Japanese Visual Media Graph: Providing researchers with data from enthusiast communities

Magnus Pfeffer
Stuttgart Media University, Germany
pfeffer@hdm-stuttgart.de

Martin Roth
Leipzig University
Germany
martin.roth@uni-leipzig.de

Abstract

The "Japanese Visual Media Graph" project aims to create a research database on Japanese visual media, including, but not limited to anime, manga, computer games and visual novels. It is aimed at researchers in Japan Studies who focus on modern media and its expressions, themes, topics, characters and reception. We envision a graph-based, highly interconnected database structure, similar to the Google knowledge graph, that is combined with a flexible search interface and analytic tools. We intend to use the data on Japanese visual media that is being created and curated by the many enthusiast communities on the web. An initial survey of several larger community websites revealed a high level of information granularity, resulting from a deep understanding of the source material and general enthusiasm for recording data by the volunteer contributors. As such, making contact with these communities and learning about their needs and motivations is one of the main project elements. We intend to engage in a meaningful discussion with representatives and administrators of the community sites in order to establish long-term cooperation that benefits both sides.

Keywords: enthusiast communities, Japanese visual media, graph-based database, metadata, ontologies

1. Introduction

With the shift to digital distribution in the past decades, it has become easier to get access to a variety of visual media. Even previously niche media and contents now have a potentially global audience, resulting in the formation of enthusiast communities that develop online platforms to discuss, analyse, catalogue and organize information on "their" kind of visual media. Some focus on a certain genre, others on a country of origin or the work of a certain author. The sheer amount of digital data on visual media which is produced daily is staggering in itself, but even more so is the level of sophistication some of these communities achieve with regards to detail, data modelling, and coordination. Using state-of-the-art technology and large-scale collaboration tools, their many voluntary contributors have created a wealth of detailed, high-quality data on visual media.

Well-known examples are the Internet Movie Database (www.imdb.com) or the Online TV Database (www.thetvdb.com), which both focus on mainstream international movie and TV programs with a very broad appeal. Consequently, their corresponding communities are large. They managed to grow continually and their online services are both well established and financially secure, as they are a highly sought platform for online advertisements. Communities focusing on media with a smaller audience have access to the same technology and collaboration tools and, likewise, create information collections of surprisingly high quality – especially since many of their community members are personally attached to their respective field of interest and its contents. Through their activities, they develop expert perspectives on visual media and gain knowledge far beyond specific contents.

The data produced by these communities is invaluable for researchers working on visual media, as it provides much-needed contextualization and relates different works, genres, and media

types. At the same time, it provides a perspective on reception and evaluation by the various audiences of visual media and offers insight into local and global media cultures.

For researchers, however, using this data involves many challenges. Firstly, it is difficult to evaluate the existing communities with regards to their workflows and data quality standards. Secondly, depending on the research question, it may be necessary to use multiple data sources, which further complicates the issue, as the researcher would have to become a data scientist and data modeller in order to do so. Thirdly, such communities are fragile and lack substantial resources beyond voluntary contributions. They often rely heavily on the willingness of their community members and users to donate money to keep servers running and to offer their software development skills to improve services. As a consequence, some of the community websites are difficult to use, offering limited search options or clunky interfaces. More importantly, we observe a lack of persistence and a considerable risk that the community site might run out of funding and simply disappear.

In this project, we aim to develop technological solutions to make the data accumulated by enthusiast communities available to researchers, and to contribute to preserving the efforts many community members have made over the past years. We will develop tools, workflows and best practice solutions for discovering, extracting, harvesting, consolidating and linking data from multiple sources. The accumulated data will be made available in a central repository and we will provide a search interface that caters to the needs of researchers in the field.

