HOCHSCHULE DER MEDIEN

Japanese Visual Media Graph

Ein Beispiel für offene Forschungsdaten

Förderung durch die

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Martin Roth und Magnus Pfeffer

Japanologentag 2022: Open Scholarship und Japanologie

HHU Düsseldorf

online

25. August 2022

Die Vortragenden

Martin Roth ist Research Fellow an der Hochschule der Medien Stuttgart und Associate Professor an der Graduate School of Core Ethics and Frontier Sciences der Ritsumeikan Universität.
Er betreut die medienwissenschaftlichen Aspekte des Projekts.

Magnus Pfeffer ist Professor für Informationsmanagement an der Fakultät Information und Kommunikation der Hochschule der Medien Stuttgart. Er betreut die informationswissenschaftlichen Aspekte des Projekts.

Übersicht der Präsentation

HOCHSCHULE DER MEDIEN

1. Einführung in das **JVMG Projekt**
2. **Datenqualität**
3. **Datenintegration und Matching**
4. **Rechtliche Harmonisierung** der Daten
5. Beispielanwendungen und **Ausblick**

Einführung in das JVMG Projekt

HOCHSCHULE DER MEDIEN

- Datenbanken der Enthusiasten-Gemeinschaften sind wichtige Ressourcen für Medienwissenschaftler*innen

- **Japanese Visual Media Graph** (JVMG) soll diese Daten zusammenführen und für quantitative Forschung nutzbar und verfügbar machen
- Gefördert durch die **Deutsche Forschungsgemeinschaft** im Rahmen des e-Research Technologies Programms

Stützpfeiler des JVMG Projekts

HOCHSCHULE DER MEDIEN

COLLABORATION WITH COMMUNITIES

We work with diverse fan and enthusiast communities to make their data available to researchers. We respect their wishes and conditions for the use of the data.

SUITABILITY FOR RESEARCH

The needs of researchers drive all aspects of the development of the graph database, from the choice of data sources and the data model to specific representation details.

OPEN DEVELOPMENT

We document the process of data integration and the associated research in an open manner. You will find regular updates on the project blog.

Arten von Enthusiasten-Gemeinschaften

HOCHSCHULE DER MEDIEN

- Breites Interesse an Japan und japanischen Medien
  - Eher allgemeine Informationen
  - Breite Abdeckung der Domäne
- Fokussierung auf eine Medienart
  - z.B. Manga, Anime, Visual Novel Spiele
  - Detaillierte Informationen zum Medium und allen damit verbundenen Aspekten
- Fokussierung auf einen Aspekt der japanischen visuellen Medien
  - z.B. visuelle Charaktere, Sprecher:innen, Musik
  - Sehr spezifische Informationen

Warum Datenbanken von Enthusiasten?

HOCHSCHULE DER MEDIEN

Enthusiast community	Works and media				Company	Characters	Work properties	Character properties	Involved people
ACDB	Work					Character	Work Tag	Character Tag	People
	10.207					107.369	1.088	4.051	5.557
AnimeClick	Animation Work	Comic Work				Character			Staff
	9.491	11.762				102.143			39.604
VNDB			Visual Novel	Release	Producer	Character	Tag	Trait	Staff
			28.190	71.349	10.394	90.077	2.585	2.777	21.164

Weitere wichtige Ressourcen

HOCHSCHULE DER MEDIEN

Database	Works and media					Characters	
Wikidata	Anime titles	Manga series	Video game	Light novel & LN series		Anime character	Manga character
	4.467	13.871	47.192	867		3.788	2.990
Media-Arts Database	Anime titles	Anime items	Game items	Manga book series	Manga magazine issues		
	12.085	~135.000	~61.000	133.779	170.670		

HOCHSCHULE DER MEDIEN

Datenqualität

Stichproben-basierte Prüfung

HOCHSCHULE DER MEDIEN

- Überprüfung einer Stichprobe der Titeldaten von Medien
  - Titel in japanischen Kanji
  - Vergleichsdaten
    - Abbildungen von Verpackungen
    - Titelinformationen im Serien-Vorspann
    - Angaben der Produzenten auf deren Website

