HOCHSCHULE DER MEDIEN

Introducing the JVMG Project

An Open Knowledge Graph for Researchers Working on Popular Japanese Visual Media

Research funded by:

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Zoltan Kacsuk

Workshop on Digital Humanities, Linked Open Data and Games Research

Stuttgart Media University

15 August 2022

Outline of the presentation

HOCHSCHULE DER MEDIEN

1. Introducing the **Japanese Visual Media Graph** (JVMG) project
2. **Data integration** and **matching**
3. **Legal harmonization** of the data
4. Measuring **data quality**
5. The **Tiny Use Case** (TUC) workflow methodology

Introducing the JVMG project

HOCHSCHULE DER MEDIEN

- Databases by fan/enthusiast communities are the **go to resource for checking information**

- **Japanese Visual Media Graph** (JVMG) project
- Project aim: Make these databases available for **large-scale quantitative research, in collaboration with the communities**
- Funded by the **German Research Foundation**'s (Deutsche Forschungsgemeinschaft) e-Research Technologies program

Key characteristics of the JVMG project

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.

Source databases

HOCHSCHULE DER MEDIEN

Fan/enthusiast community databases:

- **AnimeClick:** Wide interest in Japanese visual media and culture
- **The Visual Novel Database** (VNDB): Focused on visual novel games only
- **Anime Characters Database** (ACDB): Focus on one aspect of the domain

Other databases:

- **Wikidata:** Not focused on Japanese visual media
- **Media-Arts Database:** Collects information on manga, animation, games and media art from institutions, creators and publishers in Japan

Entity and concept numbers

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

Entity and concept numbers

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

Data integration and matching

Data integration

- Data is **cleaned and preprocessed**
- All data transformed into **RDF** form
- **Unified JVMG ontology** integrating the individual ontologies
- **Connections** (e.g. matching) between the databases
- **Can be connected** to other linked data sources

Matching: Process

HOCHSCHULE DER MEDIEN

- Process starts with **media works entities**
  - Character or people entities are ambiguous on their own, but less so if connected to a work
- Matching media entities starts with two sources and iterates
  - **Largest** dataset for a given media type is the starting point
  - Other datasets are matched against the largest set
  - Expectation: Very few entries are in the small sets, but not in the largest
- Matching is based on Japanese title (kanji form) **only**
  - This data point is available for all entities
  - Quality assessment has shown very accurate entries
  - Other information (e.g. dates) might not be consistent over multiple sources

Matching: First results

HOCHSCHULE DER MEDIEN

- Possible cases
  - **Non-ambiguous matches** between sources → merged entity (cluster)
  - **Ambiguous matches** (i.e. more than one entity with exact same title) → human disambiguation
  - **No match** → relaxed typographic matching → human checking

- Example: animeclick.it → anidb.net
  - 7467 anime entities in animeclick
  - 6334 non-ambiguous matches (~85%)
  - 397 ambiguous matches (~5%)
  - 736 cannot be matched even when relaxed (~10%)

Matching: Principal errors

- **False positive**
  - matching two non-identical entities
- **False negative**
  - not matching two identical entities

- **Risk management** for errors
  - Reliance on **kanji title only limits the false positives**
  - **Human interaction** should catch most **false negatives**
  - As <15% of entries need to be checked by humans, this can be done for all smaller databases

Matching: Additional problems

HOCHSCHULE DER MEDIEN

- Media works **granularity** of representation
  - TV series can be documented as a single entry or multiple seasons
  - Extra episodes can be part of the series or separate
  - Multi-part movies can be single or multiple entries

- Current workflow
  - Document such 1:n or n:1 matching relations
  - For the final knowledge graph, a decision has to be made for the representation
  - Information of part entities can be **aggregated** into a single entity
  - Information of single entity can be **distributed** into parts entities

