Introduction to Digital Humanities with the Japanese Visual Media Graph project

Research funded by:

DFG Deutsche Forschungsgemeinschaft German Research Foundation

Magnus Pfeffer and Zoltan Kacsuk

Graduate School of International Culture and Communication Studies, Waseda University

3 June 2025

About Magnus Pfeffer

HOCHSCHULE DER MEDIEN

- Professor at Stuttgart Media University
- CV
  - Studied Computer Science and Information science
  - Worked at a large university library as subject specialist
- Research
  - Metadata management, information retrieval
  - Ontologies, knowledge graphs

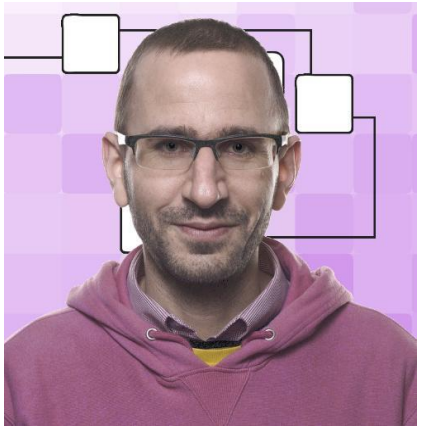
About Zoltan Kacsuk

HOCHSCHULE DER MEDIEN

- Doctor of Arts in Manga Studies, **Kyoto Seika University**, 2016

- Work on:
  - Subcultures, fan cultures, geek and otaku culture
  - Comics and manga studies
  - Data science and **digital humanities**

- Current position:
  - **Post-doctoral researcher** on the JVMG project



Stuttgart, Germany

- State capital of Baden-Württemberg
  - Population: ~500000 within city limits, 2.8 million in the urban region
  - Automotive industry
  - Highly specialized industry
  - Science
  - Media
- http://www.stuttgart.de/en/

Stuttgart Media University

- State university of applied sciences
- 21 accredited study courses (15 BA, 6 MA programs)
- 5,000 students, 130 professors
- 190 employees and associated lecturers
- 1,200 available places per year
- Library with about 90,000 items
- Large student workspaces

Outline of the presentation

1. Introducing Digital Humanities
2. The Japanese Visual Media Graph Project

HOCHSCHULE
DER MEDIEN

Introducing Digital Humanities

What is digital humanities?

HOCHSCHULE DER MEDIEN

- No set definition, often abbreviated as DH
  - **Humanities work with digital tools**
- Most of us use at least some form of digital tools:
  - word processors
  - digital library catalogues
  - online search engines
- Digital humanities projects usually involve:
  - working with **dedicated digital data sets**
    - these can be **big data**
  - using **library and information science (LIS) frameworks**
  - drawing on **software (and statistical) tools**

Different aspects of DH projects

HOCHSCHULE DER MEDIEN

- **Data storage/organization/retrieval**
  - Digitization projects
  - Database building
    - Linked Open Data (LOD) & the semantic web
  - Interface creation
- **Data processing and analysis**
- **Data visualization**

Not all DH projects involve all three of these aspects.

Common principles of DH projects

HOCHSCHULE DER MEDIEN

- Commitment to **openness and accessibility**
  - Open source technologies
  - Open licenses (e.g. Creative Commons)
  - Open access publications
- Prioritizing **collaboration**
  - Among researchers/teams
  - Between disciplines
- Adopting **approaches from IT and LIS**
  - Versioning and iterative development
  - Metadata standards
- Retaining a **critical perspective** (?)

DH and traditional humanities research

HOCHSCHULE DER MEDIEN

- Is DH really different from traditional humanities research?
  - Is it just about **digitalization**?
  - Is it just about **scale**?
  - Is it about new (born-digital) **objects of research**?
- Is DH just a fad or a **marketing buzzword** to get funding?
- What are the **strengths and weaknesses** of DH compared to traditional humanities research?
- How can DH and traditional humanities research **complement each other**?