- Stichprobe ausreichend groß für Rückschluss auf Gesamtdaten

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
VNDB English title: Visual Novels Sample: 503 Population: 28170	Correct titles	475	94.433%	89.433%	99.433%
	Typographical errors	28	5.567%	0.567%	10.567%
VNDB Original title: Visual Novels Sample: 503 Population: 28170	Correct titles	460	91.451%	86.451%	96.451%
	Typographical errors	40	7.952%	0.142%	12.952%
	Misrepresentation errors	2	0.398%	0.007%	5.398%
	Cannot be determined	1	0.199%	0.004%	5.199%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
Wikidata English title: Anime Sample: 381 Population: 1468	Correct titles	319	83.727%	78.727%	88.727%
	Misrepresentation errors	37	9.711%	4.711%	14.711%
	Missing data	20	5.249%	0.249%	10.249%
	Not anime	5	1.312%	0.341%	6.312%
Wikidata Japanese title: Anime Sample: 381 Population: 1468	Correct titles	293	76.903%	71.903%	81.903%
	Typographical errors	1	0.262%	0.068%	5.262%
	Misrepresentation errors	15	3.937%	1.022%	8.937%
	Missing data	63	16.535%	11.535%	21.535%
	Not anime	9	2.362%	0.613%	7.362%

Datenqualität: Ergebnisse

HOCHSCHULE DER MEDIEN

- Die Daten der **Enthusiasten**-Gemeinschaften weisen eine hohe Genauigkeit auf

- **Wikidata** ist im Vergleich deutlich ungenauer und es gibt zahlreiche Einträge mit unvollständigen Titeldaten

- Die **Media-Arts Database** enthält spezifische Fehler, die auf eine Nutzung von OCR-Verfahren für die Datenakquise hinweisen

HOCHSCHULE
DER MEDIEN

Datenintegration und Matching

Grundlegender Ablauf

HOCHSCHULE DER MEDIEN

- Daten der Communities werden **gefiltert und vorbereitet**
- Alle Informationen werden in als einzelne Aussagen im **Resource Description Framework Format** überführt
- **Überlappungen** zwischen den Datenquellen werden durch Abgleich (**Matching**) identifiziert
- Eine **gemeinsame JVMG Ontologie** gleicht die Bezeichnungen der beschreibenden Datenfelder an und führt Inhalte zusammen

Matching: Vorgehen

HOCHSCHULE DER MEDIEN

- Abgleich beginnt bei den Werken
  - Visuelle Charaktere oder beteiligte Personen sind für sich betrachtet nicht eindeutig, aber die Verknüpfung mit einem Werk löst diese Mehrdeutigkeiten auf
- Abgleich zwischen den Quellen ist iterativ
  - Ausgehend vom größten Datensatz für einen bestimmten Medientyp
  - Kleinere Datensätze werden mit dem größeren Eintrag für Eintrag abgeglichen
  - Erwartung: sehr kleine Anzahl bislang nicht vorhandener Einträge
- Abgleich nutzt ausschließlich die japanischen Titeldaten
  - Alle Einträge haben diese Angaben
  - Die Qualität der Titeldaten ist hoch

Matching: Praktische Ergebnisse

HOCHSCHULE DER MEDIEN

- Mögliche Fälle beim Abgleich zweier Quellen
  - **Eindeutige** Übereinstimmung → Automatisches Zusammenführen
  - **Nicht-eindeutige** Übereinstimmung → menschliche Prüfung
  - **Keine** Übereinstimmung → Erneuter Versuch mit unscharfer Suche → menschliche Prüfung der gefundenen Kandidaten

- Beispiel: animeclick.it → anidb.net, Anime Werke
  - 7467 Einträge in animeclick
  - 6334 eindeutige Übereinstimmungen (~85%)
  - 397 nicht-eindeutige Übereinstimmungen (~5%)
  - 736 ohne Übereinstimmung selbst bei unscharfer Suche (~10%)

Matching: Nicht-eindeutige Einträge

Titolo originale	Haikyu!! stop motion
Titolo inglese	Haikyu!! stop motion
Titolo Kanji	ハイキュー!!
Nazionalità	Giappone
Categoria	Serie OAV Corto
Genere	Scolastico Sport
Anno	2020
Tratto da	Manga Shounen
Stagioni	Inverno (2020)
Episodi	10
Stato in patria	annunciato
Stato in Italia	Doppiaggio inedito, Sottotitoli inedito

