 3.0?

Während im Web 1.0 größtenteils statische Inhalte zu finden sind, schaffen im Web2.0 Nutzer*innen selbst Inhalte, indem große Seiten es leichter machen, Daten ins Netz zu stellen und gemeinsam zu nutzen. Das Web 3.0 ist ein Konzept, das das Wissen im Netz maschinenlesbar machen und verlinken soll.

Welchen Nutzen hat das Semantic Web? (bzw. hätte ein umfassendes)

Die Informationen im Semantic Web sind strukturiert und maschinenlesbar. Dadurch kann man komplexe Anfragen stellen und die Informationen effizienter und automatisiert nutzen.

Warum braucht man noch ein Web?

Die Informationen sind ja schon im Netz, wir können sie nur nicht gezielt nutzen, weil sie nach menschlicher und nicht Maschinen-Logik aufgebaut sind.

Diese Woche bei der Gruppe Film:

Die Gruppenarbeit war diese Woche ein bisschen chaotisch, weil wir alle noch nicht so genau wissen wie es alles läuft. Das Thema der Woche war Suche und Recherche. Es kristallisiert sich ein bisschen heraus, dass wir einen Schwerpunkt legen müssen (vermutlich auf Filmproduktionen à la Hollywood, weil jede Art von Video einzubringen vielleicht einfach ein bisschen viel wäre.

Glossareinträge der Woche:

Semantic Web.

Linked Open Data.

Woche 3 - Resource Description Framework

03.11.2021

Vorlesung - Notizen

Diese Woche ging es um das Resource Description Framework. Vieles davon war mir schon ein Begriff, ich habe mich trotzdem über die Auffrischung gefreut. Eine Frage die in der Vorlesung aufkam: Wieso schreibt da eigentlich niemand Müll rein? Das habe ich mich auch auf jeden Fall schonmal gefragt. Aber ich glaube, dazu gehört auch zu verstehen, dass das Semantic Web halt eben nicht wie Wikipedia eine Website ist, sondern halt einfach das Internet, aber mit bestimmten Eigenschaften und Standards. Klar kann man da Müll reinschreiben, aber das liest ja dann niemand. Mechanismen, die mir helfen die Vertrauenswürdigkeit von bestimmten Informationen und Quellen einzuschätzen gibt es ja schon.

Trotzdem ist das Thema Falschinformationen im Netz relevant. Ich denke, dass die Schwelle etwas ins Semantic Web „reinzuschreiben“ schon relativ hoch ist, eben dadurch, dass die Daten im richtigen Format und auffindbar sein müssen. Es ist ja kein facebook-Post. Das so zu schreiben kommt mir jetzt auch wieder elitär vor. Darüber muss ich mir noch ein bisschen Gedanken machen. Welches Wissen ist relevant und wie WEIRD (Western Educated Industrialized Rich and Democratic) ist das Wissen im Semantic Web? Also sicherlich sehr, aber wie kann man das ändern?

Quiz

Was ist eine Ontologie?

Eine Ontologie definiert ein Metamodell für die Modellierung mit Klassen und Objekten

Wie verhindert man, dass Ungenauigkeiten und Inkonsistenzen entstehen?

Indem man eine Ontologie einsetzt und für die Daten ein bestimmtes Format vorschreibt.

Was ist RDF?

RDF steht für Resource Description Framework. In RDF werden Informationen als Tripel aus Objekt, Prädikat und Subjekt dargestellt.

Diese Woche bei der Gruppe Film:

Diese Woche haben wir gemeinsam die Aufgaben zu Suche, Recherche und Publikation bearbeitet. Es war wirklich schwer sich

Glossareinträge der Woche:

RDF.

Semantic Web Stack.

Woche 4 - Gastvortrag und praktische Übung

10.11.2021

Vortrag Dr. Kluttig

Ich fand schade, dass der Vortrag online war, ich wäre gerne mal ins Sächsische Staatsarchiv gegangen, aber das werde ich bei Gelegenheit mal nachholen.

Ein paar neue Fakten, die ich über Archive und das Archivieren gelernt habe, die mich überrascht haben/ich interessant fand:

- Archivierung ist Anfang UND Ende des Life Cycle - fast poetisch
- Archivieren ist vor allem vernichten, nur 1-2% wird tatsächlich archiviert
- Eine der schwierigsten Aufgaben ist, zu beurteilen was relevant ist.
- Das ist vor allem eine Kostenfrage, denn wenn etwas einmal archiviert ist, bleibt es für immer
- Kriterien: bestimmte Zeiträume (z.B. Wende), zeitgeschichtlich relevante Sachen, Federführung (Bsp. Gesetze/Polizeigesetz)
- Dokumente kommen in Quarantäne wegen Papierfischchen

Zwei Themen, die wir am ende noch diskutiert haben, waren Datenschutz und Falschinformationen:

Thema Datenschutz

Leipziger Meldekartei vom Polizeipräsidium, 1Mio Leute 600 Anfragen im Jahr
Ist an sich digital, aber man kann das nicht einfach offen machen, wegen personenbezogene Daten

Hier spielt die Abwägung Forschung/Privatsphäre eine Rolle

Was ist mit falschen Informationen?

