Exploring Data Quality in the JVMG project

Research funded by:

Deutsche Forschungsgemeinschaft
German Research Foundation

Zoltan Kacsuk

Phase One Wrap-Up Workshop of the JVMG project

HdM, Stuttgart, 27-28 January 2023

Outline of the presentation

HOCHSCHULE DER MEDIEN

1. Data quality dimensions
2. Data quality: **accuracy**
3. Data quality: **completeness**

HOCHSCHULE
DER MEDIEN

Data quality dimensions

Data quality dimensions

HOCHSCHULE DER MEDIEN

- Eurostat data quality dimensions:
  - relevance
  - **accuracy**
  - timeliness
  - punctuality
  - accessibility
  - comparability
  - coherence
- Further possible dimensions:
  - **completeness**
  - (reliability)
  - ...

Data quality: accuracy

Assessing data accuracy

HOCHSCHULE DER MEDIEN

- **Random sample** of anime (or visual novel) titles
  - Checked **titles** as those are among the easiest elements to find ground truth for
  - There are many data elements that **cannot be objectively assessed** for accuracy
- **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
ACDB English title: Anime Sample: 424 Population: 2435	Correct titles	312	73.585%	68.585%	78.585%
	Typographical errors	111	26.179%	21.179%	31.179%
	Misrepresentation errors	1	0.236%	0.041%	5.236%
ACDB Japanese title: Anime Sample: 424 Population: 2435	Correct titles	345	81.368%	76.368%	86.368%
	Typographical errors	77	18.160%	13.160%	23.160%
	Misrepresentation errors	2	0.472%	0.082%	5.472%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
ACDB English title: Anime Sample: 424 Population: 2435	Correct titles	312	73.585%	68.585%	78.585%
	Typographical errors	111	26.179%	21.179%	31.179%
	Misrepresentation errors	1	0.236%	0.041%	5.236%
ACDB Japanese title: Anime Sample: 424 Population: 2435	Correct titles	345	81.368%	76.368%	86.368%
	Typographical errors	77	18.160%	13.160%	23.160%
	Misrepresentation errors	2	0.472%	0.082%	5.472%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
AnimeClick English title: Anime Sample: 483 Population: 9167	Correct titles	333	68.944%	63.944%	73.944%
	Typographical errors	136	28.157%	23.157%	33.157%
	Misrepresentation errors	12	2.484%	0.131%	7.484%
	Missing data	2	0.414%	0.022%	5.414%
AnimeClick Japanese title: Anime Sample: 483 Population: 9167	Correct titles	367	75.983%	70.983%	80.983%
	Typographical errors	88	18.219%	13.219%	23.219%
	Misrepresentation errors	8	1.656%	0.087%	6.656%
	Missing data	20	4.141%	0.218%	9.141%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
AnimeClick English title: Anime Sample: 483 Population: 9167	Correct titles	333	68.944%	63.944%	73.944%
	Typographical errors	136	28.157%	23.157%	33.157%
	Misrepresentation errors	12	2.484%	0.131%	7.484%
	Missing data	2	0.414%	0.022%	5.414%
AnimeClick Japanese title: Anime Sample: 483 Population: 9167	Correct titles	367	75.983%	70.983%	80.983%
	Typographical errors	88	18.219%	13.219%	23.219%
	Misrepresentation errors	8	1.656%	0.087%	6.656%
	Missing data	20	4.141%	0.218%	9.141%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
AnimeClick English title: Anime Sample: 483 Population: 9167	Correct titles	333	68.944%	63.944%	73.944%
	Typographical errors	136	28.157%	23.157%	33.157%
	Misrepresentation errors	12	2.484%	0.131%	7.484%
	Missing data	2	0.414%	0.022%	5.414%
AnimeClick Japanese title: Anime Sample: 483 Population: 9167	Correct titles	367	75.983%	70.983%	80.983%
	Typographical errors	88	18.219%	13.219%	23.219%
	Misrepresentation errors	8	1.656%	0.087%	6.656%
	Missing data	20	4.141%	0.218%	9.141%

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%

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
Media Arts Database Japanese title: Anime Sample: 381 Population: 1468	Correct titles	369	75.306%	70.306%	80.306%
	Typographical errors	12	2.449%	0.099%	7.449%
	Misrepresentation errors	94	19.184%	14.184%	24.184%
	Cannot be determined	15	3.061%	0.124%	8.061%

Duplicate entry: 1
Not anime: 13

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
Media Arts Database Japanese title: Anime Sample: 381 Population: 1468	Correct titles	369	75.306%	70.306%	80.306%
	Typographical errors	12	2.449%	0.099%	7.449%
	Misrepresentation errors	94	19.184%	14.184%	24.184%
	Cannot be determined	15	3.061%	0.124%	8.061%

Disambiguation: 58

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
Media Arts Database Japanese title: Anime Sample: 381 Population: 1468	Correct titles	369	75.306%	70.306%	80.306%
	Typographical errors	12	2.449%	0.099%	7.449%
	Misrepresentation errors	94	19.184%	14.184%	24.184%
	Cannot be determined	15	3.061%	0.124%	8.061%

Title in MADB: ガラスの花と壊す世界 Vitreous Flower & Destroy the World D.backup

Probably mostly ~~OCR~~ errors

Data checked	Decision	Count	Percentage	CI lower bound	CI upper bound
Media Arts Database Japanese title: Anime Sample: 381 Population: 1468	Correct titles	369	75.306%	70.306%	80.306%
	Typographical errors	12	2.449%	0.099%	7.449%
	Misrepresentation errors	94	19.184%	14.184%	24.184%
	Cannot be determined	15	3.061%	0.124%	8.061%

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:
  - Data **recording protocols** of the MADB
  - Disambiguation information **added to titles**
  - Disambiguation information **in different format** than the original
  - **Errors in the sources** the data is taken from

HOCHSCHULE DER MEDIEN

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:
  - 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 enthusiast 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

HOCHSCHULE DER MEDIEN

- Matched with our list:
  - ACDB: ~3210 titles
  - AnimeClick: ~7104 titles
- Not matched with our list:
  - Wikidata: ~1515 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 the Media Arts Database
  - Chinese, Korean and global anime in the enthusiast databases and Wikidata
  - Western animation titles in the Media Arts Database

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/