The Benefits of RDF and External Ontologies for Heterogeneous Data

A case study using the Japanese Visual Media Graph

Senan Kiryakos
Stuttgart Media University
Germany
kiryakos@hdm-stuttgart.de

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

Abstract

Across numerous fan created and curated websites, there exists a wealth of semantically rich descriptive data for a variety of Japanese visual media, such as anime, manga, and video games. The amount and granularity of these data makes it valuable for domain researchers, but the semantic heterogeneity and lack of interconnectedness makes analysis cumbersome. Seeking to address this issue, the Japanese Visual Media Graph (JVMG) project aims to create a type of global database built using RDF and external ontologies to better enable data-driven research in the domain. We discuss the benefits this approach has when compared to the local relational databases used at the fan-sites, such as enabling the easy creation of aggregate resources using data from multiple providers, and the ability to constantly update and alter a schema over time. This encourages the incorporation of additional data in the future, while still maintaining rich data semantics and provenance. While many of the benefits are discussed in the context of the JVMG project or the Japanese visual media domain, we discuss how this database approach may be similarly advantageous to other projects that seek to create aggregate resources, or collect heterogeneous data from diverse sources.

Keywords: ontology-based data access; semantic heterogeneity; knowledge graph; fan-curated data; digital humanities

1 Introduction

For researchers interested in Japanese visual media (e.g. anime, manga, video games), the largest collection of semantically rich descriptive data exists not at existing memory institutions, but across a multitude of fan-created and

curated websites. These data are abundant, highly granular, multilingual, and provide a rich source for which data-driven research can be conducted. The nature of these data make it valuable for researchers studying various aspects of the Japanese visual media domain, such as themes, genres, and characters, as well as trends, receptions, and influences. In attempting to better meet the needs of researchers, the Japanese Visual Media Graph (JVMG) project (Pfeffer & Roth, 2019) seeks to utilize this community-created data to e develop a graph-based, highly interconnected database (i.e. a type of knowledge graph), combined with flexible and powerful querying and analytical tools.[1] It should be noted that while the database may contain some links to visual media found at the original data providers, the primary contents of the database are descriptive metadata and not the visual materials themselves.

The desire to utilize this fan-created data in an interconnected database poses several issues, but also unique opportunities. Relying on multiple data providers means the data is fairly heterogeneous, available in different formats, covering diverse entity levels, described according to different data models, and using different vocabularies. Thus, the integration of this disparate data into a single database involves significant data transformation, property mapping, the identification of related or identical resources, and the creation of a unified global ontology. Though a challenge, this integration presents us with unique opportunities, such as merging related resources with data sourced from multiple providers, developing a more formal data model and ontology for the domain of Japanese visual media, and exploring different ways of making these data accessible to, and able to be queried by, researchers.

In detailing this database approach, this paper outlines the issues and opportunities encountered by the JVMG project thus far, and highlights the advantages and disadvantages presented when using a database that relies on linked data formats and external ontologies. We discuss the benefits that this approach has had for primary aspects of our project specifically, such as the integration of heterogeneous data, and the creation of aggregated resources,

[1] The database is currently for internal use and will be made more publicly available later this year. Up-to-date information on the state of the project and means of access will be published on the project homepage https://jvmg.iuk.hdm-stuttgart.de/

while also presenting this approach as one that may be of interest to projects with similar goals regarding data integration.

Following this introduction, Section 2 describes the data providers, the type of data being collected and integrated, and data processing. Section 3 outlines the differences between databases currently used with these data and the approach we use, along with discussing two specific advantages. Lastly, we conclude with thoughts and future plans in Section 4.

2 Data Providers

In order to aid those seeking to conduct data-driven research on the Japanese visual media domain, a large amount of descriptive data is required. The best source of these data can change depending on the domain. Memory institutions, such as libraries or archives, for example, contain a wealth of data for more traditional literary materials, though this is often of a limited granularity due to factors such as cataloguing rules. Past research (e.g. Kiryakos et al., 2017) has shown that for media types like anime, manga, and video games, the best source of data is instead found across a variety of hobbyist webpages. The amount of data is not only vast, but extremely granular, as communities with various interests and perspectives describe the same materials in distinct ways.