Quellenkritik ist eine wichtige Frage, jede Person hat auch das Recht Sachen richtig zu stellen. Aber da müssen auch die Nutzer/innen, Forscher/innen gut arbeiten.

Quiz

Wie ist der Aufbau von URIs?

URIs bestehen aus fünf Teilen: scheme (Schema oder Protokoll), authority (Anbieter oder Server), path (Pfad), query (Abfrage) und fragment (Teil), wovon scheme und path Pflicht-Teile und der Rest optionale Teile.

Das häufigste Schema ist Http

Was ist die Struktur eines Tripels im RDF?

Triples haben die Struktur: Subjekt, Prädikat, Objekt

Wie kann mithilfe von RDF eine Ressource beschrieben werden und welche Struktur besitzt diese Beschreibung?

Ressource wird mithilfe eines bzw. mehreren Triples beschrieben Struktur: Graph

Welche Rolle spielen Ontologien im Web of Data?

Ontologien bieten eine formale Beschreibungssprache und Framework um Begriffe für Maschinen bzw. Software interpretierbar zu machen. Ontologien machen auch die Strukturierung der Daten im Web of Data konsistenter und einfacher zu verstehen.

Seminar

Im Seminar haben wir eine erste praktische Übung gemacht, das war für mich auf jeden Fall eine gute Wiederholung. Außerdem war gut, dass die Beispiele in jupyter notebooks und in python waren, das sind beides vertraute Umgebungen für mich, aber die Verbindung mit den linked data technologien war noch neu für mich, das war super. In meiner Bachelorvorlesung zu linked data haben wir immer nur einzelne ttl Dateien erstellt und Anfragen geschrieben, aber diese Beispiele schienen mir schon etwas praxisnäher.

Glossareinträge der Woche:

Archiv.
URI

Woche 5 - DBPedia

26.11.2021

Vorlesung - Notizen

DBPedia ist wirklich ein interessantes Projekt, das ich schon aus dem Bachelor kannte. Deswegen war für mich nicht viel Neues dabei. Ich habe deshalb mal ein paar SPARQL Anfragen über den DBPedia SPARQL-Endpoint geschrieben, ein bisschen was ausprobiert und dabei SPARQL mal wiederholt.

Unten sieht man die queries, die linke ruft die dbpedia informationen zum Radfahrer Günther Hoffmann ab, die rechte findet alle Radfahrer, die 1968 bei den olympischen Spielen gefahren sind und noch leben.

The image displays two side-by-side screenshots of the DBPedia SPARQL Query Editor. Both screenshots show the browser address bar with the URL 'https://dbpedia.org/sparql/'.

The left screenshot shows the 'SPARQL Query Editor' interface. The 'Default Data Set Name (Graph IRI)' is set to 'http://dbpedia.org'. The 'Query Text' area contains the following SPARQL query:

```
SELECT *
WHERE
{
  ?athlete rdfs:label "Günter Hoffmann (cyclist)"@en
}
```

The 'Results Format' is set to 'HTML'. There are 'Execute Query' and 'Reset' buttons at the bottom.

The right screenshot shows the same interface. The 'Default Data Set Name (Graph IRI)' is set to 'http://dbpedia.org'. The 'Query Text' area contains the following SPARQL query:

```
SELECT
?athlete
WHERE
{
  ?athlete dct:subject dbc:Living_people ;
  dct:subject dbc:Cyclists_at_the_1968_Summer_Olympics.
}
```

The 'Results Format' is set to 'HTML'. There are 'Execute Query' and 'Reset' buttons at the bottom.

Ich hab dann noch ein bisschen rumexperimentiert, aber hatte leider nicht so viel Zeit.

Quiz

Was ist DBpedia und warum nutzt man DBpedia?

DBpedia transformiert Wikipedia-Artikel gemäß der Konzepte von Linked Open Data für das Semantic Web. Man nutzt es um auf die Informationen aus Wikipedia automatisiert zuzugreifen.

Nennen Sie eine wesentliche Gemeinsamkeit und einen Unterschied von DBpedia und Wikidata.

