Japanese Visual Media Graph: Project Update

Prof. Magnus Pfeffer, Stuttgart Media University
<pfeffer@hdm-stuttgart.de>

Overview

- Project starting point
- Lesssons and results from project phase 1 (2019-2022)
- Project phase 2 (2023-2025)

Project starting point

Observations

- “Manga, Anime and Games” are a growing field of study
  - Interest from Japanese and Media Studies
  - Need for data-driven research options
- Lack of comprehensive ressources
  - Rapidly growing number of media
  - Library catalogs and similar publisher-sourced data
  - Small encyclopedic efforts
- Vibrant enthusiast communities
  - Active sind the late 1990s
  - Collect and curate information on Japanese Visual Media

Main project goals

- Create a research database on Japanese visual media, including, but not limited to, anime, manga, computer games and visual novels
- Aimed at researchers in Japan Studies who focus on modern media and its expressions, themes, topics, characters and reception
- Use a graph-based, highly interconnected database structure, similar to the Google knowledge graph, that is combined with a flexible search interface and analytic tools
- Use the data on Japanese visual media that is being created and curated by the many enthusiast communities on the web

Lesssons and results from project phase 1 (2019-2022)

Working with enthusiast communities

- Can be hard to contact
  - Emails not answered
  - Non-standard communication (IRC, Discord, Social Media)
- Need time to establish a connection
  - Are we even real? Or a troll on the internet?
  - Need to explain our academic motivations
- Very supportive once personal contact was made
  - In-person workshops with researchers were helpful

Data sources

- The visual novel database (vndb.net)
  - International community
  - Focussed on the visual novel game genre
- AnimeClick (animeclick.it)
  - Italian community
  - Broad interest in Japanese media and culture
- Anime Characters Database (animecharactersdatabase.com)
  - International community
  - Focussed on character appearance and personality
- Media-Arts Database
  - Japanese project, utilizing publisher data

Data licensing

- Heterogenous licensing by the communities
  - Low awareness of copyright in database context
  - Non-standard license agreements and site policies
  - Incompatible licenses
- Legitimate concerns of the communities
  - Wholesale copying of sites to divert traffic and profit by advertisements
  - Recognition of their work by proper attribution data sources
- License must cover all research use cases
  - Combining with other data, publication of analysis and research
- Solution: CC-BY-SA-NC 4.0, dual licensing where needed

Data quality

- Extensive analysis of data accuracy
  - Based on title information (Japanese and English) of anime works
  - Comparison to “ground truth”: Media photos, OP title cards, publisher page
  - Sample size large enough to be in a 5% margin at 95% probability
- Results very encouraging
  - Very low error rate in enthusiast communities’ data
  - Most common type of error is typographical inconsistencies
  - Disambiguation information stored in “original title” field
- Wikidata is much worse in comparison
  - up to 10% misrepresentation errors, up to 16% missing data

Matching and merging data: general process

- 1. Merge works based on title information (in Japanese kanji)
  - Use very strict string comparison first
  - Relax string comparison to address typographic inconsistencies next
  - Resolve ambiguous matches manually
- 2. Compare information on creators per work cluster
  - Match candidates can be found with relaxed name matching
- 3. Use creator clusters to find new candidates for matching works
- 4. Use n-gram comparison for remaining unmatched works
  - Manual resolution required

Matching and merging data: current state

- Done
  - Anime and manga work clusters using strict and relaxed matching
  - Manual resolution of ambiguous matches of work titles of anime
  - Matching between multiple sources
- Ongoing
  - Creator clusters using Japanese names
  - Problem: creator data in media-arts database not normalized
- Future
  - Visual character clusters

Small use cases for research

- Idea
  - Showcase research using the knowledge graph
  - Limited scope to reach results more quickly
  - Used to direct the development in the project

- Examples
  - Census of characters in Japanese visual media
  - Testing a point of Hiroki Azuma's "Otaku: Japan's Database Animals"
  - Recurring patterns in character creation in visual novels

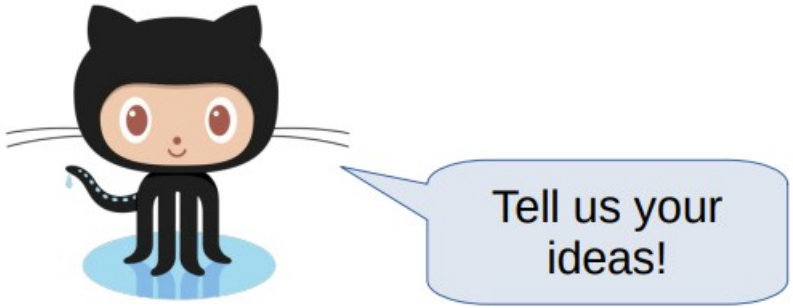
Software

- Testing of different approaches
  - RDF triple store with seperate web frontend and external search index
  - Integrated RDF/linked data hosting solution
  - Integrated graph database solution
- Performance issues
  - Viewing large result sets in web frontend
- Licensing issues
  - Proprietary licenses with unusual restrictions

