HOCHSCHULE DER MEDIEN

Offene Forschungsdaten am Beispiel des Projekts „Japanese Visual Media Graph“

Teil 1: Übersicht über das Projekt

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Magnus Pfeffer
Open Up! Webinar-Reihe
Hochschule der Medien, 16. Nov. 2021

Aufbau des Vortrags

HOCHSCHULE
DER MEDIEN

- Projektidee
- Grundlagen Knowledge Graph
- Die Fan-Gemeinschaften
- Datenqualität und Datenintegration
- Rechtliche Fragen
- Datenbank und Frontend
- Die Methode “Tiny Use Case”

HOCHSCHULE
DER MEDIEN

Projektidee

Ansatz

HOCHSCHULE DER MEDIEN

- Beobachtung: Fan-Gemeinschaften sammeln seit teilweise 20 Jahren Informationen zu Anime, Manga und Computerspielen und es gibt keine vergleichbaren Daten aus anderen Quellen

- Japanese Visual Media Graph (JVMG) Projekt
- Ziele:
  - Aufbereiten der Daten für quantitative Forschung
  - Integration der Daten in ein einheitliches Datenmodell
  - Dauerhafter Zugang mit klaren Lizenzbedingungen

Kernelemente

HOCHSCHULE DER MEDIEN

- Aktive Zusammenarbeit mit den Communities
  - Respektieren der Wünsche und Lizenzen
  - Angebote auch für die Communities
- Forschende als primäre Zielgruppe
  - Datensammlung nicht als Selbstzweck
  - Konkrete Szenarien (use cases) steuern die Entwicklung der Angebote
- Transparenter Entwicklungsprozess
  - Dokumentation von Entwurfsentscheidungen
  - Bereitstellen als Open Source

HOCHSCHULE
DER MEDIEN

Grundlagen Knowledge Graph

Graph allgemein

HOCHSCHULE DER MEDIEN

- Abstrakte Struktur mit Knoten und Kanten (=Verbindungen)

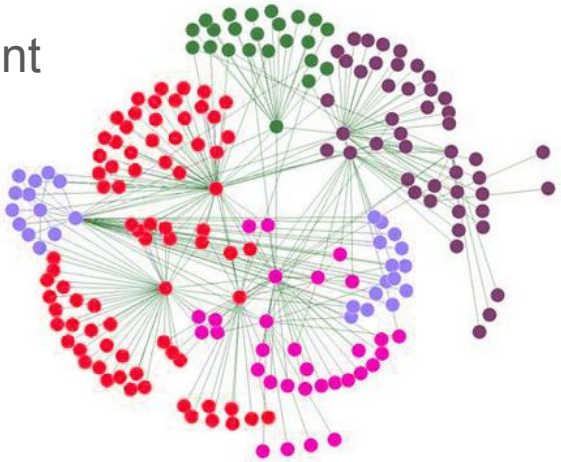
- Übertragung auf die Wissensrepräsentation
  - Knoten: Objekte oder Konzepte (= “Entitäten”)
  - Kanten: Beziehungen zwischen den Entitäten

- Eigenschaften
  - Einzelne Aussagen über “die Welt” als Tripel (Knoten, Kante, Knoten) darstellbar
  - Aussagen über die gleichen Entitäten können in einem Graph zusammengefasst werden
  - Die Summer der Aussagen im Graph ermöglicht neue Erkenntnisse

Beispiel: Wer kennt wen?

HOCHSCHULE DER MEDIEN

- Repräsentation sozialer Beziehungen
  - Entitäten: Personen
  - Beziehung: “ist bekannt mit”
- Es reicht aus, wenn jede Person ihre Bekannten benennt
  - Aus den Aussagen wird der Graph gebildet
- Analysemöglichkeiten (Beispiele)
  - Wessen Bekanntenkreise überlappen sich, obwohl sich die Betroffenen selbst nicht kennen?
  - Wer kann viele Personen über direkte und indirekte Beziehungen (hier: Bekannte von Bekannten) erreichen?