Gemeinsamkeiten:

Versuchen Informationen aus Wikipedia strukturiert zu speichern

Abfrage per SPARQL

Unterschiede:

Aufbau der strukturierten Informationen (RDF vs. JSON, XML, SQL)

Repräsentation von Fakten (Triple vs. Statements und Claim)

Wieso sind mit DBpedia und SPARQL verschachtelte Abfragen möglich?

Weil die Wikipedia-Daten als Triple aufbereitet werden und somit untereinander verlinkt sind (Durch die Verbindung der gleichen Variable). Auch die Verknüpfung durch Verkettung von Triple Pappern und das Wiederverwenden von Variablen (Joins), Vereinigung durch UNION (Vereinigungsmenge) und OPTIONAL ist möglich.

Diese Woche bei der Gruppe Film:

Im Seminar ging es um Mediendatenbanken, Metadaten und Medienformate. Das war besonders interessant, da video ein ziemlich komplexes Medium ist, alleine schon dadurch dass es eine Audio und eine Videospur gibt. Die Metadaten sind auch nicht ganz trivial, da viele schwer zu formalisieren sind.

Glossareinträge der Woche:

SPARQL.

DBPedia.

Woche 6 - OpenRefine

01.12.2021

In dieser Woche haben wir OpenRefine ausprobiert. Das ist ein Tool um Daten zu bereinigen und zu bearbeiten. Mit einem entsprechenden Plugin funktioniert auch die rdf-serialisierung.

Ich fand es ein bisschen schade, dass ich alleine in der Gruppe Linux war. Aber es war ganz interessant, das auszuprobieren. In meiner Bachelorvorlesung zu linked data mussten wir manuell rdf dateien schreiben und dann daraus queries ausführen, das war zwar gut fürs Verständnis, aber es ist nicht besonders anwendungsbezogen, weil man das ja schlecht für einen großen Datensatz machen kann. Dafür sind dann tools wie openrefine da.

The screenshot shows the OpenRefine web interface. At the top, there's a navigation bar with the OpenRefine logo and a 'New Version' notification for v3.5.2. Below that, a 'Create Project' dialog is open, showing a table of data. The table has columns for episode number, season, title, air date, and other metadata. Below the table, the 'Parse data as' section is visible, showing options for character encoding (US-ASCII) and parsing settings like 'Parse next' and 'Store blank rows'.

Project name	Tags
simpsons episodes csv	

Episode	Season	Title	Air Date	Runtime	Score	View Count	Image					
4	17	Two Cars in Every Garage and Three Eyes on Every Fish	1990-11-01	7F01	2	4	17	26.1	64959	8.1	1457	http://static-media.fxx.com/img/FX_Networks_..._FXX/660/859/Simpsons_02_01.jpg
5	19	Dead Putting Society	1990-11-15	7F08	2	6	19	25.4	50691	8	1366	http://static-media.fxx.com/img/FX_Networks_..._FXX/662/811/Simpsons_02_08.jpg
6	21	Bart the Daredevil	1990-12-06	7F06	2	8	21	26.2	57605	8.4	1522	http://static-media.fxx.com/img/FX_Networks_..._FXX/662/811/Simpsons_02_06.jpg
7	23	Bart Gets Hit by a Car	1991-01-10	7F10	2	10	23	24.8	56486	7.8	1340	http://static-media.fxx.com/img/FX_Networks_..._FXX/673/547/bart_gets_hit_by_car.jpg
8	26	Homer vs. Lisa and the 8th	1991-02-07	7F13	2	13	26	26.2	58277	8	1329	http://static-media.fxx.com/img/FX_Networks_..._FXX/930/879/Simpsons_02_13.jpg

Quiz

Wozu dient OpenRefine?

Ein Programm, womit große Datenmengen durchsucht, bereinigt, in andere Formate transformiert und mit anderen Datenmengen verbunden werden können. Datenformate zb: CSV, XLS, JSON oder XML

Welche Schritte muss ich bei der Extraktion von RDF Daten aus einer CSV Datei beachten?

- Normalisierung der Daten (z.B. Whitespace entfernen, Datumsformate anpassen)
- Zuweisung einer Spalte zu einem Schema

Warum eignet sich OpenRefine für die Bearbeitung von Daten?

Openrefine ist freie Software, das heißt für alle kostenlos und verfügbar.

Diese Woche bei der Gruppe Film

Im Seminar haben wir uns mit der manuellen Bearbeitung und passenden Software für Videos auseinandergesetzt. Ich persönlich war ein bisschen erschlagen von der schiereren Fülle von Tools und Möglichkeiten. Aber in der Gruppe war es ein bisschen einfacher sich einen Überblick zu verschaffen.

