Exploring the research utility of fan-created data in the Japanese visual media domain *

Senan Kiryakos[0000−0001−5980−4511] and Magnus Pfeffer[0000−0002−2141−6721]

Stuttgart Media University, Stuttgart, Nobelstraße 10, 70569 Stuttgart, Germany
{kiryakos,pfeffer}@hdm-stuttgart.de

Abstract. Researchers wishing to study the Japanese visual media domain do not currently have access to a large set of descriptive data for analysis. To remedy this, the Japanese Visual Media Graph project (JVMG) seeks to build a knowledge graph consisting of descriptive metadata sourced from various online fan communities, described using RDF ontologies. To better understand how this informal, crowdsourced data can be both described using a formal ontologies, and made useful to researchers, this paper presents a summary of the properties of community data from a number of fan sites, and discusses the impact it has on the creation of a unified dataset and on possible research use cases. We find that the data sources are of high quality and coverage. They complement each other well and our central ontology should enable these connections between related resources across communities. Certain niche research topics are enabled by the use of community created data alone, but others encourage the incorporation of additional authoritative sources.

Keywords: Ontologies · Knowledge Graphs · Visual Media · Fan Communities · Data Integration.

1 Introduction

Japanese visual media, such as animation, manga, and video games,is a topic of interest for researchers of a variety of disciplines. As global interest in Japanese visual media has grown commercially, so too has academic interest[1]. This interest ranges from studying the domain broadly[2], to research on fandom interaction with the domain online[3]. This latter group includes studies on the utility of user-generated content[4], and intersects with more general studies on fandom and participatory culture[5, 6]. Those wishing to conduct data-driven research on topics such as these are faced with the fact that there is no central database on the domain and authoritative data, e.g. from libraries, is limited in detail and coverage. On the other hand there are a lot of fan communities that create and curate a significant amount of unique and granular information on the domain of Japanese visual media. But these communities are focused on different aspects of

* This work has been funded by Deutsche Forschungsgemeinschaft (German Research Foundation) with a grant from the funding program e-Research Technologies.

the domain and while data across these communities often deals with the same resources, the fact is that these draw from different data sources, use distinct data models and vocabularies and describe the resources at different granularity levels. This results in a level of heterogeneity that can be an obstacle for researchers wishing to analyze the domain broadly, or who desire as much data as possible. The Japanese Visual Media Graph project (JVMG)[1] seeks to address these issues through the creation of a central knowledge graph which combines data from multiple community providers and a unified, RDF-based ontology.

Foundational to the creation of this knowledge graph is the collection and analysis of descriptive data from a number of online fan communities. This not only serves as a necessary step in the technical development, but also functions as a type of limited domain analysis, resulting in a unified ontology that better reflects the ways that different communities understand and describe their data. Additionally, this allows for fundamental aspects and limitations of the domain to be better communicated to researchers, such as levels of data quality and authority, descriptive granularity of various resource types and entities, and domain coverage, to better guide and inform their prospective studies.

In early phases of the project we identified and analyzed over 70 online community databases and selected several based on quality and quantity of data, diversity in coverage and languages, and unique site-specific information. In instances where the data was not already openly available, agreements on data exchange were made, followed by the collection and processing of data for each community separately. In order to preserve as much of the semantic richness and diversity of the sources, a Resource Description Framework (RDF) based ontology consisting of a class structure and vocabulary is created for each dataset. The processing consists primarily of a transformation from the source formats, typically SQL tables, to an RDF serialization based on the respective ontologies. While more labor intensive than simply mapping the heterogeneous community data to a central ontology, this provides several advantages[7]. Most importantly, it means that the semantic richness of each source dataset is maintained, as the meanings and constraints of individual properties reflect the original data, while also allowing for easier alterations or updates based on changes made to the original source data.

In this paper, we will provide an overview of the individual communities from which data was collected in order to show how a more thorough understanding of the domain based will both impact our unified ontology, and impact the types of research and analysis that can be performed using this data.