As such, we rely on these communities, several of whom we have made formal collaboration agreements with, to utilize their data. These partners include the Anime Characters Database, AnimeClick, and the Visual Novel Database (VNDB). Other sources that allow for open data use, such as Wikidata, are also utilized. Though the coverage and contents of these sites differ, they generally contain traditional descriptive data, such as titles, creators, and dates, along with more granular data such as genres and tags, character lists and traits, relationships between related entities (e.g. characters or creative works), and organizational provenance. The amount of some of these data is shown in table 1. This is the type of information that we seek to leverage to make data-driven research questions regarding Japanese visual media more easily investigated. Before doing so, however, work on the data must be performed, as the communities provide us with data in different ways, using different formats, and according to their own individual data models and vocabularies. Though a detailed description of the data

transformation pipeline is outside the scope of this paper, an outline of this transformation will help better illustrate how the data comes to be integrated and connected in our database.

TABLE 1: Approximate numbers of various data collected from providers.

Visual Novel Database					
Visual Novel Works	Unique Releases	Characters	Descriptive Work Tags	Descriptive Character Tags	Agents
28,000	73,000	91,000	2600	2800	31,000
AnimeClick					
Anime	Manga	Characters	Related Works	Authors	Other Staff
9400	11,000	102,000	15,000	28,000	39,000
Anime Characters Database					
Characters	Works	Character Relations	Descriptive Character Tags	Descriptive Work Tags	Voice Actors
101,000	11,000	17,000	3800	1100	4900

2.1 Data Transformation and Ontology Creation

Our data is sourced in a variety of ways, e.g. as data dumps or through APIs, depending on the provider, and is typically either in the form of SQL tables or JSON. As our database uses Resource Description Framework (RDF) serializations, data transformation is the first technical step, though this is preceded by an analysis of each provider to better understand their data model. Conversion is done using Python, and is then processed using the RDFLib package. The resulting output is a series of RDF files able to be integrated and queried using the query language SPARQL. While other existing methods of creating RDF output based on relational data exist (see e.g. Arenas et al. 2012; Das et al. 2012), our method differs, as we do not create a mapping layer over the relational data, but instead ingest and transform it directly.

This processing is a central part of the data integration process, as RDFLib allows us to define and output RDF triples using non-RDF input data we

receive from our data providers. Through an analysis of the data, an ontology is created for each data provider, which includes a class structure based on how resources are presented by a provider, and an accompanying RDF vocabulary file. At its most basic, an example of this vocabulary conversion would be the 'title' column header in an SQL table from the VNDB becoming `<http://mediagraph.link/vndb/ont/title>` - a property in our namespace that represents that same title column header. An excerpt of this original tabular data from VNDB with limited properties is shown in table 2, with its RDF transformation shown in figure 1. While this RDF data is able to be queried using SPARQL, a separate RDF vocabulary file combined with our web frontend allows for these data to be displayed and browsed in a more human-readable way via a web browser; a sample page is shown in figure 2.

TABLE 2: A sample of tabular data for a visual novel from VNDB.