Knowledge Graph

HOCHSCHULE DER MEDIEN

- Erweiterung des einfachen Graphen
  - Beschreiben von Eigenschaften der Entitäten
    (d.i. Aussagen über die Entitäten, die keine Beziehung zu einer anderen Entität enthalten)
  - Neue Arten (“Klassen”) von Entitäten
  - Neue Arten von Beziehungen

- Am Beispiel “Wer kennt wen”
  - Eigenschaften: Name und Alter der Personen
  - Neue Entitäten: Wohnorte, Hobbies, Schule, Firma
  - Neue Beziehungen: “ist befreundet mit”, “arbeitet bei”, “wohnt in”, “interessiert sich für”

- Ermöglicht differenziertere Fragen und Analysen

HOCHSCHULE
DER MEDIEN

Die Fan-Gemeinschaften

animeclick.it

HOCHSCHULE DER MEDIEN

- **Breites** Interesse an japanischen visuellen Medien und Japan **allgemein**
  - Manga, Anime
  - Fernsehserien
  - Kultur, Essen und Trinken

- Kernangebote
  - Informationen zu den Medien und Themen
  - Möglichkeiten zum Austausch und Diskussion
  - Präsentation von eigenen Werken (Fan-Art, Cosplay)

- Partnerseite “gamerclick.it” zu Spielen

AnimeClick

The Visual Novel Database

HOCHSCHULE DER MEDIEN

- **Enges** Interesse an Computerspielen des **Genres** “Visual Novel”

- Kernangebot: Datenbank
  - Daten zu den Spielen und den unterschiedlichen Versionen
  - Daten zu Charakteren und ihren Rollen
  - Daten zu Firmen und beteiligten Personen
  - Inhaltliche Beschreibung über Schlagwörter und Zusammenfassungen

- Diskussionsforen

Anime Characters Database

HOCHSCHULE DER MEDIEN

- Interesse an einem **bestimmten Aspekt** der Domäne

- Kernangebot: Visuelle Suchmaschine
  - Aussehen und Kleidung der Charaktere

- Informationen zu Serien und Sprecher:innen
- Spielerische Interaktion mit der Website

Anime Characters Database

Kennzahlen

HOCHSCHULE DER MEDIEN

Visual Novel Database					
Visual Novel Works	Unique Releases	Characters	Descriptive Work Tags	Descriptive Character Tags	Agents
28,000	73,000	91,000	2600	2800	31,000
AnimeClick					
Anime	Manga	Characters	Related Works	Authors	Other Staff
9400	11,000	102,000	15,000	28,000	39,000
Anime Characters Database					
Characters	Works	Character Relations	Descriptive Character Tags	Descriptive Work Tags	Voice Actors
101,000	11,000	17,000	3800	1100	4900

HOCHSCHULE
DER MEDIEN

Datenqualität und Datenintegration

Datenqualität

HOCHSCHULE DER MEDIEN

- Stichproben des Datums “Titel” aus mehreren Quellen
- Überprüfung durch Vergleich mit Abbildungen der Medien oder Angaben auf Webseiten der Verlage/TV-Stationen/Produzenten

- Ergebnis
  - Die meisten Abweichungen betreffen Sonderzeichen und Leerzeichen
  - Mitunter Probleme bei der Zeichenkodierung
  - **Sehr wenige** echte Fehler (fehlende Titelelemente, komplett falscher Titel)

Datenaufbereitung

HOCHSCHULE DER MEDIEN

- Zugang zu den Daten der Websites
  - i.d.R. Datenbank oder Daten-API
- Vorverarbeitung und Bereinigung
- Transformation der Daten in Aussagen
  - Verwendung des Resource Description Framework (RDF)
  - URIs als Identifikatoren für die einzelnen Entitäten
  - Beschreibende Daten zu den verwendeten Attributen ebenfalls im RDF-Format
- Laden der RDF-Daten in eine gemeinsame Datenbank