Glossareinträge der Woche:

OpenRefine
Metadaten

Woche 7 - Gastvortrag

08.12.2021

Vortrag Frau Berta von der DDB

Diese Woche hat Frau Berta von der Deutschen Digitalen Bibliothek einen Vortrag gehalten. Der war auch interessant, allerdings schwer verständlich für mich, da ich Verbindungs- und Audioprobleme hatte.

Die DDB ist wie der Name schon sagt eine digitale Bibliothek, die den Nutzer*innen Zugriff auf verschiedene digitalisierte Medien wie Musik, Bücher und Filme, aber auch Kunstwerke, Fotografien und Akten ermöglicht. Sie agiert als Portal, Aggregator und Netzwerk. So wird ein Austausch für verschiedene Arbeitsgruppen und Institutionen möglich, mit dem Ziel einen demokratisierten Zugang zu Wissen und Ressourcen zu ermöglichen und eine digitale Plattform für das kulturelle Erbe Deutschlands zu erschaffen.

Mein Fazit: Ich fand es wirklich schade, dass ich nicht ganz alles mitbekommen habe. Frau Berta wirkte sehr engagiert und wenn ich eins mitgenommen habe, dann auf jeden Fall gewachsenen Respekt für die Leute, die sich unermüdlich mit der Konservierung, Aufbereitung, Bereitstellung und auf eine Weise Demokratisierung von Wissen beschäftigen. Im Vortrag ging es ja auch viel um Datenbereinigung und die Herausforderungen die es da gibt.

Quiz

Was ist die Europeana und in welcher Verbindung steht die DDB zu Europeana?

Europeana ist eine virtuelle Bibliothek, die das wissenschaftliche und kulturelle Erbe Europas in Form von multimedialen Daten zugänglich machen soll.

Die DDB ist Aggregator für Europeana:

- „große Schwester“ von DDB
- deutscher Aggregator - übermittelt Metadaten und Objekte aus deutschen Kultureinrichtungen an Europeana

Was bedeutet Datenqualität und was muss dabei beachtet werden?

Maß für Erfüllung der Anforderungen an eine modellierte Information(sstruktur).
Qualität von Daten abhängig von: Korrektheit, Konsistenz, Zuverlässigkeit,
Vollständigkeit, Genauigkeit, Aktualität, Redundanzfreiheit, Relevanz, Einheitlichkeit,
Eindeutigkeit, Verständlichkeit

Diese Woche bei der Gruppe Film:

Diese Woche haben wir einen Halbzeitüberblick gemacht.
Wir haben dazu auch die Struktur unserer bisherigen Ergebnisse überarbeitet. Das hat mir sehr geholfen, weil ich die Übungen bisher immer sehr chaotisch und unübersichtlich fand.

Glossareinträge der Woche:

DDB

Europeana

Woche 8 - Ontologies as Common Language

15.12.2021

Vorlesung - Notizen

Diese Woche haben wir Ontologien und RDF-Schema besprochen.

Ich kannte viele Begriffe schon vorher. Allerdings fand ich trotzdem die Wiederholung des rdf-schema nützlich, da ich viele Konzepte nicht mehr auf dem Schirm hatte.

Am Ende der Vorlesung haben wir dazu eine praktische Übung gemacht. Ich fand es schade, dass wir dafür nicht so viel Zeit hatten. Dazu hätte ich mir eine längere und vielleicht etwas umfangreichere Übung gewünscht. Gute Modelle und Ontologien zu schreiben ist eine schwierige Aufgabe und meiner Meinung nach kann man das nicht genug üben, weil es auch in anderen Bereichen extrem hilfreich ist.

Ich habe mich daran erinnert, dass ich mal angefangen habe eine Einführung in Ontology Engineering (Dr. C. Maria Keet - Introduction to Ontology Engineering) zu lesen, die ich mir von meiner Nachbarin ausgeliehen hatte. Habe jetzt nochmal gesucht und festgestellt, dass es open access ist. Ich nehme mir vor das mal zuende zu lesen.

Quiz

Was ist eine Ontologie (in der Informatik)?

Ontologie: Lehre von Sein

Ontologien in der Informatik sind formal geordnete Darstellungen für eine Menge von Begriffen und ihre Beziehungen. Dadurch lässt sich Wissen formalisieren und klassifizieren.

Was sind wichtigen Merkmale um den Begriff Ontologie einordnen zu können?

- definiert und beschreibt Ausdrücke
- modelliert eine Domäne

- basiert auf einem Konsens
- maschinenlesbar
- hohe Aussagekraft bei steigender Komplexität

Was sind RDF-S, OWL und SHACL und wozu kann es jeweils verwendet werden? Was gibt es für Einschränkungen?