Visual Novel table					
id	**title**	**original**	**pid**	**l_wp**	**l_wikidata**
7014	Dangan Ronpa Kibou no Gakuen to Zetsubou no Koukousei	ダンガンロンパ 希望の学園と絶望の高校生	1761	Danganronpa:_Trigger_Happy_Havoc	Q1035512
Producer table					
id	**name**	**original**	**type**	**l_wp**	**l_wikidata**
1761	Spike	スパイク	co	Spike_%28company%29	Q2541040

```
@prefix vndb: <http://mediagraph.link/vndb/ont/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://mediagraph.link/vndb/vns/7014> a vndb:VisualNovel ;
    rdfs:label "Dangan Ronpa Kibou no Gakuen to Zetsubou no Koukousei"@en,
        "ダンガンロンパ 希望の学園と絶望の高校生"@ja ;
    vndb:l_wp "https://en.wikipedia.org/wiki/Danganronpa:_Trigger_Happy_Havoc" ;
    vndb:l_wikidata <https://www.wikidata.org/wiki/Q1035512> ;
    vndb:original "ダンガンロンパ 希望の学園と絶望の高校生" ;
    vndb:title "Dangan Ronpa Kibou no Gakuen to Zetsubou no Koukousei" ;
    vndb:producedBy <http://mediagraph.link/vndb/producers/1761> ;
    vndb:vndbLink <http://vndb.org/v7014> .

<http://mediagraph.link/vndb/producers/1761> a vndb:Producer ;
    rdfs:label "Spike" ;
    vndb:l_wikidata <https://www.wikidata.org/wiki/Q2541040> ;
    vndb:l_wp <https://en.wikipedia.org/wiki/Spike_%28company%29> ;
    vndb:name "Spike" ;
    vndb:original "スパイク" ;
    vndb:type "co" ;
    vndb:vndbLink <http://vndb.org/p1761> .
```

FIG. 1. Table 2 after RDFLib data processing and transformation.

FIG. 2. Web frontend screenshot for RDF data shown in figure 1. Note the ‘Producer’ data is a separate page linked to via the ‘Produced by – Spike’ value URL.

Data transformation and ontology creation for each provider allows the loading and querying of each dataset separately, providing users with the ability to access datasets from individual or multiple providers. Creating 'local' schemas in this way also allows for a more straightforward relation between the providers' data model and a unified 'global' schema, which must occur at some point during data integration (Cruz & Xiao, 2005; Doan et al., 2012). This relation mapping between the ontology of each data provider and our own unified ontology is performed primarily using the Web Ontology Language (OWL) and Simple Knowledge Organization System (SKOS), allowing us to map properties and concepts between datasets.

In addition to the benefits discussed in Section 4, an intrinsic advantage resulting from this transformation is that much of the formerly tabular data is turned into web-resolvable URIs, enabling their inclusion in other web applications, and as a browser-based entry point for the database. This "things, not strings" (Singhal, 2012) approach is fundamental to providing the benefits of knowledge graphs, and our data transformation brings these benefits to a large amount of data in the Japanese visual media domain. This applies not only to properties, such as the 'title' example mentioned previously, but to their values, resulting in unique web identifiers for single creative works, franchises, their creators, and their contents.

3 Benefits and Drawbacks of an External Ontology Database Approach

The data sourced from our providers belongs to schema-bound, local databases (i.e. tabular relational databases). This approach requires the schema to be developed first, and results in a data model that is difficult to later alter. This also means the data is viewed from a single perspective, and can discourage the integration of new data not considered during initial development. This also makes it difficult to incorporate data from additional sources, as the schemas will inevitably differ, requiring adjustment based on these new sources.

The database approach we instead use is that of a schema-less, global database combined with external ontologies. As they are separate from the data itself, the ontologies can easily be altered after their initial creation, which more readily supports the incorporation of new heterogeneous data. The external ontologies here refer to individual RDF ontologies, mentioned

in Section 2.1. The varying granularity of the data we use also encourages the use of RDF-based ontologies, as they are better able to specify domain knowledge and semantic richness when compared to relational databases (Buron et al., 2019; Munir & Sheraz, 2018). This is not to say that the approach we use is preferred over relational databases in all cases, but rather this approach has provided significant benefits for the JVMG project specifically. We feel these benefits are worth highlighting so that other similar projects (e.g. those working with heterogeneous data or those seeking to create aggregate resources) may be informed on the advantages and disadvantages of this approach. The following section discusses some of these benefits in further detail.