2 Data Structures and Vocabularies of Community Data

Community data that the JVMG has already or seeks to incorporate into its knowledge graph can be organized into three general categories. First, communities that cover the Japanese visual media domain broadly, or contain data for

[1] https://jvmg.iuk.hdm-stuttgart.de/

the domain as part of a larger subset. These include the Media Arts Database[2], Wikidata[3] and AnimeClick[4]. Second are sources that cover a particular medium within the domain, such as anime or video games. Examples of this type are AniDB[5], and the Visual Novel Database[6], for anime and visual novels, respectively. The last group are communities that focus on a single aspect of the domain, but without a medium restriction, such as the Anime Characters Database[7].

The communities we have collaborated with all offer large amounts of data for different resource types, a varying but often high amount of descriptive granularity, and structured, though informal, data models. To illustrate these aspects, this section will feature a representative of each group and briefly discuss important features of each, such as the coverage and scope, primary entities and relationships, and descriptive properties. A table summarizing the data quantity of these three communities is shown in Table 1.

Table 1. Approximate quantities of sample community data.

Community	Works	Characters	Tags	Producers
AnimeClick	120,000	102,000	5,000	67,000
Visual Novel Database	28,000	91,000	5,400	31,000
Anime Characters Database	11,000	101,000	5,000	4,900

2.1 The Visual Novel Database

The Visual Novel Database (VNDB) is an online database for the visual novel genre of video games. This community focuses solely on visual novels (VNs), but features an extremely high granularity for most aspects of the genre, such as creative works, characters, and contributors. Interestingly, this granularity extends to the carriers of VNs (i.e. physical items for sale, or digital downloads) rather than only the content, which is the primary focus of most other communities. Both the content and carrier are represented by distinct entity types, Visual Novel and Release, with each having its own set of distinct properties and values. Content entities include those such as Characters and descriptive Trait and Tags, along with entities representing contributor roles such as Staff and Producer.[8]

[2] https://mediaarts-db.bunka.go.jp/
[3] https://www.wikidata.org/
[4] https://www.animeclick.it/
[5] https://anidb.net/
[6] https://vndb.org/
[7] https://www.animecharactersdatabase.com/
[8] A description of the dataset, along with the RDF ontology created by the authors for the VNDB is available at https://doi.org/10.5281/zenodo.5506936

Though the scope of VNDB is limited to a single medium, it is by far the most thorough dataset for that medium. Both creative works and their contents are described in extreme detail, and connections between relevant entities have been established. Relationships between entities are plentiful, with all related characters, releases, contributors, and other VNs being connected via a single umbrella VN entity. The granular description of contents is perhaps the greatest strength of VNDB, as the Tag and Trait hierarchies are robust, logically connected, and occasionally informally defined through links to external resources, primarily Wikipedia. Other communities may apply a trait such as "criminal" to a character, whereas VNDB features the more specific "blackmails" trait, which is a part of the "Engages In – Crime – Blackmail" hierarchy, is defined via a scope note, contains a link to the Wikipedia article on Blackmail, and provides "extortion" as an alias. This granularity extends to Tags attributed to VNs, where the parent group "Theme" extends to "Drama–Health Issues–Psychological Problems–Eating Disorder". Though this hierarchy is not a traditional subject authority file, building formal vocabularies of descriptive tags for niche mediums with user-generated content has previously been undertaken[4]. While the focus on a single medium is a clear limitation for a database covering Japanese visual media broadly, and descriptive data regarding content is arguably subjective, the opportunity to extend VNDB's Tag and Trait hierarchy to a wide range of applicable Japanese visual media, along with its fairly thorough coverage of an entire medium, are important reasons for its inclusion in a unified community dataset.