RDF-S = RDF-Schema

OWL = Web Ontology Language

SHACL = Shapes Constraint Language

RDF Schema ist ein Teil von RDF, der in RDF beschrieben wird. Es handelt sich um einen objektorientierten Ansatz

OWL basiert auf RDF Schema und fügt weitere Beschreibungsmöglichkeiten hinzu (Transitivität, Kardinalität)

SHACL ist eine Sprache zur Validierung von RDF-Graphen anhand spezifizierter Bedingungen. Es ist ein Shapemodell mit Constraints. Die Bedingungen/Constraints werden per RDF angegeben.

Diese Woche bei der Gruppe Film:

Diese Woche haben wir uns im Seminar mit Ähnlichkeiten und Beziehungen zwischen medialen Objekten beschäftigt. Interessant fand ich, dass es sozusagen technische Ähnlichkeiten gibt, also zum Beispiel ein gleiches Dateiformat, aber auch semantische Ähnlichkeiten, die schwerer zu formalisieren sind. Also wenn ein Film zum Beispiel einen anderen adaptiert oder parodiert. Noch schwerer wird es bei subtilen Referenzen.

Glossareinträge der Woche:

Ontologie

RDF-Schema

Woche 9 - Linked Open Data

05.01.2021

Vorlesung - Notizen

Diese Woche ging es unter anderem um das von W3C herausgegebene 5 Star Model für Open Linked Data. Die wollte ich nochmal kurz zusammenfassen:

- 1. Star:** Daten sind öffentlich zugänglich und unter einer öffentlichen Lizenz. (Bsp. eine pdf-Datei)
- 2. Star:** Daten sind öffentlich und strukturiert. Allerdings sind die Daten nur mit proprietärer Software nutzbar. (Bsp. eine Excel-Datei)
- 3. Star:** Daten sind öffentlich, maschinenlesbar und in einem nicht-proprietären Format. (Bsp. eine csv-Datei)
- 4. Star:** Daten sind öffentlich, maschinenlesbar, in einem nicht-proprietären Format und es werden URIs verwendet um Dinge zu identifizieren.

5. Star: 4 Star + Die Daten sind mit anderen Datensätzen verlinkt. Dadurch tritt dann der Netzwerk Effekt ein.

Quiz

Warum sind nicht-proprietäre Datenformate wichtig?

Nicht-proprietäre Formate sind freie, bzw. offene Formate wie zB. CSV und HTML. Um diese Dateien zu verarbeiten und zu nutzen sind keine kostenpflichtigen Lizenzen nötig. Es ist wichtig nicht-proprietäre Formate zu verwenden, damit der Zugriff auf bestimmte Daten nicht vom Geld abhängt.

Wieso benötigen wir strukturierte Datenformate im Web?

- Maschinenlesbarkeit
- Automatisierte Verarbeitung
- Standardisierung der der Auswertung
- Darstellung von komplexen Daten

Maschinenlesbarkeit **Wie können wir dafür sorgen, dass unsere Daten verlinkbar sind?**

Durch die Verwendung von URIs und die sinnvolle Strukturierung von Daten

Warum sollten wir auf andere Daten verlinken?

- Kontext herstellen
- Netzwerkeffekt
- verwandtes Wissen finden (Aggregation von Wissen)

Diese Woche bei der Gruppe Film:

Das Thema der Woche war Klassifikation und Anreicherung. Wir haben wieder festgestellt, dass das gar kein triviales Thema ist. Es gibt wirklich viele Arten Videos zu klassifizieren (und wir haben uns ja nur auf Film beschränkt).

Das zweite Thema Metadaten und Zusatzinformationen war auch spannend und gleichermaßen kompliziert. Einige Metadaten sind sehr deutlich (Regisseur, Altersfreigabe etc). Aber vor allem in Zeiten des Streaming hat sich die Notwendigkeit einiger anderer herausgestellt. Besonders um die Auffindbarkeit in Datenbanken zu erhöhen sind gepflegte Metadaten wichtig.

Außerdem muss man unterscheiden zwischen Metadaten die sich auf eine Datei im ganzen beziehen (z.B. Altersfreigabe) und Metadaten die mit timestamps versehen sind (z.B. Kapitelmarken).

Spannend fanden wir da auch die Möglichkeiten solche komplexeren Metadaten automatisch zu extrahieren (z.B. Szenenübergänge oder welche/r Charakter/Schauspieler*in in dieser Szene anwesend ist.)