Datenintegration

HOCHSCHULE DER MEDIEN

- “Matching”
  - Finden von Entitäten, die von mehreren Datenquellen beschrieben werden
  - Finden von Attributen, die dieselben Eigenschaften beschreiben

- “Merging”
  - Zusammenführen der gematchten Entitäten
    → **Neue** Entitäten mit Attributen und Beziehungen aus mehreren Quellen
  - Zusammenführen redundanter Attribute
  - Gegebenenfalls: Auflösen von Widersprüchen in den Daten

- Datenbank enthält danach sowohl die Originaldaten der Communities als auch die vereinheitlichten Daten

HOCHSCHULE
DER MEDIEN

Rechtliche Fragen

Was wollen die Communities?

HOCHSCHULE DER MEDIEN

- Grundsätzliche Bereitschaft, die kollaborativ erstellten Daten zu teilen

Aber:

- Bedürfnis, sich vor einem kompletten “Abgreifen” durch Dritte zu schützen
- Befürchtung, dass andere Angebote die eigenen Nutzer zu sich ziehen
- Wunsch nach Anerkennung der eigenen Arbeit

→ sehr **unterschiedliche** Lizenzen und Bedingungen für die Nutzung

Anforderung an offene Forschungsdaten

HOCHSCHULE DER MEDIEN

- Lizenz muss alle forschungsrelevanten Nutzungsformen erlauben
  - Speicherung, Kopie, Weitergabe
  - Veränderung, Ergänzung, Zusammenführung mit eigenen Daten
  - Analysen aller Art

- Lizenz muss international gültig sein
  - Berücksichtigung unterschiedlicher Rechtsräume (USA, EU, Japan, …)

- Im Projektkontext:
  - Einschränkung auf nicht-kommerzielle Nutzung akzeptierbar
  - Die Lizenzbedingungen dürfen sich nicht gegenseitig ausschließen
    - Ideal: gemeinsame Lizenz für alle Daten

Lösung im Projekt

HOCHSCHULE DER MEDIEN

- **Creative Commons (CC) BY-NC-SA 4.0**
  - BY: Angabe der Herkunft verpflichtend
  - NC: nicht-kommerzielle Nutzung
  - SA: muss unter gleichen Bedingungen weitergegeben werden

- Eigenschaften
  - CC Lizenzen sind international ausgelegt
  - Version 4.0 berücksichtigt Datenbanken explizit
  - “Kleinster gemeinsamer Nenner” der vorhandenen Regelungen zur Datennutzung von Seiten der Communities

Umsetzung der gemeinsamen Lizenz

HOCHSCHULE DER MEDIEN

- **Separate** Lizenzvereinbarungen mit den einzelnen Communities
  - Die meisten stellen ihre Daten unter weniger restriktiven Lizenzen zur Verfügung
  - Eine weitere Lizenz mit mehr Einschränkungen wird als unproblematisch gesehen

- Nutzung und Lizenzierung **ausgewählter** Datenelemente
  - Keine personenbezogenen Daten (Usernamen, individuelle Äußerungen in Foren, …)
  - Keine Datenelemente, die ein eigenen Urheberrecht bekommen könnten (längere Texte, inhaltliche Zusammenfassungen)

Kompatibilität mit weiteren Quellen

HOCHSCHULE DER MEDIEN

Datenquelle	Lizenz	Kompatibilität mit der CC BY-NC-SA 4.0 Lizenz
Anime Characters Database	-	Individuelle Vereinbarungen zur Lizenzierung ausgewählter Elemente unter der CC BY-NC-SA 4.0 Lizenz
AnimeClick	-	
The Visual Novel Database	ODbL	
Media-Arts Database	CC BY 4.0	**Ja, nur Angabe der Quelle**
Wikidata	CC0	**Ja, keine Einschränkungen**
AniDB (nur die zum Download angebotene Titelliste)	CC BY-NC-SA 4.0	**Ja, da identisch**