3.1 Creating Aggregate Resources

Each data provider has a separate dataset and receives its own unique ontology for use in our database, allowing for the browsing and querying of individual datasets when required. As much of the data across providers describes the same or related resources, researchers would also benefit from the ability to query combined datasets which contain this related, merged data. This is particularly beneficial to our project due to the high granularity and differing perspectives of the community data providers. If our main source of bibliographic data were libraries, for example, we would expect a significant amount of duplicate data due to restrictive cataloguing rules, and the benefits of aggregating data would be limited. The advantages of aggregation become more clear when understanding the nature of diverse fan-created source data, some of which focuses on particular aspects of Japanese visual media rather than the creative work itself, such as the characters (i.e. the Anime Characters Database).

Our RDF-based, external ontology approach allows for the easy creation of these aggregate resources. During the merging process, all of the matching data from all datasets related to a single related resource is aggregated, including duplicate or conflicting data. In a next step, a curated subset of attributes is created for each entity type, and existing data is mapped to these attributes. Duplicate and conflicting information is resolved in this step by different means: a preferred piece of information can be selected (e.g. the canonical title) and the remaining pieces are listed as variations (e.g. title variants). As ontologies are external to the data, we can choose a unique part

of the namespace to be used with these aggregate resources. Any researcher working with the domain data can then solely focus on this consolidated ontology, as it will cover the whole domain and make the heterogeneous data accessible in a uniform way. As the original provider data as well as the curated data is available using the same entity URL, but with different ontologies, both provenance and all original values from the different sources are preserved and are immediately available. Though the creation of a separate ontology for the merged dataset requires additional labour, it makes the data more convenient to browse, and allows for the modelling of the domain as a whole, rather than the modelling of single datasets from individual providers. As this step takes place after the creation and analysis of both individual provider and aggregate datasets, the resulting ontology reflects a more thorough understanding of the complete domain and its data; this is a key advantage of the schema-less approach presented here, as in more traditional databases, this step would have come prior to merging, and unexpected changes to the schema may then need to take place, requiring difficult ontology or data pipeline adjustments.

The creation of aggregated resources also allows for an easier managing of different data quality levels, as data is additive and conflicting statements are more visible. As we do not edit source data, and researchers may be interested specifically in conflicting or contentious data across communities, these conflicts and other quality issues are generally left as is. That said, in order to assess data quality and possible types of mistakes and to ease matching across providers, a random selection of statements is currently manually checked against ground truth where available (e.g. the visual media itself, credit rolls) as a part of a study on fan community data quality.

One drawback inherent to this data integration step is that when merging data from multiple providers, care must be taken to ensure they are describing not only the same resource, but the same entity level of that resource. Entity levels here refer to a separation of 'levels' of a resource that may be described by a given provider. For example, one provider may describe an entire franchise or intellectual property as a single resource, another may describe its manifestation in a single medium such as anime. A prominent example of this entity partitioning can be seen on Wikipedia, where the franchise *Pokémon* has separate articles for the umbrella franchise, the series of video games, single games within that series, etc. One popular model defining such

entity levels is the FRBR entity-relationship model (IFLA, 2009), which defines Work, Expression, Manifestation, and Item entities. Though we do not currently define unique entities in the data we collect, we do analyze the data structure and properties used by each provider so that we can determine whether or not descriptions are about identical or related entities.

The significance of this issue is dependent on the domain and heterogeneity of the data being merged. For our project, particular attention must be paid here, as entity levels described for Japanese visual media amongst enthusiast communities frequently vary (Kiryakos & Sugimoto, 2019), and because relationships across different entity levels is quite common in Japanese visual media (Lee et al., 2018). If granular data from each provider is maintained (e.g. if a merged resource contains information on both manga volumes and anime episodes), it becomes difficult to define what exact entity or concept that resource is describing. As each resource in our database is given a unique URI, merged resources can be referred to globally in Linked Data or other web contexts. If an aggregate resource contains data from various entity levels, we cannot claim that its URI represents a common entity type, such as a franchise or single manifestation, as is possible if data is more strictly partitioned and merged. This problem can arise with any project integrating heterogeneous data, but it should be noted that our approach does not provide a solution for this, and that entity identification and separation must be performed if desired.