Glossareinträge der Woche:

5-Star Model

nicht-proprietäre Datenformate

Woche 10 - Klassifikation/Anreicherung, Datenqualität, Weiterentwicklung/Reparatur/Collaboration

12.01.2021

FAIR Data

Sollen ein nachhaltiges Forschungsdatenmanagement sicherstellen

F – Findbarkeit (Daten erhalten global einzigartige Identifier und Metadaten)

A – Open Access (Daten können durch offenes, freies Protokoll benutzt werden)

I – Interoperabilität (Daten nutzen eine geteilte Sprache/Grundlage für die Repräsentation ihrer Daten)

R – Reuse/Nachnutzbarkeit (Daten und Metadaten müssen ausreichend beschrieben werden für Wiederverwendung)

Datenqualität ist echt wichtig, habe ich auch schonmal bemerkt, da kann halt ein ganzes Projekt dran scheitern.

Spannend fand ich auch den Abschnitt zum Thema semantische Wikis, gerade das Ontowiki fand ich interessant, damit will ich mich gerne noch mehr auseinandersetzen.

Diese Woche bei der Gruppe Film:

Leider konnte ich diese Woche im Seminar nicht dabei sein, aber zum Glück wird ja alles im Wiki dokumentiert und meine Gruppe war so nett mir das wichtigste mitzuteilen.

Die Themen waren Qualität, Kollaboration und Publikation. Alles Themen zu denen man jeweils eigentlich mehrere Stunden füllen könnte, daher ist das Thema Publikation auch ein bisschen hinten runtergefallen.

Interessant fand ich wieder, vor allem beim Thema Qualität die Unterscheidung zwischen den technischen Aspekten (Bildqualität) und den eher schwierig zu messenden Aspekten (ist es ein guter Film? Kann man das messen? Welche Metrik funktioniert?).

Glossareinträge der Woche:

FAIR Data Prinzipien

Wiki

Woche 11 - Gastvortrag

19.01.2021

Vortrag Sebastian Tramp von eccenca

Der Vortrag diese Woche hat mir nicht so gut gefallen, wie die Anderen. Herr Tramp hat sehr enthusiastisch von seiner Arbeit gesprochen, aber inhaltlich fiel es mir manchmal schwer zu folgen, da ich die Struktur des Vortrags nicht so einfach nachzuvollziehen fand. Am Ende hat es aber doch wieder gut zusammengepasst. Einige Dinge im Vortrag waren neu für mich. Insbesondere die betriebswirtschaftliche Sicht auf Wissen fand ich interessant. Aus dem akademischem Umfeld haben wir da sicherlich manchmal andere Perspektiven. Aber aus meinem Umfeld kenne ich schon anekdotische Hinweise darauf, dass in vielen Unternehmen das vorhandene Wissen nicht effizient genutzt wird. (Das weiß nur die Ute wie das geht, aber die ist grade im Urlaub. Und wenn Ute in Rente geht, weiß es niemand mehr) Vor allem die Balance zu finden zwischen sinnvoller digitaler Transformation und Benutzbarkeit für die Mitarbeiter*innen ist sicherlich eine Herausforderung und interessante Aufgabe. Außerdem wurde ausführlich der Weg zu Fair Data erklärt. An dieser Stelle wurden sehr viele verschiedene Modelle gezeigt. In einem ersten Schritt hat Herr Tramp die Schritte Kuratierung (Knowledge Graph erzeugen), Exploration (Suche und Erkundung) und Provisionierung (Bereitstellung) unterschieden. Danach hat er noch andere Modelle zum Workflow beschrieben, unter anderem den Enterprise Knowledge Graph Project Flow, ein iteratives Modell, das auch unserem Linked Data Lifecycle ähnelt. Besonders interessant fand ich den letzten Abschnitt zum Thema Ontology Engineering. Mit dem Thema Ontologien habe ich mich ja schon im Bachelor an vielen Stellen auseinandergesetzt, vor allem da Logik und Semantik auch meine Schwerpunkte waren. Allerdings kam mir alles immer sehr theoretisch vor, das hat mir zwar Spaß gemacht, aber wo man das jetzt tatsächlich anwendet war immer so ein bisschen eine Frage. Zum Beispiel hat Herr Tramp erwähnt, dass Reasoning zum Beispiel in der Forschung eine große Rolle spielt, aber in der Praxis nicht, weil es einfach nicht skaliert. Hätten wir mehr Zeit gehabt, hätte ich gerne noch nachgefragt was es so für Themengebiete gibt, die so unterschiedlich in der Forschung und Praxis sind.

Quiz

Aus welchen Schritten besteht der iterative Enterprise Knowledge Graph Project Flow bei eccenca?