HOCHSCHULE
DER MEDIEN

Datenbank und Frontend

Software

HOCHSCHULE DER MEDIEN

- Datenverarbeitung
  - Primär eigene Python-Programme
  - Protégé: Editor für die beschreibenden RDF-Daten, Open Source
- Datenintegration
  - Nur eigene Python-Programme
- Datenbank
  - Apache Fuseki: Einfacher RDF-Triple-Store, Open Source
  - Spezielle Graph-Datenbanken haben für die aktuellen Anwendungen keine konkreten Vorteile

Frontend: Anforderungen

HOCHSCHULE DER MEDIEN

- Anzeige und Navigation der Entitäten
  - Klare Kennzeichnung der Herkunft der Daten
- Suchfunktion
  - RDF-basierte SPARQL-Abfragesprache von Fuseki unterstützt
  - Zusätzlich Suchindex für Stichwortsuche
- Anbindung weiterer Analysetools
  - Eigene Entwicklungen
  - Anwendungsbezogener Datenexport / Beliefern von Schnittstellen

→ Eigene Entwicklung als Open Source

Frontend

HOCHSCHULE DER MEDIEN

Dark Mode Search crosstab graphs languages

Goku

Property	Value	
label acdb	Goku	
type acdb	Character	
ACDB Link acdb	https://www.animecharactersdatabase.com/characters.php?id=15533	
Age acdb	Adult	
Animal Ears acdb	No	
Appears In acdb 15	• Dragon Ball (Series) • Dragon Ball GT • Dragon Ball Super • Dragon Ball Z • Dragon Ball Z: Battle of Gods • Dragon Ball Z: Bojack Unbound • Dragon Ball Z: Broly - The Legendary Super Saiyan	• Dragon Ball Z: Cooler's Revenge • Dragon Ball Z: Dead Zone • Dragon Ball Z: Lord Slug • Dragon Ball Z: Revival of 'F' • Dragon Ball Z: Super Android 13! • Dragon Ball Z: The Tree of Might • Dragon Ball Z: The World's Strongest • Dragon Ball Z: Wrath of the Dragon
Author acdb	1	
Character Role acdb	Protagonist	
Character Tag acdb	• arm guards	• karate • spiky hair

Die Methode “Tiny Use Case”

“Tiny Use Case” (TUC)

HOCHSCHULE DER MEDIEN

- Entwickelt im Rahmen des *Database Infrastructure for Global Games Culture Research (diggr)* Forschungsprojekt

- Anwendung der Idee der agilen Entwicklung von Software
  - Klar abgegrenzte Dimension der Aufgabe
  - Schnelles Erstellen von Prototypen
  - Entwicklungszyklen fokussieren auf einen Aspekt und verbessern den Prototypen inkrementell

- Hier: Konkrete Fragestellung aus der Medienwissenschaft soll mit Daten beantwortet werden

TUCs im Projekt

HOCHSCHULE DER MEDIEN

- Brücke zwischen den Anforderungen der medienwissenschaftlichen Forschung und der informationswissenschaftlichen Sicht auf die Daten
  - Eignung der Daten für konkrete Fragestellung
  - Einschätzung der Datenqualität
  - Nachvollziehbarkeit der Datenorganisation und -strukturen
- Konkrete Zwischenergebnisse
  - Dokumentation auf Projektblog
  - Präsentation und Diskussion auf Fachkonferenzen
- Schnelle und gezielte Weiterentwicklung des Frontends

TUCs im Projekt

- Der zweite Teil des Vortrags stellt die Ergebnisse eines TUCs vor

HOCHSCHULE
DER MEDIEN

HOCHSCHULE DER MEDIEN

Vielen Dank für Ihre Aufmerksamkeit!

E-Mail: pfeffer@hdm-stuttgart.de

Projekt-Website:
https://jvmg.iuk.hdm-stuttgart.de/

Direktlink JVMG-Datenbank:
https://mediagraph.link/