Many resources

HOCHSCHULE DER MEDIEN

- **Datasets**
  - ROIS-DS CODH datasets, JVMG, etc.
- **Software tools**
  - Gephi, Voyant Tools, etc.
- **Online services**
  - Aggregation, publication and search platforms: e.g. CrossAsia
  - Online software tools: e.g. Voyant Tools
- **Other resources**
  - DH is in a development/growth phase

Datasets can be grouped by

HOCHSCHULE DER MEDIEN

- **Type of data** (numbers, text, image, audio, video, geolocation data, descriptive metadata, etc.)
- **Format** and **encoding**
  - Required **technical skills**
- **“Born-digital” vs digitized data**
  - Level of processing for digitized data (OCR, annotation, etc.)
- **Research domain** (literature, linguistics, politics, history, …)
- **Content** of data
- Type of **license**
- **Intended use** (archive, training data, etc.)
- **Language**
- …

ROIS-DS Center for Open Data in the Humanities

Many projects (see: http://codh.rois.ac.jp/index.html.en)
Datasets (http://codh.rois.ac.jp/dataset/):

- **Text data:**
  - Dataset of Pre-Modern Japanese Text
  - Dataset of Edo Cooking Recipes
- **Image data:**
  - Collection of Facial Expressions (KaoKore)
  - Ukiyo-e Face Datasets
  - Edo Shopping Guide
  - Edo Sightseeing Guide
  - Dataset of Modern Magazines
  - (North China Railway Archive (http://codh.rois.ac.jp/north-china-railway/)
- **Training data:**
  - Kuzushiji Dataset
  - Seal Script Dataset
- **Map (geolocation) data**
  - Edo Maps Beta
  - Geoshape Repository

The Japanese Visual Media Graph Project

HOCHSCHULE DER MEDIEN

Data-driven methodologies

From digital humanities to **computational humanities**

- Quantitative **textual analysis** - distant reading, culturomics
- **Web native** approaches - ‘digital methods’, webometrics
- **Algorithmic** data **extraction** - cultural analytics
- **Metadata** analytics

Metadata analytics

HOCHSCHULE DER MEDIEN

- Roots in **bibliometrics** and **scientometrics**
- Based on **descriptive metadata** in library and information science
- Large metadata databases built by **enthusiast communities** online*
- Understanding cultural objects and processes **at scale**

Introducing the JVMG project

HOCHSCHULE DER MEDIEN

- Observation: Databases by enthusiast communities are often the **most up-to-date online resource for checking information**

- Project aim: Make these databases available for **large-scale quantitative research, in collaboration with the communities**
  - Funding: German Research Foundation (DFG), 6 years duration