- Analyse der Daten und Verstehen der Domäne
- Erstellen des semantischen Modells (Ontologien, Vokabulare)
- Mapping der Daten auf die Ontologie
- Enrichment, Cleaning and Interlinking
- Review

Diese Woche bei der Gruppe Film:

In unserer Seminargruppe haben wir an der Zusammenfassung der Arbeitsaufträge aus den vergangenen Wochen weitergearbeitet.

Woche 12 - Zusammenfassung

26.01.2021

Gedanken

Zusammenfassend muss ich leider sagen, dass ich nicht so viel Neues gelernt habe, wie ich mir zu Beginn des Semesters erhofft hatte. Vor allem im Bereich Linked Data hätte ich mir praktische Übungen gewünscht, die vielleicht auch ein bisschen umfangreicher sind. Eine Idee wäre zum Beispiel das Wiki in den Übungen von Anfang an als Semantisches Wiki aufzubauen. Oft war für die praktischen Übungen nur wenig Zeit. Insgesamt hatte ich das Gefühl die Vorlesung und die Übung bauen nicht immer gut aufeinander auf. Es war ein bisschen als gäbe es eine Vorlesung über das Semantic Web und eine über Medienarten und deren Lifecycle. Vor allem dadurch, dass wir die in der VL betrachteten Konzepte in den Übungen kaum angewendet haben, war es schwierig die Konzepte dann mental zusammenzubringen.

Gut war aber auf jeden Fall, mal ein bisschen über meinen Tellerrand hinauszuschauen und mich ein bisschen mit den Medien(arten) auseinanderzusetzen. Besonders die Gruppenarbeiten im Seminar haben mich immer wieder ein bisschen aus meiner „comfort zone“ herausgeholt, da ich mich noch nicht so viel mit dem Medium Video beschäftigt hatte.

Meine Erfahrung mit dem Lernportfolio ist ein bisschen durchwachsen. Natürlich hilft es, seine Gedanken aufzuschreiben und sich selbst nochmal mit den Themen der Vorlesung auseinanderzusetzen. Aber dadurch, dass es eben auch eine Prüfungsleistung ist, steht man trotzdem unter einem gewissen Druck.

Auf der anderen Seite hätte ich vermutlich nicht das Durchhaltevermögen gehabt, wenn es keine Prüfungsleistung wäre ein komplettes Portfolio zu erstellen. Und das ist auch nicht zu vernachlässigen. Es hat mich wirklich dazu motiviert mich jede Woche zusätzlich zur Vorlesung mal hinzusetzen und Inhalte zu wiederholen (und das will was heißen bei mir!)

Insgesamt kommt es mir auch ein bisschen wie eine Fleißübung vor, da sich einige Dinge immer wiederholen, wie zum Beispiel die wichtigen Begriffe aus Vorlesungen, Quizfragen und Glossareinträge. Besonders die fest vorgegebene Form, die für jede Vorlesung eingehalten werden musste, fand ich eher beschränkend als hilfreich und generell wenig abwechslungsreich. Also es gab oft Wochen in denen ich mich gerne noch mit etwas zusätzlich beschäftigt hätte und das dann auch ins Lernportfolio aufgenommen, aber durch die Quizfragen und Definitionen war meine Zeit und der Platz dann schon relativ ausgeschöpft. Außerdem habe ich natürlich die Abgabe und die Note im Hinterkopf, das heißt der Pflichtteil muss gemacht werden, aber danach hatte ich keine Lust mehr auf Kür. Ich fand es schwierig eine Balance zu finden zwischen den Anforderungen und meinen eigenen Interessen.

Besser hätte ich gefunden, wenn wir jede Woche eine kleine spezifische und anwendungsorientierte Aufgabe bekommen hätten, bei der man aktiv das Gelernte anwenden muss (und dabei ja auch die Vorlesung rekapitulieren und seinen eigenen Lernprozess reflektieren muss).

Fazit Gruppenarbeit Film

Von der Gruppenarbeit bin ich im Endeffekt nicht so ganz begeistert. Ich fand zwar die Arbeitsaufträge ganz interessant, aber da wir in unserer Gruppe alle sehr wenig Vorwissen hatten, waren wir manchmal ein bisschen verloren. Oft haben wir einzeln recherchiert und uns anschließend über unsere Ergebnisse unterhalten. Ich habe den Eindruck, dass der Lerneffekt davon nicht so enorm groß war.

Auch fand ich, dass man die Zeit im Seminar hätte besser nutzen können, z.B. für praktische Übungen zur Anwendung von Methoden/Verfahren aus der Vorlesung. Oft war es schwer, einen Zusammenhang herzustellen zwischen den Inhalten aus der Vorlesung und den im Seminar besprochenen Themen.