Matching granularity example: Aggretsuko

HOCHSCHULE DER MEDIEN

	AniDB	AnimeClick	Anime Characters Database
4 web seasons	Yes	Yes	Yes
1 announced(?) web season	Yes	No	Maybe?
1 web christmas special	Yes	Yes	Maybe?
1 TV series	Yes	Yes	Yes

Legal integration of the data

Problems and challenges

HOCHSCHULE DER MEDIEN

- **Licensing practices** of the communities
  - **Lack of awareness** of or disregard for copyright issues
  - Varying and often **incompatible licenses**
- **Concerns** of the communities
  - **Wholesale copying** of their work
  - **Traffic subverted** from their sites
  - **Lack of acknowledgment** of their work
- **Licensing needs** of the JVMG project
  - The license has to be **open**
  - Need to find the **lowest common denominator**
  - Have to cover **most jurisdictions**

Overview of JVMG project data sources

HOCHSCHULE DER MEDIEN

Data source	License	Compatibility with the CC BY-NC-SA 4.0 license
Anime Characters Database	-	CC BY-NC-SA 4.0 license provided for the JVMG project by **individual agreement** for the parts used in each case
AnimeClick	-	
The Visual Novel Database	ODbL	
Media-Arts Database	CC BY 4.0	**yes**
Wikidata	CC0	**yes**
AniDB (publicly available anime titles data dump only)	CC BY-NC-SA 4.0	**identical**

HOCHSCHULE
DER MEDIEN

Measuring data quality

Assessing data accuracy

HOCHSCHULE DER MEDIEN

- **Random sample** of anime (or visual novel) titles
- **Sample sizes determined** so that statistical estimates can be drawn for the population parameters
- **Manual checking** of sample elements against ground truth or official websites, etc.

Accuracy main findings

HOCHSCHULE DER MEDIEN

- **Enthusiast community databases** have a **very high accuracy**
  - Most errors are only **typographical errors**
- **Wikidata** suffers from a variety of problems
  - Large number of problems with the **title contents**
  - Largest number of **missing data**
  - **Miscategorization** errors
- **Media-Arts Database** errors are mostly due to the **automatic processing** of the material and excess information added to the title field
  - **OCR** results need to be checked and edited manually
  - A lot of **disambiguation** information added to titles

The Tiny Use Case (TUC) workflow methodology

General idea for the TUC workflow methodology

HOCHSCHULE DER MEDIEN

- Pioneered by the **diggr** (Databased Infrastructure for Global Games Culture Research) research project team
- Inspiration from **agile** software development principles
  - Cycle of continuous incremental innovations and assessments
- Each TUC **3-4 months** long

Learning from TUCs

- Exploration of the **data's usability for research**
  - Limits of the data
  - **Data quality** issues
- Needs of the researchers in relation to the **frontend**
- **Bridging disciplinary boundaries**
  - between library and computer science on the one hand, and humanities and social science on the other

Examples of Tiny Use Cases

HOCHSCHULE DER MEDIEN

1. Investigating Japanese **Visual Novel Characters**
2. Testing one of the points from Hiroki Azuma's **"Otaku: Japan's Database Animals"**
3. **Exploring recurring patterns in character creation** in visual novel games
4. **Examining the concept of media mix** by looking at networks of co-appearing characters
5. **Census of characters** in Japanese visual media

Future work

HOCHSCHULE DER MEDIEN

- Planned collaborations and expansion of the knowledge graph:
  - Databased **anime production** studies
  - Exploring **fan fiction** and cultural evolution
  - Working with **genre and trope** definitions
- Providing training and research assistance:
  - Developing **tutorial materials**
  - **JVMG lab** events

HOCHSCHULE DER MEDIEN

Thank you for your attention!

Get in touch at: rothm@hdm-stuttgart.de, pfeffer@hdm-stuttgart.de, kacsuk@hdm-stuttgart.de, malmsheimer@hdm-stuttgart.de

Visit our project website:
https://jvmg.iuk.hdm-stuttgart.de/

Visit the JVMG database:
https://mediagraph.link/