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

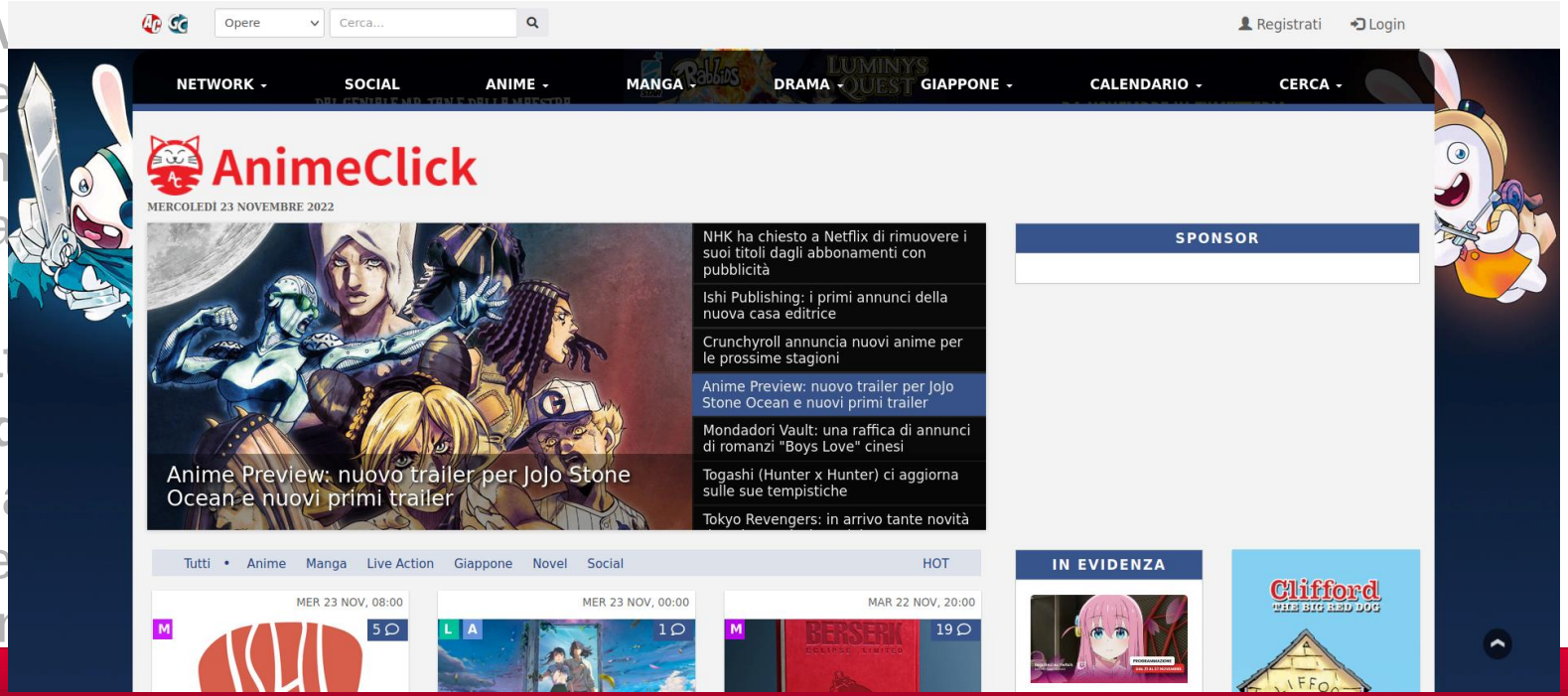
Source databases

Fan/enthusiast community databases:

- **AnimeClick:** Wide interest in Japanese visual media and culture
- **The V**
  game
- **Anim**
  doma