While we consider most of these tools and workflows to be universal in application, we will limit the prototypical development to the domain of contemporary Japanese visual media, focusing on anime, manga and video games, which have become a mainstay of international popular culture and a pillar of Japan's recent soft power strategy, often referred to as "cool Japan" (McGray, 2002; Oyama, 2016; Valaskivi, 2013). This field offers a particularly rich and challenging subject for our project, with which we can, in turn, contribute substantially to ongoing research efforts in the scientific community. The creative industry in Japan displays a strong tendency towards extensive story worlds and character-centred cross-media franchising, sometimes referred to as "media-mix" (Steinberg, 2012; Picard & Pelletier-Gagnon, 2015; Nozawa, 2013), as well as a close relation to fan creations and fan activities (Condry, 2013). The complex, vast network of elements and works emerging from these practices necessitates an approach which is able to draw detailed connections between various contents, themes, and characters across multiple media, content layers and data sources. In addition, a significant share of information and content is not available outside of Japan, which makes gathering and validating information about respective subjects especially difficult. What initially sparked our interest and motivated us to develop this research project was that the existing community-based data sources reflect this complexity and scale. Japanese visual media are the main subject of some of the most active communities on the web, and these communities display an astonishing level of detail, care, and theoretical and technical knowledge. In this project, we aim to explore possibilities and practical solutions to preserving their efforts and making the vast amount of data on visual media in Japan available to researchers. This means entering into dialogue with the communities, as well as data scientists and researchers focusing on Japan. Our project aims at developing respective strategies and starting a productive dialogue between all parties involved. Our emphasis on the community perspective implies that we also need to reflect on the vague notion of "Japaneseness" that serves as common identifier for the community members. This identifier defines the boundaries of our data, regardless of other, i.e. academic definitions of "Japan" (Steinberg & Zahlten, 2017; Consalvo, 2016). At the same time, the project makes the negotiations and discussions of such boundaries within enthusiast communities available for further analytic considerations. Given that enthusiast communities have diverse structures and aims, which differ from the scientific community, one focus of our project is to start a productive dialogue with such communities and develop best practice solutions for collaborating with them.

1.1 Areas of particular interest

Based on the discussions amongst the applicants and with the help of a series of qualitative interviews with researchers of contemporary visual media, we have identified several major areas of interest, which provide welcome challenges for our project. These interviews were conducted either in person at conferences or at workshop meetings, or via email or telephone. We explained the idea of a centralized database that aggregates data from multiple enthusiast web communities and makes it available for research and analysis through an interface that is developed with researchers in mind.

A vast and complex field of study beyond individual capacities

Contemporary Japanese visual media are characterized by vast universes around specific themes, which develop through an interplay of various media over time. Respective media have also become a focus of research in Japanese studies, with an increasing number of publications, journals and conferences dedicated to its topics. Two such examples would be Mechademia (https://www.mechademia.net) and the Mutual Images Journal (http://www.mutualimages-journal.org). However, as published visual media grows at a non-linear rate, so too does the interconnectedness of said media. With many larger media franchises spanning different media types and a confusing amount of sub-series and related works, researchers struggle to stay informed about their source material. An integrated data repository can allow researchers to trace particular elements or subjects across different media and in their temporal development.

Lack of usable, persistent, and comprehensive resources

While there are a few encyclopaedic efforts, these are limited in scope, focusing on a singular type of media, and rarely updated (Clements & McCarthy, 2006; Gravett, 2004). Others are commercial in nature and seldom possess the rigour necessary for academic work. No exhaustive bibliographic resources exist, and library catalogues focus only on small local collections; also, their bibliographic descriptions are not suited for searching the media contents. In addition, most of our respondents knew of only a few enthusiast websites and generally considered the cost of assuring the quality of the contents too high in comparison with the potential benefits. The clunky search interfaces and the risk of the websites disappearing without adequate backup are additional obstacles for researchers. An integrated and persistent data repository with documented high-quality standards and a user-friendly interface can reduce the cost for researchers significantly.

Strong need for more research options

A central database of short bibliographic descriptions of all entities published in Japan exists for some media, and it can be considered as a first major step in the right direction (Kiryakos & Sugimoto, 2016). But instead of searching for individual works by title or author, researchers would rather explore the interconnected media by genres, themes, tropes or the contributing persons. Some of our interviewees went further; they pointed out that as most visual media is driven by its unique characters. A database which not only includes the media and a description of its contents, but also individual characters, would open up completely new venues of research. Researchers have taken note of the methods that are employed in other media-related fields and long for tools to apply them to their own interests, like network analysis in literature (Moretti, 2011). Most interviews resulted in possible research questions that might be answered with the database and most respondents urged us to make a prototype of the database available as soon as possible so they could test it and give early feedback. When asked, most were more than willing to participate in the development in workshops or similar events. Against this background, our project aims to develop a broad spectrum of showcase uses and best practice approaches to the data in close collaboration with the research community.