Titolo originale	Haikyū!!
Titolo inglese	Haikyu!!
Titolo Kanji	ハイキュー!!
Nazionalità	Giappone
Categoria	Serie TV
Genere	Commedia Scolastico Sport
Anno	2014
Tratto da	Manga shōnen
Stagioni	Primavera (2014) Estate (2014)
Episodi	25
Stato in patria	completato
Stato in Italia	Doppiaggio completato, Sottotitoli completato

Herausforderung: Granularität der Daten

HOCHSCHULE DER MEDIEN

- Medienwerte können unterschiedlich repräsentiert werden
  - Fernsehserien als einzelner Eintrag oder als mehrere Staffeln
  - Sonderepisoden als Teil der Serien oder einzelner Eintrag
  - Mehrteilige Kinofilme als einzelner oder separate Einträge

- Lösungsansatz
  - Dokumentation von 1:n oder n:m Zuordnungen
  - Entscheidung für die Darstellung im JVMG-Wissensgraphen
    - Informationen aus mehreren Einträgen in einen einzelnen aggregiert
    - Informationen aus einem separaten Eintrag auf mehrere verteilt

HOCHSCHULE
DER MEDIEN

Rechtliche Harmonisierung der Daten

Herausforderungen

HOCHSCHULE DER MEDIEN

- **Lizenzierungspraxis** der Gemeinschaften
  - Kein Bewusstsein für die Problematiken des Urheberrechts
  - Unterschiedliche und oft inkompatible Lizenzen

- **Befürchtungen** der Gemeinschaften
  - Vollständiges Übernehmen ihrer Arbeit
  - Weniger Zugriffe auf ihre Websites
  - Keine Anerkennung ihrer Arbeit

- **Erfordernisse** des JVMG-Projekts
  - Offene Lizenzen für alle Daten
  - Abdeckung aller Nutzungsarten der Forschung
  - Weltweite Gültigkeit der Lizenzen

Lizenzen: Übersicht

HOCHSCHULE DER MEDIEN

Datenquelle	Lizenz	Kompatibilität mit der CC BY-NC-SA 4.0 Lizenz
Anime Characters Database	-	Gesonderte CC BY-NC-SA 4.0 Lizenz für die vom JVMG-Projekt genutzten Teile der Daten
AnimeClick	-	
The Visual Novel Database	ODbL	
Media-Arts Database	CC BY 4.0	**Ja**
Wikidata	CC0	**Ja**
AniDB (frei verfügbare Anime Titeldaten)	CC BY-NC-SA 4.0	**Identisch**

HOCHSCHULE DER MEDIEN

Beispielanwendungen und Ausblick

Projektinterne Anwendungsfälle (Tiny Use Cases)

HOCHSCHULE DER MEDIEN

1. Kartographie von **Visual Novel Charakteren**
2. Überprüfung von Hiroki Azuma Thesen aus **“Otaku: Japan’s Database Animals”**
3. **Wiederkehrende Muster in Charakteren in** visual novel games
4. **Media Mix** aus der Perspektive von Netzwerken gemeinsam auftretender Charaktere
5. **Zensus von Charakteren** in Japanese visual media

Nächste Schritte

HOCHSCHULE DER MEDIEN

- Geplante Zusammenarbeit und Erweiterung des Knowledge Graphs:
  - **Anime production** studies
  - **Fan fiction** als kulturelle Evolution
  - **Genre and trope** als Diskurs und Inhaltsbeschreibung
- Training und Unterstützung für potentielle Nutzende
  - **Tutorials** und Dokumentation von Beispielanwendungen
  - **JVMG lab** als konzentriertes Format

HOCHSCHULE DER MEDIEN

Vielen Dank für Ihre Aufmerksamkeit

Kontaktieren Sie uns gern: rothm@hdm-stuttgart.de, pfeffer@hdm-stuttgart.de

Projektwebseite und Blog:
https://jvmg.iuk.hdm-stuttgart.de/

JVMG Datenbank:
https://mediagraph.link/