2.2 The Anime Characters Database

The Anime Characters Database (ACDB) is a community database dealing primarily with anime characters, though characters sharing an 'anime aesthetic' from other mediums, such as original art, video games, and manga, are also included. ACDB refers to itself as a 'visual search engine', and indexes characters according to a specific set of traits, such as hair and eye color, age, and type of clothing worn. This results in a uniform set of available traits for characters, and less ambiguity due to the limited numbers of descriptive traits available, but also results in a limited level of granularity compared to what a character may receive on VNDB. For example, characters can have their general hair length identified, but not a specific cut or style name. In addition to the central `Character` entity are `People` entities that represent voice artists, `Work` entities which are the source material for characters, and the descriptive `Character Tag` and `Series Tag` entities describing characters and works respectively. [9]

Though descriptive granularity is limited when compared to VNDB, ACDB contains data for a variety of mediums, and this has significant implications for the types of relationships between entities in the dataset. While both VNDB and ACDB contain relationships between creative works and characters, ACDB

[9] A description of the dataset, along with the RDF ontology created by the authors for the ACDB is available at https://doi.org/10.5281/zenodo.5508699

covering a variety of mediums beyond VNs means that descriptive data applied to a character is applied to many more instances of that character than those in VNDB. Similarly, while both contain recursive relationships between creative works, these relationships in ACDB extend beyond related VNs found in VNDB. One unique outcome of this is that ACDB can connect `Work` entities representing specific medium instances to a `Work` entity representing a broad series or franchise, something not possible in VNDB, as it by nature does not feature entities representing multimedia franchises. While the descriptive properties are still useful and do contribute unique data, relationships between entities are the most significant contribution of ACDB to a unified community dataset.

2.3 AnimeClick

AnimeClick is an Italian language fan site for Japanese anime, manga, and live action drama. The site is a general fandom wiki, and covers these mediums broadly, without a particular focus. Data are represented by four primary entities - `Animation Work`, `Comic Work`, `Character` and `Staff`. The `Animation Work` entity includes various animated formats, such as TV series, films, or original video animations, while `Comic Works` are primarily manga. The granularity of the data varies between entity types, with work entities being fairly detailed, and information on characters being more limited. Relationships between relevant entities are present, including adaptations of `Animation` and `Comic Works` to one another, and `Characters` to the `Staff` that voiced them. While some descriptive data is of limited utility due to it being in Italian, there is also a significant amount of useful language-agnostic data, such as episode counts, completion statuses, and connections between related resources. [10]

AnimeClick's dataset can be seen as a type of broad middle ground between the types of datasets represented by VNDB and ACDB. While VNs are largely absent from AnimeClick's dataset, it does describe multiple mediums, similar to ACDB. Descriptive granularity for characters is quite low, while works are described in greater detail. AnimeClick also features the parent-child relationship found in ACDB, connecting individual entities to a high-level franchise. Unlike ACDB, relationships are also present between members of a given franchise, with their relationship type defined, indicating sequels, prequels, derivative works, etc. Descriptive data for mediums beyond those covered in VNDB, provided by additional relationships between entities than those provided by ACDB, are the primary contributions that AnimeClick provides to a unified community dataset. Additionally, the ability to incorporate Italian data that is able to contribute usable and relatable data to a largely English dataset encourages the inclusion of other international communities, both for any additional data they may provide, and to expand the audience of the JVMG database.

[10] A description of the dataset, along with the RDF ontology created by the authors for AnimeClick is available at https://doi.org/10.5281/zenodo.5508683

3 Discussion and Impact of the Data Analysis

After having summarized each community's data in Section 2, this section explains the impact that the analysis of the data has had, and continues to have, on the ongoing development of the JVMG project.

3.1 Impact on the JVMG Ontology

The intention of the JVMG is to present a unified view on the domain of Japanese visual media and as such a unified domain model is needed. The analysis of the models derived from the fan databases has been very helpful in determining key aspects of the unified domain model. In particular, analyzing VNDB revealed that, while limited to only describing VNs and their contents, the granularity was unmatched by other sources, even when compared to other sources, such as Wikidata or the Media Arts Database. The benefits of this for researchers interested in VNs is clear, but immediate advantages for those interested in Japanese visual media more broadly, or simply other mediums, are less so. However, the phenomenon of multimedia franchises or 'media-mix' in Japan is extremely common, meaning a lot of VNDB's data, particular for characters and producers, can be used in conjunction with other datasets, once relationships between them have been established. For example, the granular trait data based on a character's appearance in a VN may be able to be applied to the same character's appearance in a manga, anime, or live action adaptation.