1.2 Overview of enthusiast communities

In order to better assess the available resources, their contents, communities and data structures, we conducted a survey of enthusiast websites in the domain of Japanese visual media. This survey was done as part of an international seminar at Stuttgart Media University, and the students collected information on more than 40 different sites, from smaller niche communities to large databases with enormous amounts of information. To give a short impression of the kind of site and their contents we describe a few examples:

- **anidb.net** is an English-language fan effort to catalogue all anime that have ever been produced for TV, direct-to-video, movie theatres or web streams. It contains information about more than 10.000 individual works and records information on the level of the individual episode. It also contains information on the related persons and their individual role/contribution as well as a large thesaurus to describe the media content. Information on individual characters is also present, such as who are protagonists or side characters in the works. One of the main strengths of anidb.net, besides its size, is the number of descriptive relations documented between works.
- **myanimelist.net**, despite its naming, covers both manga and anime. It is also an English-language site and while it does not offer as many attributes and related persons for the individual entry in comparison to anidb.net, it carefully documents the interrelatedness between manga and anime adaptions of the same work. Thus it can be a valuable resource which can be used to "bridge" more specialised resources focusing on individual media types.
- **vndb.org**, the visual novel database, focuses solely on this subtype of computer games. Visual novels typically feature lengthy text-based storylines and very limited player interaction, which can lead to different effects on story development. This focus allowed the community of over 100.000 users to collect data and descriptions of over 20.000 different software titles with more than 65.000 characters. It denotes both Japanese and international releases, offers a rich tagging system to describe the contents and has a large forum for user interaction and discussion. As visual novels are often the starting point of larger franchises, having access to information on this kind of media will be helpful for many research questions.
- **animexx.de**, the website of the largest anime and manga association in Germany, hosts an extensive database with metadata on manga, anime, and other visual media, as well as related fan works and cosplay.

2. Project Objectives

The project has four distinct objectives, covering the areas of community interaction, data model and database usage. Specifically, they include:

Data sharing agreements with enthusiast communities

In order to have a successful partnership, it is paramount to learn more about the motivation and needs of the enthusiast communities, and the challenges they face. We have already started exchanging ideas and meeting with communities. This process, even though at an early stage, has helped revealing the very different approaches to data accumulation, structuring and, most importantly, to sharing employed by the individual communities: while some block any non-human access by web crawlers and other automatic harvesting programs, others do not regulate the access, with two sites even have a sophisticated API to access raw data. Licensing information is sometimes missing, incomplete or in one case contradictory. These insights are exciting all the more, as there is little precedent or best practices available for collaborating with fan communities in the context of data.

Our goal is to come to an agreement with the broader Japanese visual media enthusiast community in regards to licenses and conditions of data sharing. Importantly, this means respecting the wishes of the volunteers that worked together to create the data and identifying

aspects of our proposed central data repository that can be of immediate or mid-term benefit to the communities. We will seek to create a win-win situation for both parties and establish continuous cooperation with each community.

Creating an adaptable data model

In order to ingest all kinds of data and provide rich search functions, the data model needs to accommodate the media itself, the entities involved in its creation, and the contents and relationships between the individual works and their elements. While existing data can be used as a reference, not all aspects have been modelled in any database so far. The combined model will be complex and will not be described as a traditional metadata schema, but as a data graph containing typed entities with a pre-described set of attributes and relations between these entities. Rules will describe the type of relation (1:1, 1:many, many:many) as well as the valid types of entities that can be a member of the relation. Such a graph-based model is both easily adaptable and very expressive, so it will be possible to describe and link all aspects of Japanese visual media. The data model will be mapped onto an ontology that can be expressed in RDF, so that all elements of the database can be made available as Linked Open Data.