Software

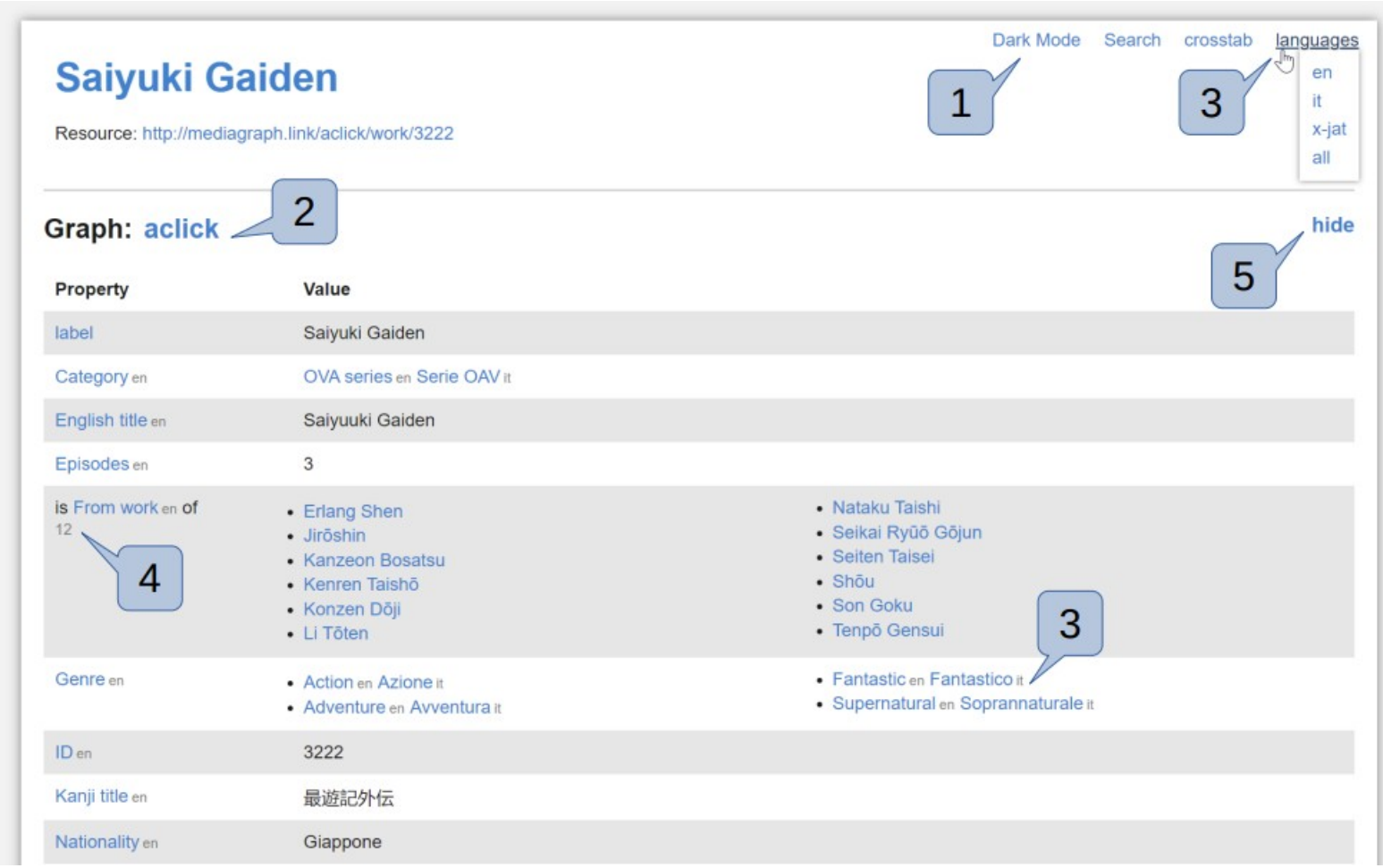
- Integrated solutions not a good fit for our data and use cases