Other da

- **Wikid**
- **Media**
  game
  Japar



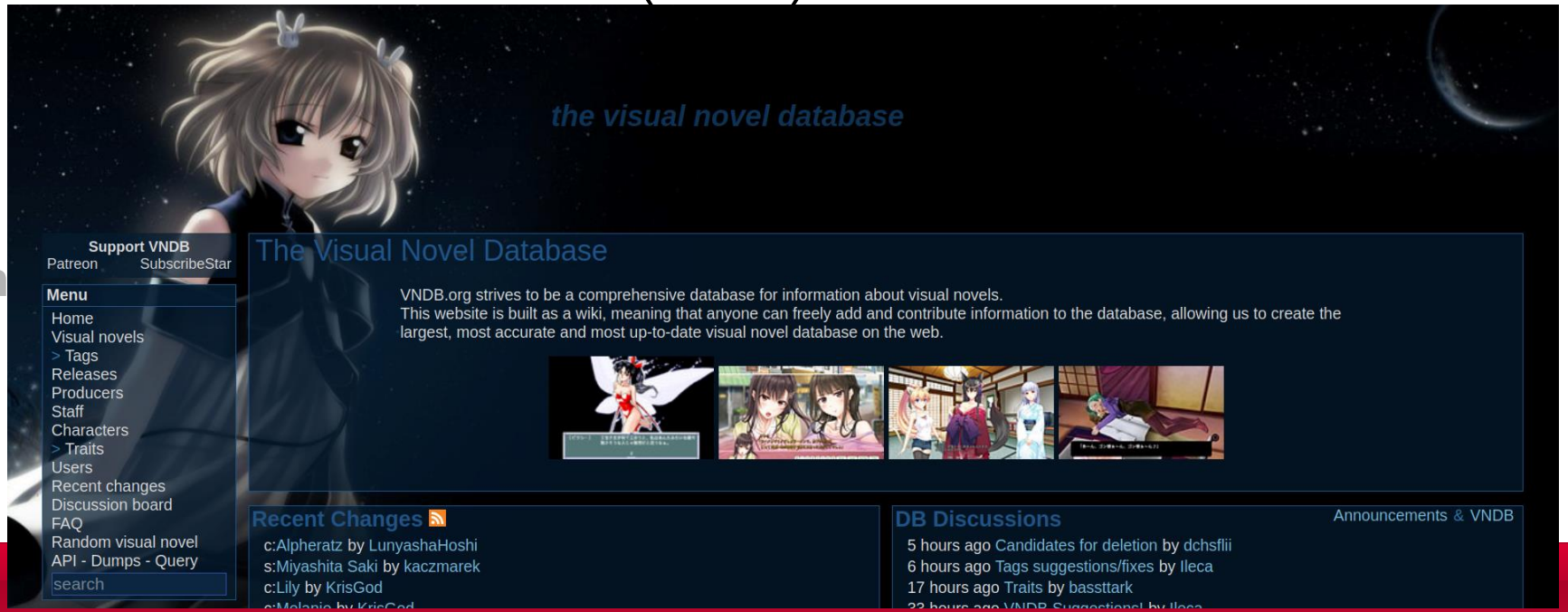
Source databases

Fan/enthusiast community databases:

- **AnimeClick:** Wide interest in Japanese visual media and culture
- **The Visual Novel Database** (VNDB): Focused on visual novel
-

Oth

-
-



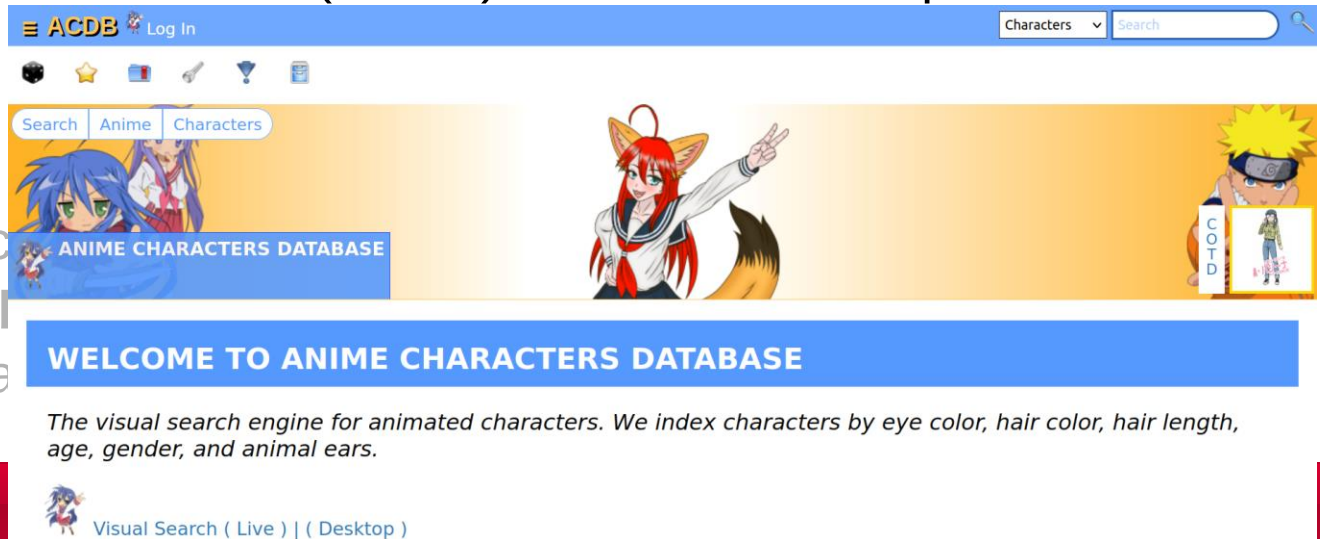
Source databases

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 foc
- **Media-Arts Datal** games and media Japan