Creating a single, searchable research database hub

In order to provide researchers with the means to search and analyse all data available in various community databases, we will collect and integrate respective data sources into one searchable database hub. Adapting graph data to a traditional database schema often results in information loss or performance problems in storing and retrieval. As native graph databases have become more mature in the past years, it is viable for our project to store the data directly into one such database. The selection of an adequate open-source software for this task as well as the configuration and - if necessary – the extension of the web-based search interface is part of this objective. There is also a need to formulate workflows and repeatable processes that ingest data from the multitude of enthusiast sites into the central database. Instead of developing new tools, the focus here is on reusing existing, open-source tools for web harvesting or database ingestion. All workflows should be documented and well described so they can be run at the site of the university library.

Evaluation of the data quality, model, and interface

Building an infrastructure for research should not be driven by the necessities of information technology, but by the needs of researchers. In order to determine the success of our work, we aim to conduct a series of small-scale research projects in our project team and actively encourage beta tests by external researchers. In an initial step, a member of the project team will formulate requirements and define search strategies for representative research questions, which will then serve as test cases for the developed prototypes. We will establish a tight feedback loop between researcher and data scientist in each step of data modelling, data integration and search interface development to guide the development efforts. In the second step, prototypes of the database and search interface will be made open to beta testers who will also test their individual research scenarios and give feedback to the project team. In this, we follow the request of the interested external researchers interviewed in preparation of this project.

3. Project Partners and Timeline

The project is conducted by an interdisciplinary research team comprised of members from the fields of information science, media science and Japan studies, located at the partner institutions Stuttgart Media University and Leipzig University Library. The project has started on May 1st, 2019, with an intended duration of 36 months. Important milestones include a workshop with members from different enthusiast communities, which was successfully held in early July 2019, a working prototype that is available to interested researchers after 18 months into the project,

and a larger workshop with both researchers that used the prototype at the end of the project duration.

All code developed for any part of the media graph database will be published with an open-source licence on GitHub. A project blog is available on jvmg.iuk.hdm-stuttgart.de with news and updates.

Acknowledgements

The Japanese Visual Media Graph project is funded by the Deutsche Forschungsgemeinschaft (DFG) in the Scientific Library Services and Information Systems (LIS) / E-Research Technologies programme.

References

Clements, J., & McCarthy, H. (2006). The Anime Encyclopedia: A Guide to Japanese Animation Since 1917 (3rd edition). Berkeley, Calif.: Stone Bridge Press.

Condry, I. (2013). The soul of anime: Collaborative creativity and Japan's media success story. Durham: Duke University Press.

Consalvo, M. (2016). Atari to Zelda: Japan's Videogames in Global Contexts. Cambridge, Mass.: MIT Press.

Gravett, P. (2004). Manga: 60 Years of Japanese Comics. London: Laurence King Publishing.

Kiryakos, S., & Sugimoto, S. (2016). Aggregating manga metadata across institutions: lessons learned in the application of EDM. In IConference 2016 Proceedings (p. 3). iSchools.

McGray, D. (2002). Japan's Gross National Cool. Foreign Policy, (130), 44–54.

Moretti, F. (2011). Network theory, plot analysis. Literary Lab Pamphlet 2. Stanford University Literary Lab. Retrieved from https://litlab.stanford.edu/LiteraryLabPamphlet2.pdf

Nozawa, S. (2013). Characterization. Semiotic Review, (3). Retrieved from https://www.semioticreview.com/ojs/index.php/sr/article/view/16

Oyama, S. (2016). Japanese creative industries in globalization. In L. Hjorth & O. Khoo (Eds.), Routledge Handbook of New Media in Asia (pp. 322–332). Abingdon Oxon UK: Routledge.

Picard, M., & Pelletier-Gagnon, J. (2015). Introduction: Geemu, media mix, and the state of Japanese video game studies. Kinephanos: Journal of Media Studies and Popular Culture, 5, 1–19.

Steinberg, M. (2012). Anime's media mix: Franchising toys and characters in Japan. Minneapolis: University of Minnesota Press.

Steinberg, M., & Zahlten, A. (2017). Introduction. In M. Steinberg & A. Zahlten (Eds.), Media Theory in Japan. Durham: Duke University Press.

Valaskivi, K. (2013). A brand new future? Cool Japan and the social imaginary of the branded nation. Japan Forum, 25(4), 485–504.