Exploring the commonalities and differences in descriptive metadata databases compiled by online fan and enthusiast communities and public administration agencies using the Japanese Visual Media Graph

Research funded by:

DFG Deutsche Forschungsgemeinschaft
German Research Foundation

Zoltan Kacsuk, Magnus Pfeffer, and Martin Roth
FanLIS 2022: Fan futures – beyond the archive
City, University of London, Online, 19-20 May 2022

Outline of the presentation

HOCHSCHULE DER MEDIEN

1. Introducing the **JVMG project** & source databases
2. Data quality: **accuracy**
3. Data quality: **completeness**
4. Comparing **ontologies**
5. Summary: what can we learn about **fan information behaviour** through these results?

Introducing the JVMG project & source databases

HOCHSCHULE DER MEDIEN

Introducing the JVMG project

HOCHSCHULE DER MEDIEN

- Databases by fan 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**
- Funded by the **German Research Foundation**'s (Deutsche Forschungsgemeinschaft) e-Research Technologies program

Source databases

HOCHSCHULE DER MEDIEN

Fan 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

Data quality: accuracy

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: 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
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%

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%

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%

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%

Accuracy main findings

HOCHSCHULE DER MEDIEN

- **Fan 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

Data quality: completeness

Problems with measuring data completeness

HOCHSCHULE DER MEDIEN

- Need to know the **complete population**
  - VNDB aims for completeness in the visual novel field, but there is no alternate source to measure this by
  - For the other databases we decided to look at **anime titles**, BUT there is:
    - No agreement on **what constitutes anime**
    - No agreement on **what constitutes the unit** to be recorded
    - **No actual list** of all anime works/titles
- Need to know the **aims of the database**
  - Not all communities are interested in collecting all titles
  - **Researchers' needs** in relation to completeness are **often different** from the fan communities'

Creating a complete list of anime

HOCHSCHULE DER MEDIEN

- What constitutes anime?
  - At least partially **made in Japan**?
  - **Released in Japan**?
  - **Stylistic features**?
- What constitutes the unit?
- Our list:
  - Work in progress
  - Currently **12.202** individual titles
  - Only anime that was at least partially produced in Japan with official release there

Results in relation to completeness of anime titles

- **Matched** with our list:
  - ACDB: ~3210 titles
  - AnimeClick: ~7104 titles
- **Not matched** with our list:
  - Wikidata: ~4467 entities
  - Media-Arts Database: ~12.085 entities (includes non-anime as well)
- Other results:
  - **Deprecated titles**: ACDB & AnimeClick both 8 titles
  - Deprecated titles cannot appear in Media-Arts Database
  - **Chinese, Korean and global anime** in the fan databases and Wikidata

HOCHSCHULE
DER MEDIEN

Comparing ontologies

Core entities

- Described with formal objective attributes (usually ‘literals’)
- Very limited number of core entities in the domain:
  - **Media**
  - **People** involved in the creation of the media
  - **Companies** involved in creation/publication/dissemination of media
  - **Visual characters** (every community agrees that highlighting characters is important!)
- In some other media centered domains treating protagonists in such a key way is not necessarily the case
  - E.g. IMDB does not have characters separately

Core concepts

HOCHSCHULE DER MEDIEN

- To describe media:
  - Genre
  - Tags
- To describe characters:
  - Traits
  - Tags
- Different approaches:
  - **VNDB:** very detailed closed ontology of hierarchical traits/tags
  - **ACDB:** closed ontology only for core traits plus free-form tagging
  - **AnimeClick:** curated list of genres, closed non-hierarchical list of tags

Entity and concept numbers for fan databases

HOCHSCHULE DER MEDIEN

Fan 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 for non-fan databases

HOCHSCHULE DER MEDIEN

Fan community	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		

Relationships

- Connections between core entities and core concepts
  - Between two entities
    - E.g. character appears in media
  - Between two concepts
    - Hierarchical relationships between concepts
  - Between an entity and a concept
    - E.g. Character has certain personality trait
- Connections **between two entities** or between **two concepts** is usually **based on objective observation**
- But, connections **between an entity and a concept** is often a **subjective decision**
  - Cannot be validated
  - Tools in place to create a convergence towards a consensus
    - E.g. VNDB uses numbers and font size to display community consensus for tags

The object-work-franchise relationship

HOCHSCHULE DER MEDIEN

- **Three layers:** physical object – abstract work – franchise
- **VNDB:** physical media to abstract work (bottom two layers)
- **AnimeClick and ACDB:** Clustering media into media franchises (top two layers)
- **Media-Arts DB:** bottom up from physical objects towards franchise (franchise still work in progress)
- **Wikidata:** Very fragmented (no agreement on a common ontology)

Literals to entities

HOCHSCHULE DER MEDIEN

- **Wikidata** strive to remove literals
  - E.g. Kyoto is an entity, whereas in the fan databases Kyoto is always a literal
- Turning a literal into an entity is a lot of work
- Fan databases create entities of the concepts, but will not create separate entities for geographical entities for example
  - Taking the effort to turn concepts into entities is what enables a large part of the work the JVMG project is premised to support
- A smaller community might not create entities for people, but as the database grows they shift towards entity creation
- Hypothesis: **What gets turned into entities is driven by what needs to be searched for**, thus the development of the databases is driven by the search needs of the communities

Summary: what can we learn about fan information behaviour through these results?

Relating our examination to previous work on fan information behaviour

- Similar to Price (2019) and Price & Robinson (2021):
  - We **compare** fan databases among each other as well as with non-fan databases
  - As in the case of tag network analysis, we rely on the **data itself** to draw conclusions
- Previous work often focusing on fanfiction sites
  - We compare databases that serve **other types of information needs**

Price, L. (2019). Fandom, Folksonomies and Creativity: the case of the Archive of Our Own. *The Human Position in an Artificial World: Creativity, Ethics and AI in Knowledge Organization* (pp. 11–37). Ergon Verlag.

Price, L. and Robinson, L. (2021). Tag analysis as a tool for investigating information behaviour: comparing fan-tagging on Tumblr, Archive of Our Own and Etsy. *Journal of Documentation*, 77(2), pp. 320–358.

Results in relation to fan information behavior

- Work on data quality helps highlight the accuracy of these databases
  - **Dedication of the fans** is clearly visible -> especially in comparison with other sites that work with the same information
- Work on data completeness highlights that online fan communities have varying approaches to:
  - What should be recorded in their databases
  - What layers of the object-work-franchise hierarchy they focus on
  - And these don’t necessarily overlap with the needs of researchers working on the domain

Results in relation to fan information behavior

- **Fan databases are similar** to each other and different from other databases
  - They are driven by similar needs (finding further works) in relation to the description of the media
  - Wikidata is different -> they want to entitify the whole world
  - Media-Arts Database doesn’t care about characters, plus all domains are completely separate (agglomeration of data and not integration)
- **Flexibility and structure** both present in the fan databases’ use of concepts
  - Similarities with Bullard’s (2018) ‘curated folksonomies’
  - Conforming to the results of Johnson (2014), Price (2019) and Price & Robinson (2021)
- **Centrality of characters** in the domain of popular Japanese visual media
  - Corroborates Azuma’s claim from *Otaku: Japan’s Database Animals* (2009 [2001])

Bullard, J. (2018). Curated Folksonomies: Three Implementations of Structure Through Human Judgment. *Knowledge Organization*, 45(8): 643-652

Johnson, S. F. (2014). Fan fiction metadata creation and utilization within fan fiction archives: Three primary models. *Transformative works and cultures*, 17(1).

HOCHSCHULE DER MEDIEN

Thank you for your attention!

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

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

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