HOCHSCHULE DER MEDIEN

and culture

ual novel

aspect of the

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

HOCHSCHULE DER MEDIEN

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

HOCHSCHULE DER MEDIEN

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

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

HOCHSCHULE
DER MEDIEN

Data integration and matching

Data integration

HOCHSCHULE DER MEDIEN

- 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

- 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

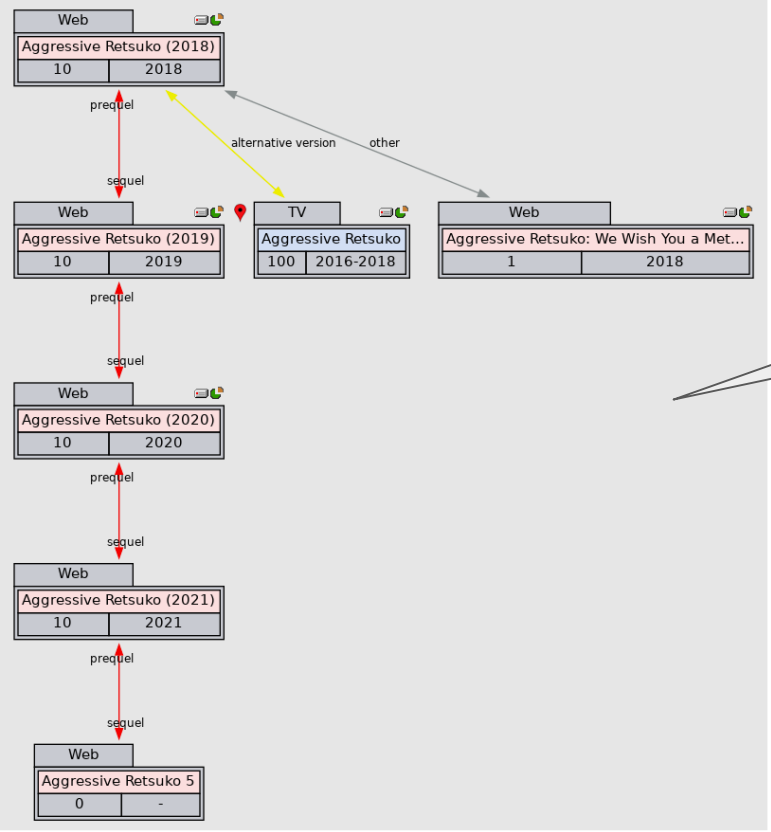
Matching: Additional problems

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



AniDB:

4 web seasons with 10 episodes each
1 announced(?) web season
1 TV series with 100 (short) episodes
1 web christmas special

Image source: anidb.net

Matching: Granularity example

AnimeClick:
4 web seasons with 10 episodes each
1 TV series with 100 (short) episodes
1 web christmas special