The need to facilitate and establish relationships between communities were also the takeaways of both the ACDB and AnimeClick datasets. Though descriptive granularity from these communities is limited when compared to VNDB, their inclusion of more media types and their ability to contribute unique data means that the linking of data from communities is greatly expanding the amount of information available to JVMG database users, and not simply providing redundant data from multiple sources. The media-mix mentioned previously is again important here; opportunities for connections between related works, creators, and characters from VNDB, ACDB, and AnimeClick are plentiful, so long as relationships can be identified and established. The importance of these relationships has informed the development of the JVMG RDF ontology, affecting both its properties and classes:First, the vocabulary needs properties that are able to link related entities. While this type of property is common in many ontologies, such as the `relation` or `isVersionOf` properties from the DCMI Terms vocabulary[11], the granularity of community data means that relationship types are often defined, e.g. sequels, prequels, or adaptations. To maintain the semantic richness of the source data, the granularity of relationship properties in our ontology should be equal to community data that the ontology is describing. While relational properties can link related resources while maintaining entity separation, combining related data into a single entity is beneficial for the visual browsing of multi-community data via a single access point. Research has

[11] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

shown that this is a preferred method for accessing a group of related media by various audiences[8, 9], and its use for Japanese visual media can be seen online in Wikipedia articles that describe multimedia franchises rather than single medium instantiations. The ontology should therefore also provide classes able to represent a similar set of related data from varying mediums and entities. This has been the subject of past research[10], and Section 2 touched upon how AnimeClick and ACDB have versions of this, but an extension of this concept and its modeling in the JVMG dataset should be implemented. The merging of data from multiple sources into franchise entities can also create problems relating to redundant or conflicting data stemming from communities using different primary source data, differing transliteration styles for Japanese titles or names, etc. Providing a way to mitigate these issues by creating properties allowing for the labeling of specific data as ground truth, where it has been able to be accurately identified, is also a feature that the ontology should support.

3.2 Impact on Researchers

The impact that community data has on research avenues was determined in part by three primary researchers that are a part of the internal JVMG team, and whose areas of interest include digital humanities and media studies. While additional external researchers have been collaborative partners as the project has progressed, with more planned in the future, feedback from these internal researchers during ongoing development was helpful in guiding various factors of the database; some of their findings and concerns are discussed here.

One example research question focussed on the relationship between character archetypes and the audience. Here, the VNDB data proved extremely valuable, as this medium often deals with romance as a theme, with the player taking the role of the protagonist. Due to the high granularity of the data, the researcher was able to identify and define distinct archetypes through analyzing co-occurring traits, and hypothesize the reasons for the frequency of certain archetypes and the effects they may have on the audience. While the preliminary analysis used only the VNDB dataset, it can be expanded on other media using the connections between the data sources.

Another researcher conducting more broad data-driven research found that despite the accumulation of multiple datasets providing a significant quantity of granular data, issues still arose when attempting to definitively address their various hypotheses. Research into topics such as the existence or lack of temporal changes in the relative frequency of shared character traits, and comparing networks of creators of large multimedia franchises, were impeded by either a limited level of data granularity, or by missing or conflicting data. These critiques suggest that the knowledge graph needs both additional granular data, as well as some amount of authoritative data and quality control.

As we are constantly seeking to incorporate additional data into the JVMG, the former of these concerns will hopefully be addressed as the project progresses, but the amount of extreme granular data for Japanese visual media is often limited, and there may simply be research questions that desire a subset of data

that is not readily available. With regards to data quality concerns, we have already begun to address this through the inclusion of data from the Media Arts Database (MADB). This source is a Japanese government funded database that includes descriptive metadata for anime, manga, and video games. Unlike the fan community data, MADB's data is, depending on the medium, sourced directly from producers, publishers, and library catalogues. While this will not result in a lot of additional granular data, specifically for the contents of creative works, it will be able to act as an authoritative source for important fields such as titles and names, addressing some data conflicts between different communities and acting as the source of ground truth mentioned in Section 3.1.

