

Enthusiast websites as a source for data on visual media

Research funded by:



Magnus Pfeffer
Core Cultural Metadata Model (CCMM) Workshop
DCMI Conference 2022
05 October 2022

Outline



- Japanese Visual Media Graph (JVMG) project
- Enthusiast communities
- Data quality
- Legal aspects



Introducing the JVMG project

 Databases by fan/enthusiast communities have collected huge amounts of data on Japanese visual media



- Japanese Visual Media Graph (JVMG) project proposal
- <u>Project aim:</u> Make these databases available for large-scale quantitative research, in collaboration with the communities
- 3 year grant funded by the "e-Research Technologies" program of the German Research Foundation





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

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

Database		Wo	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	Anime titles	Anime items	Game items	Manga book series	Manga magazine issues		
Database	12.085	~135.000	~61.000	133.779	170.670		



Assessing data accuracy

- 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.

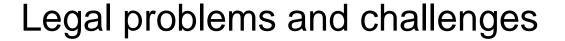
Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
VNDB English title: Visual Novels Sample: 503 Population: 28170	Correct titles	475	94.433%	89.433%	99.433%
	Typographical errors	28	5.567%	0.567%	10.567%
VNDB Original title:	Correct titles	460	91.451%	86.451%	96.451%
Original title: Visual Novels Sample: 503 Population: 28170	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	Correct titles	319	83.727%	78.727%	88.727%
English title: Anime Sample: 381	Misrepresentation errors	37	9.711%	4.711%	14.711%
Population:	Missing data	20	5.249%	0.249%	10.249%
1468	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%



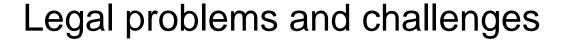
Quality summary

- Enthusiast website data is of very high quality
- Wikidata in comparison is much worse, especially because of missing and incomplete data
- Media-Arts Database shows specific errors that suggest OCR has been used in the data acquisition





- 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



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		
AnimeClick	-	by individual agreement for		
The Visual Novel Database	ODbL	the parts used in each case		
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		



Thank you for your attention!

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

Visit our project website:

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

Visit the JVMG database:

https://mediagraph.link/