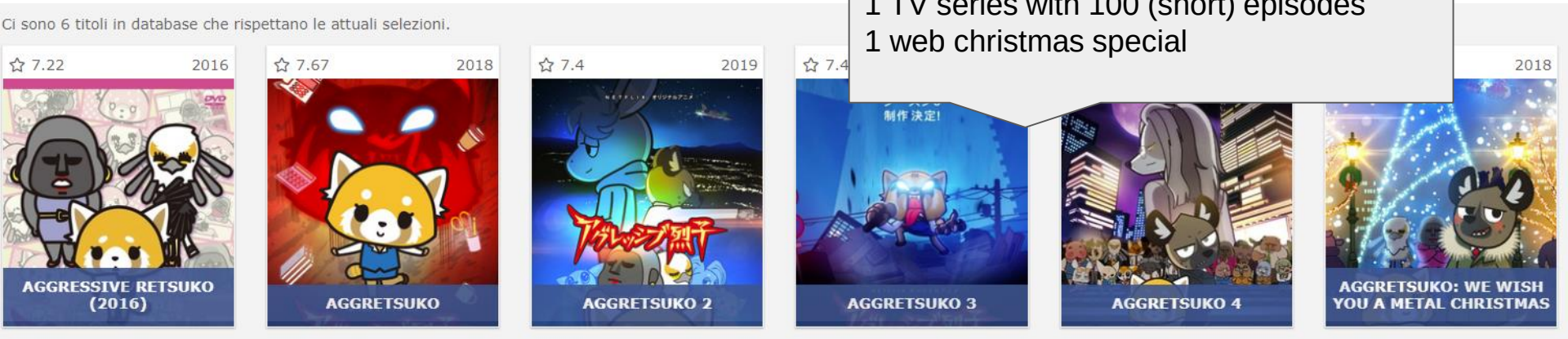


Image source: www.animeclick.it

Matching: Granularity example

Aggretsuko

ONA

Aggressive Retsuko

Anime

Anime Characters Database:
1 TV Anime (no episode information)
1 Web ONA (no episode information)

Image source: www.animecharactersdatabase.com

Matching: Granularity example

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

Problems and challenges

HOCHSCHULE DER MEDIEN

- **Licensing needs** of the JVMG project
  - The license has to be **open, allowing all kind of research**
  - 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

