HOCHSCHULE DER MEDIEN

Japanese Visual Media Graph:

Bündelung des Wissens von Fan-Gemeinschaften in einem domänenspezifischen Knowledge Graph

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Magnus Pfeffer, Zoltan Kacsuk, Martin Roth

8. Jahrestagung des Verbands
»Digital Humanities im deutschsprachigen Raum«
Universität Potsdam & Fachhochschule Potsdam, online, 07.–11. März 2022

Aufbau des Vortrags

HOCHSCHULE
DER MEDIEN

- Ausgangspunkt
- Fan-Gemeinschaften
- Datenintegration
- Rechtliche Fragen
- Technische Basis
- Untersuchung zur Datenqualität
- Der “Tiny Use Case” Ansatz
- Ausblick

HOCHSCHULE
DER MEDIEN

Ausgangspunkt

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 Gemeinschaften
  - Respektieren der Wünsche und Lizenzen
  - Angebote auch für die Gemeinschaften
- 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

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

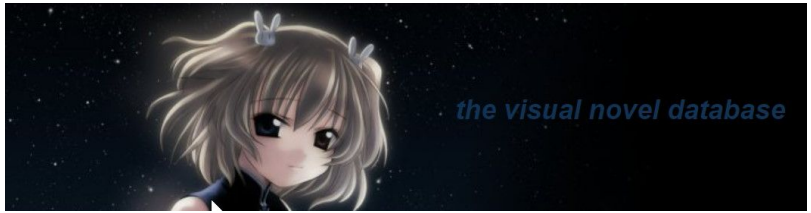
The Visual Novel Database

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

Gemeinschaft	Werke und Medien				Firmen	Charaktere	Eigenschaften	Eigenschaften	Personen
ACDB	Work					Character	Work Tag	Character Tag	People
	10207					107369	1088	4051	5557
AnimeClick	Anime Work	Comic Work				Character			Staff
	9491	11762				102143			39604
VNDB			Visual Novel	Release	Producer	Character	Tag	Trait	Staff
			28190	71349	10394	90077	2585	2777	21164

HOCHSCHULE
DER MEDIEN

Datenintegration

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 Gemeinschaften als auch die vereinheitlichten Daten

Zentrale Ontologie

HOCHSCHULE DER MEDIEN

- Spezifische Ontologie für die Daten jeder Quelle
- Identifikation der Schnittmengen
- Zusammenführung in ein gemeinsames Modell für die Domäne als Ganzes

HOCHSCHULE
DER MEDIEN

Rechtliche Fragen

Was wollen die Gemeinschaften?

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 Gemeinschaften

Umsetzung der gemeinsamen Lizenz

HOCHSCHULE DER MEDIEN

- **Separate** Lizenzvereinbarungen mit den einzelnen Gemeinschaften
  - Einige 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 eigenes Urheberrecht bekommen könnten (längere Texte, inhaltliche Zusammenfassungen)
  - Keine Datenelemente, die von Seiten der Gemeinschaften nicht geteilt werden sollen

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

Technische Basis

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: 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
  - Zusätzlicher Suchindex für Stichwortsuche mit Elasticsearch
- 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