Trotzdem hat die Gruppenarbeit im Seminar meinen Horizont erweitert und ich glaube auch, dass ich auf viele Sachen eine neue Perspektive gewinnen konnte, vor allem auf das Medium Video und die komplexen Arbeitsschritte die mit dem Lebenszyklus verbunden sind.

Glossar

Medien. Der Begriff Medium leitet sich vom lateinischen Wort ‚medium‘ ab, was ‚Mitte‘ bedeutet. Medien sind Mittel zum Transport von Informationen von einem Sender zu einem Empfänger.

Media Lifecycle Management. Alle Arbeitsschritte die mit dem Lebenszyklus eines Mediums von Erstellung bis Speicherung zu tun haben. Die Schritte unterscheiden sich je nach Medienart.

Semantic Web. Das Semantic Web oder Web 3.0 ist eine Erweiterung des World Wide Web nach den Standards des W3C, mit dem Ziel die Daten im Web zu formalisieren und maschinenlesbar zu machen.

Linked Open data. Linked Data beschreibt strukturierte Daten, die maschinenlesbar und verlinkt sind, d.h. auf denen man strukturierte Anfragen ausführen kann. Sind diese Daten Open Access, dann spricht man von Linked Open Data.

RDF. RDF steht für Resource Description Framework. In RDF werden Informationen als Tripel aus Objekt, Prädikat und Subjekt dargestellt

Semantic Web Stack. Der Semantic Web Stack beschreibt die Architektur des Semantic web und welche Technologien verwendet werden. Er baut auf RDF auf.

Archiv. Ein Archiv ist eine Institution oder Organisationseinheit, in der Archivgut zeitlich unbegrenzt im Rahmen der Zuständigkeit des Archivs oder des jeweiligen Sammlungsschwerpunktes aufbewahrt, benutzbar gemacht und erhalten wird. Das kann digital oder analog geschehen.

URI. Ein Uniform Resource Identifier (Abk. URI, englisch für einheitlicher Bezeichner für Ressourcen) ist ein Identifikator und besteht aus einer Zeichenfolge, die zur Identifizierung einer abstrakten oder physischen Ressource dient. Ressourcen wie Webseiten, sonstige Dateien, E-Mail-Empfänger oder webapps können identifiziert werden.

SPARQL. SPARQL ist eine graphenbasierte Abfragesprache für Abfragen von Inhalten aus dem Beschreibungssystem Resource Description Framework (RDF), das in Datenbanken zur Formulierung logischer Aussagen über beliebige Dinge genutzt wird.

DBpedia. DBpedia ist ein Gemeinschaftsprojekt, das Wikipedia-Artikel gemäß der Konzepte von Linked Open Data für das Semantic Web transformiert.

OpenRefine. OpenRefine ist ein Programm zur Bereinigung und Umwandlung von Daten. Die Software wurde im Jahr 2010 von Google veröffentlicht, aber ist inzwischen als Freie Software verfügbar.

Metadaten. Metadaten oder Metainformationen sind strukturierte Daten, die Informationen über Merkmale anderer Daten enthalten.

DDB. Die Deutsche Digitale Bibliothek ist eine virtuelle Bibliothek, die verschiedene deutsche Kultur- und Wissenschaftseinrichtungen vernetzen und über eine gemeinsame Plattform öffentlich zugänglich machen soll.

Europeana. Die Europeana ist eine europäische virtuelle Bibliothek, die das kulturelle Erbe Europas konservieren und zugänglich machen soll. Die DDB ist ihr deutscher Aggregator.

Ontologie. Ontologien in der Informatik sind formal geordnete Darstellungen für eine Menge von Begriffen und ihre Beziehungen. Dadurch lässt sich Wissen formalisieren und klassifizieren.

RDF-Schema. RDF Schema ist eine Erweiterung von RDF und stellt ein Vokabular für die Modellierung von Daten in RDF zur Verfügung.

5-Star Model. Das 5-star model für Linked Open Data wird verwendet um die Qualität von Daten im netz bezogne auf ihre nutzbarkeit zubeschreiben.

nicht-proprietäre Datenformate. Dateien in nicht-proprietären Dateiformaten kann man verwenden, bearbeiten und teilen ohne dafür eine kostenpflichtige Software zu benutzen.

FAIR Data. FAIR Data: Findbar, Open Access, interoperabilität, Reuse

Wiki. Ein Wiki ist eine Website, deren Inhalte von den Nutzern direkt im Webbrowser bearbeitet und geändert werden können. Wikis ermöglichen kollaboratives Arbeiten.

Mindmap