Web interface

Web interface requirements

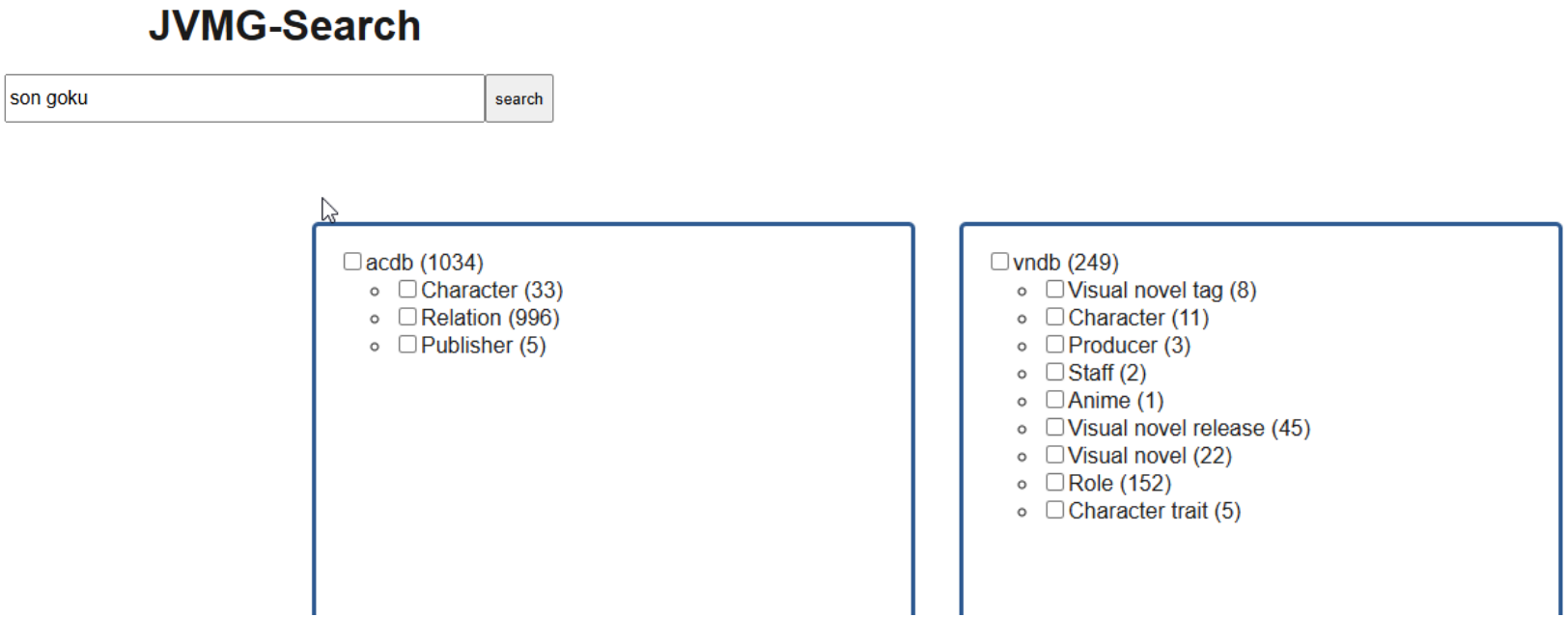
HOCHSCHULE DER MEDIEN

- Display and Navigation
  - Readable labels for all properties
  - Clickable links for all connected entities
  - Clear attribution of data origin
- Search
  - RDF-specific search using SPARQL
  - Keyword search using index
- Analysis Tools
  - User-task specific export of data / data delivery for APIS
  - Individually developed extensions

→ Decision to develop a new open source web frontend

Web-Frontend: Facetted search interface

HOCHSCHULE DER MEDIEN



Web interface: Presentation and navigation

Hair Color (Hair)

Resource: http://mediagraph.link/vndb/trait/2

Graph: vndb

Property	Value
Alias en	Hair Colour en
Applicable en	false
Default spoil status en	0 en
Group en	Hair en
Name en	Hair Color en
Searchable en	false
Trait parent en	Hair en
is Trait parent en of 15	• Black (Hair) en • Blond (Hair) en • Blue (Hair) en • Brown (Hair) en • Cyan (Hair) en • Dyed (Hair) en • Green (Hair) en • Grey (Hair) en • Multicolored (Hair) en • Orange (Hair) en • Pink (Hair) en • Red (Hair) en • Teal (Hair) en • Violet (Hair) en • White (Hair) en
VNDB link en	https://vndb.org/i2
label	Hair Color (Hair) en
sexual status en	false
type	Character trait en

Dark Mode Search Create Index crosstab languages

Shinpi no Hou 神秘の法 Shinpi no Hou

Resource: http://mediagraph.link/jvmg/mediawork/10

Graph: anidb

hide

Property	Value
label	Shinpi no Hou x-jat
AniDB link en	https://anidb.net/anime/9081
Official title en	神秘の法 ja
Primary title en	Shinpi no Hou x-jat
Synonym en	• Happy Science Movie 6 x-unk • The Mystical Laws en

Graph: http://mediagraph.link/graph/cluster

hide

Property	Value
http://mediagraph.link/jvmg/ont/hasMember	• Shinpi no Hou x-jat • 神秘の法
label	Shinpi no Hou

Graph: madb

hide

Property	Value
https://mediaarts-db.bunka.go.jp/data/property#periodDisplayed	10/6より公開。日活配給。
https://mediaarts-db.bunka.go.jp/data/property#source	アニメアワード投票用データ ／ 年間パーフェクト・データ（AWD+APDY）
actor	【Sho Shishimaru】獅子丸 翔 （KOYASU TAKEHITO） / 【Leika Chan/Theta】張 麗華/シータ （AYUMI FUJIMURA） / 【Konohana©no©Sakuya©Bime】木花開耶姫 （RYOKA YUZUKI） / 【Utica】ユチ