Keitai Shoujo ケータイ少女

Property	Value
label vndb	• Keitai Shoujo • ケータイ少女
type vndb	Visual novel
Description vndb	Chihiro is a second year high school student. He is a bit impatient since he's got no girlfriend even though he gets closer to Christmas. One day, while he is on the mobile phone internet, he finds an interesting web site and downloads it. When he finishes it, his phone suddenly flashes..., and a tiny girl appears in front of him. "I'm Rin. What's your name? You have no girlfriend, don't you? I'll be in trouble if I don't make you happy." She tells him that if he doesn't find a girlfriend before Christmas, she won't be able to go back to her world. Will he be able to find a girlfriend before Christmas for himself and Rin...? \n\n[From [url=http://www.himeyashop.com/product_info.php/products_id/6137]Himeya Shop[/url]]
Has release vndb	• Keitai Shoujo PC ケータイ少女PC • Shouji Shaonv PC 手機少女 PC
Original vndb	ケータイ少女
is Primary role of vndb	• Fujimiya Momoka • Gotou Miya • Mishima Ichiru • Rin • Tomoe Sayo • Yamada Ayano
is Staff for of vndb	• KIZAWA studio • Koshimizu Ami • Yasukawa Shougo
Tag vndb	• Male Protagonist • No Sexual Content • Protagonist's Childhood Friend as a Heroine • Twin Tail Heroine
Title vndb	Keitai Shoujo
VNDB link vndb	http://vndb.org/v1010
Visual novel length vndb	Medium (10 - 30 hours)
Wikidata link vndb	https://www.wikidata.org/wiki/Q11075743

1 2 3 4 5

HOCHSCHULE DER MEDIEN

HOCHSCHULE
DER MEDIEN

Untersuchung zur Datenqualität

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)

Datenfeld	Angabe	Anzahl	Anteil	Konfidenz-Grenze	Konfidenz-Grenze
ACDB English title Stichprobe: 424 Auswahlgrundlage: 2438	Korrekter Titel	312	73.585%	68.585%	78.585%
	Typografische Abweichungen	111	26.179%	21.179%	31.179%
	Falsche Angabe	1	0.236%	0.041%	5.236%
ACDB Japanese title Stichprobe: 424 Auswahlgrundlage: 2438	Korrekter Titel	345	81.368%	76.368%	86.368%
	Typografische Abweichungen	77	18.160%	13.160%	23.160%
	Falsche Angabe	2	0.472%	0.082%	5.472%

Datenfeld	Angabe	Anzahl	Anteil	Untere Grenze	Obere Grenze
English title Stichprobe: 483 Auswahlgrundlage: 967	Korrekter Titel	333	68.944%	63.944%	73.944%
	Typografische Abweichungen	136	28.157%	23.157%	33.157%
	Falsche Angabe	12	2.484%	0.131%	7.484%
	Fehlende Daten	2	0.414%	0.022%	5.414%
Japanese title Stichprobe: 483 Auswahlgrundlage: 967	Korrekter Titel	367	75.983%	70.983%	80.983%
	Typografische Abweichungen	88	18.219%	13.219%	23.219%
	Falsche Angabe	8	1.656%	0.087%	6.656%
	Fehlende Daten	20	4.141%	0.218%	9.141%

Datenfeld	Angabe	Anzahl	Anteil	Konfidenz Grenze	Konfidenz Grenze
VNDB English title Stichprobe: 503 Auswahlgrundlage: 28771	Korrekter Titel	475	94.433%	89.433%	99.433%
	Typografische Abweichungen	28	5.567%	0.567%	10.567%
VNDB Original title Stichprobe: 503 Auswahlgrundlage: 28771	Korrekter Titel	460	91.451%	86.451%	96.451%
	Typografische Abweichungen	40	7.952%	0.142%	12.952%
	Falsche Angabe	2	0.398%	0.007%	5.398%
	Nicht ermittelbar	1	0.199%	0.004%	5.199%

Der “Tiny Use Case” Ansatz

“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

HOCHSCHULE
DER MEDIEN

Ausblick

Nächste Schritte

HOCHSCHULE DER MEDIEN

- Aktuelles Projekt
  - Abschluss des Match/Merge Prozesses
  - Veröffentlichung eines Entwurfs der gemeinsamen Ontologie
  - Veröffentlichung und Dokumentation aller Software
  - Veröffentlichung aller Daten zum Download
- Nächste Projektphase
  - Weiterentwicklung des Prototypen zu einem stabilen Services
  - Erweiterung der Datengrundlage um zusätzliche Aspekte
  - Zusammenarbeit mit anderen Forschenden → neue Use Cases
  - Dokumentation und Workshops für Wissenschaftler:innen

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/