- Own solution
  - Apache Fuseki RDF triple store with SPARQL interface
  - Custom web frontend with optimized queries for large results
  - Elasticsearch index for search function
  - Open source



https://github.com/Japanese-Visual-Media-Graph

Frontend features



1: CSS selection

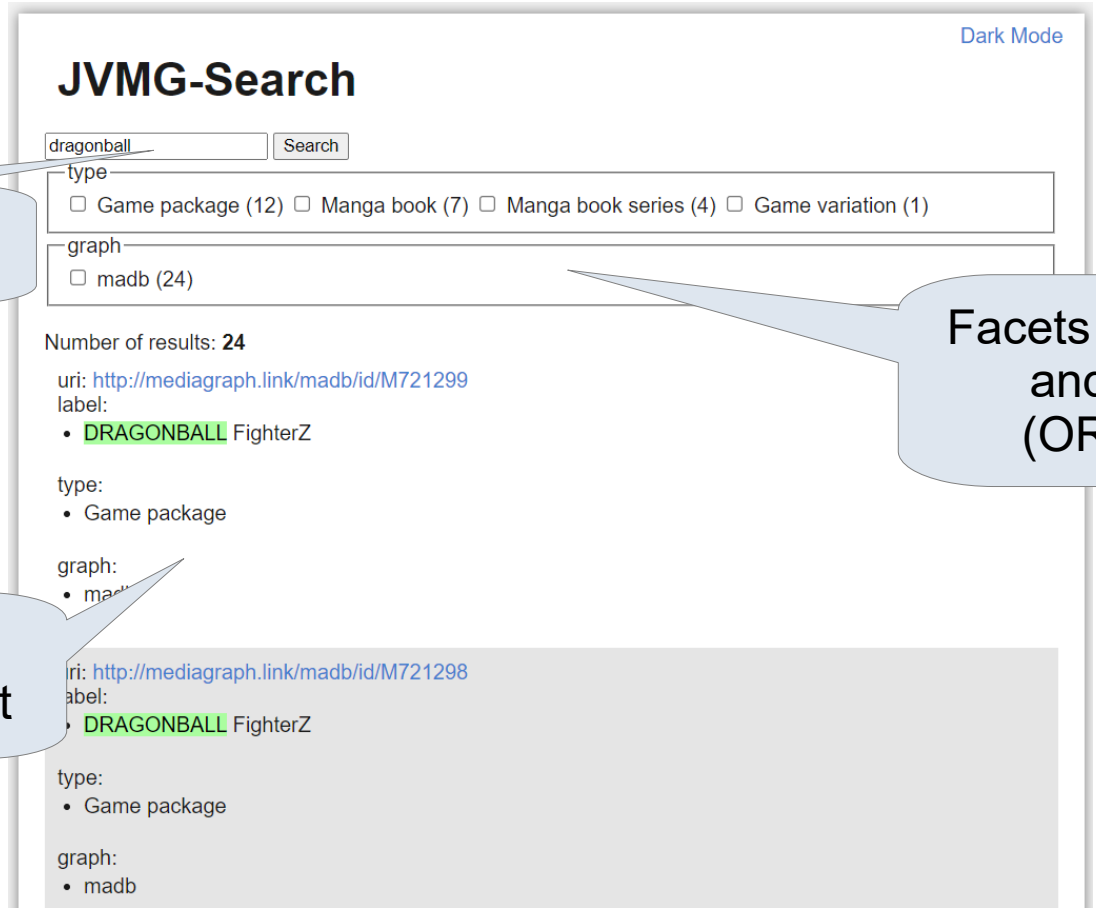
2: Source indicator

3: Language filter

4: Expandable sets

5: Source filter

Search function



Search all indexes

Facets for data source and entity type (OR-combined)

Result list with match highlight

Project phase 2 (2023-2025)

New partners for research

- Federico Pianzola (University of Groningen)
  - Fan works
- Martin Hennig (University of Tübingen)
  - Media genres
- Bryan Hartzheim (Waseda University Tokyo), Stevie Suan (Hosei University Tokyo)
  - Creator careers

A locally running software stack

- Server mediagraph.link has limited resources
- Local copy of database useful for larger analysis tasks

- Goal
  - Local copy using container or other technology
  - Easy to install and run on any operating system (Win, Mac, Linux)
  - Integration into local research environment (Jupyter Labs)

Materials and support for researchers

- Documentation
  - Complete documentation of all ontologies
  - Video guides and online tutorials for common tasks
  - Example complex analysis tasks using Python and Jupyter Labs

- Workshops
  - Single-day workshops at conferences
  - Multi-day interactive workshops at locations Stuttgart and Kyoto
    - Open to all interested researchers

Thank you for your attention.