4 Conclusion and Future Work

In this paper, we have examined the current landscape of descriptive data for the Japanese visual media domain from the perspective of fan communities online, and the role that this data has had in informing the JVMG. We have identified three categories of fan communities and presented a representative example for each. We have shown that the data models can be connected to merge the information that is present in the individual datasets into a larger model. Also, while some data might be redundant, given the diverse interests of the individual communities, a significant part of the data augments the other sources and allows for the creation of a richer, more granular description of the included entities.. This better understanding of the domain informed key aspects of the JVMG, including how the central ontology should enable as many meaningful connections between community data as possible, and how we can better meet the needs of prospective researchers by continuing to incorporate more granular and more authoritative data into the knowledge graph.

Plans for future work are to address the issues revealed during the data analysis and impact assessments, i.e. concerns with data quality / authority, and the ability to establish relationships between community data. A separate analysis on data quality within the communities is currently underway, and the previously mentioned incorporation of the Media Arts Database dataset will directly address some authority concerns. The identification of related data available to connect is an ongoing process, and we are continuing to explore ways in which this can be done. Matching fields such as titles and creators has been successful, but we are currently exploring other ways to identify eligible data, such as via characters or matching Wikipedia links. After additional data has been connected and data quality further addressed, we hope to work with additional outside researchers and collect more feedback to further expand and improve the knowledge graph.

References

1. Fennell, D., Liberato, A. S. Q., Hayden, B., Fujino, Y.: Consuming Anime. Television & New Media **14**(2), 440-456 (2013). https://doi.org/10.1177/1527476412436986

2. Berndt J.: Anime in Academia: Representative Object, Media Form, and Japanese Studies. Arts.**7**(4), 56-69.(2018). https://doi.org/10.3390/arts7040056
3. Lee, J. H., Shim, Y., Jett, J.: Analyzing User Requests for Anime Recommendations. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '15), pp. 269-270. Association for Computing Machinery, New York (2015). https://doi.org/10.1145/2756406.2756969
4. Windleharth, T. W., Jett, J., Shmalz, M., Lee, J. H.: Full Steam Ahead: A Conceptual Analysis of User-Supplied Tags on Steam. Cataloging & Classification Quarterly **54**(7), 418-441 (2016). https://doi.org/10.1080/01639374.2016.1190951
5. Mittell, J.: Sites of participation: Wiki fandom and the case of Lostpedia. Transformative Works and Cultures **3**, (2019). https://doi.org/10.3983/twc.2009.0118
6. Popova, M.: Fan studies, citation practices, and fannish knowledge production. Transformative Works and Cultures **33**, (2020). https://doi.org/10.3983/twc.2020.1861
7. Kiryakos, S., Pfeffer, M.:The Benefits of RDF and External Ontologies for Heterogeneous Data: A Case Study Using the Japanese Visual Media Graph. In: Information between Data and Knowledge: Proceedings of the 16th International Symposium of Information Science (ISI 2021), pp. 308-320. Glückstadt: Verlag Werner Hülsbusch Regensburg, (2021). https://doi.org/10.5283/epub.44950
8. Lee, J. H., Clarke, R. I., Rossi, S.: A qualitative investigation of users' discovery, access, and organization of video games as information objects. Journal of Information Science **42**(6), 833-850 (2016). https://doi.org/10.1177/0165551515618594
9. Talleråas, K., Dahl, J.H.B., Pharo, N.: User conceptualizations of derivative relationships in the bibliographic universe. Journal of Documentation **74**(4), 894-916 (2018). https://doi.org/10.1108/JD-10-2017-0139
10. Kiryakos, S., Sugimoto, S.: Building a bibliographic hierarchy for manga through the aggregation of institutional and hobbyist descriptions. Journal of Documentation **75**(2), 287-313 (2019). https://doi.org/10.1108/JD-06-2018-0089