Web interface: Overview page

HOCHSCHULE DER MEDIEN

Overview

Anime Characters Database Graph

Character A character found in ACDB.	131029
Publisher Within ACDB, Publisher refers to a title or series to which a character belongs, and not an entity that is engaged in a publishing role.	12381
Tag A class for descriptive details about an anime or manga.	331
Character Tag A class for descriptive details about a character	1240
Individual A real-world individual associated with a character.	9308

AnimeClick Graph

Character Character of a manga or anime	57579
Author Author of an anime or manga	59502
Staff A person or group involved in the creation or publication of a creative work.	187017
Episode Single episode of an anime	212396
Creative work Creative work includes visual novels and animes	36604
Tag A class for descriptive details about an anime or manga.	2885

Media Arts Database Graph

Agent	60383
Animation TV program	124733
Animation video package	9191
Game package	48118
Manga book	331862
Manga other	5805
Media Art exhibition or performance	14335

Visual Novel Database Graph

Anime Indicates the anime of a visual novel	1831
Visual novel A specific visual novel.	45612
Character A visual novel character.	126578
Staff An individual staff member that has contributed to the creation of a visual novel.	33267
Visual novel release Physical or online release of a visual novel	107961
Character trait A descriptive character trait found in VNDB	3218

HOCHSCHULE
DER MEDIEN

Example research applications

Example research applications with the JVMG knowledge graph

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

HOCHSCHULE
DER MEDIEN

Future directions

Current developments

HOCHSCHULE DER MEDIEN

- Functionality
  - New search assitant for complex queries
  - Data exports as lists and subgraphs
- Data
  - Information from TVTropes.org
  - Importing the latest Media-Arts Database update (new format)
  - Integrated graph the spans all sources
- Services
  - Online documentation with tutorials
  - JVMG „Lab“ workshops in Stuttgart and Kyoto

Future work

HOCHSCHULE DER MEDIEN

- Funding ends in 2026
  - Possible follow-up funding opportunities?
  - Continuation as self-funded academic community effort?

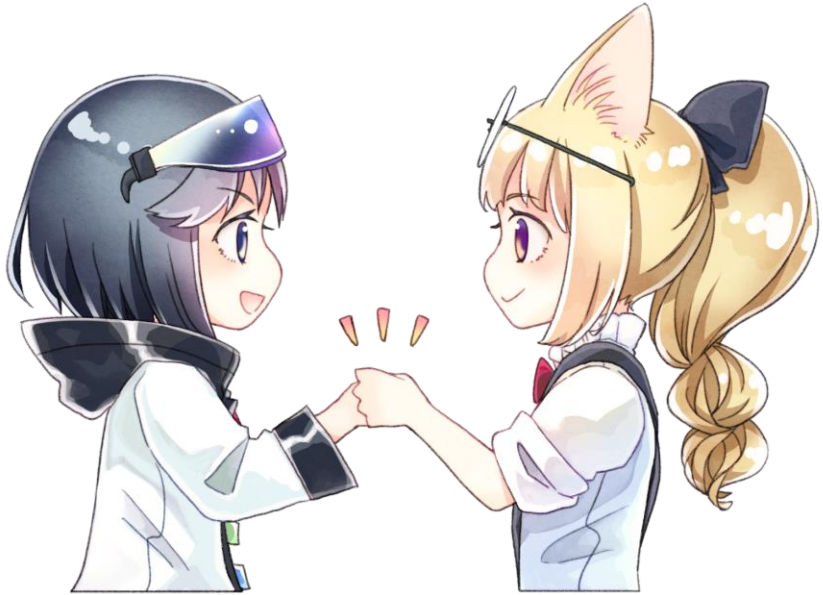
- More data
  - Media types: Music, CD Drama, Light Novels, Games
  - Data sources: More communities

- More details
  - Content descriptions
  - Item-related data

HOCHSCHULE DER MEDIEN

Thank you for your attention!

Get in touch at: kacsuk@hdm-stuttgart.de

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

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