3.2 External Ontologies

The ontology being separate to the data itself provides a level of flexibility that is not present with traditional relational databases. The class structure and vocabulary is able to be changed whenever desired, either to reflect an updated understanding of the data, or to incorporate new data requiring the addition of new properties. Through the use of multiple external ontologies mapped to one another, a single dataset can be viewed and presented in distinct ways. Datatype constraints are also not rigid, allowing for different value types, e.g. URI and string, to exist for a single property. This becomes important when merging data, as one provider may describe an author using a plain text string, whereas another may use a URI linking to something such as the Virtual Internet Authority File.

As our approach utilizes RDF, the use of existing, widely used vocabularies, such as those provided by Dublin Core and Schema.org, is also possible. Depending on the domain, such vocabularies may be sufficient for modeling single or aggregate datasets. As the Japanese visual media domain is fairly niche and the source data is highly granular, the creation of a new ontology was preferred (though some related niche media controlled vocabularies exist, such as the Comic Book Ontology). Though this requires significantly more labour, it allows for a given dataset to be modeled and described in an exact manner, and results in a more thorough understanding of the domain. The creation of a new RDF ontology, particularly for a previously uncovered domain, also enables its inclusion in other Linked Data applications.

A drawback with this approach is that despite the flexibility and conveniences that come with using external ontologies, the resulting schema can be difficult to enforce, unlike relational database tables, particularly once resource merging takes place. For example, one may define what value types are valid for a given set of attributes based on an analysis of existing data. If new data is integrated and merged that provides alternative values for these attributes, the schema definition will be invalid (though the data will remain retrievable). One remedy to this is to regularly update the ontology to reflect the newly integrated data, which external ontologies easily support. If one wishes to define a more strict data model, one method is to use RDF validation and constraint methods, such as Shape Expressions (Prud'hommeaux, et al., 2014) and Shapes Constraint Language (Corman, et al., 2018), which allows for the definition of conditions that must be present in an RDF graph to be considered conformant.

4 Conclusion

Through the development of a database which utilizes RDF and external ontologies, the JVMG project is poised to meet its goal of serving researchers studying Japanese visual media. This database approach allows for the easy and ongoing integration of heterogeneous data from multiple providers, while maintaining semantic richness, and modelling the domain in the process. Additionally, this approach allows for the seamless creation of aggregate resources, which merge related data from multiple providers. The use of external ontologies encourages the integration of new data sources, further

improving the coverage and granularity of the database, as ontologies can be easily adapted to incorporate additional heterogeneous data over time. The resulting database provides researchers with a single source with which to browse or query a wealth of granular, interconnected data on the Japanese visual media domain.

Future work seeks to build on this development in multiple ways. First, we plan to obtain more data from various web resources, particular those that offer existing query or collection methods, such as Wikipedia, Wikidata, and various Fandom pages. While a 'coverage ceiling' will eventually be reached, this extra data can still be useful, e.g. for finding additional connections between entities. Also, we will focus on making the data more accessible to researchers who do not have the knowledge to create large SPARQL queries on the fly. In order to learn more about data-driven research methods, we employ "tiny use cases"[2], i.e. small research questions that require data from the domain. Mixed groups of researchers from media/Japanese studies and information/computer science respectively work together on these questions and document their approaches and findings. These can be used to familiarize external researchers with the database and its ontologies as well as introduce typical data analysis workflows and techniques. Once more external researchers have started doing active research using the project's data, we will have a better understanding of what features and functions could be added that go beyond predefined queries or query wizards.

5 Acknowledgements

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

6 References

Arenas, M., Bertails, A., Prud'hommeaux, E., & Sequeda, J. (2012). A Direct Mapping of Relational Data to RDF, W3C Recommendation.

Buron, M., Goasdoué, F., Manolescu, I., & Mugnier, M.L. (2019). Ontology-Based RDF Integration of Heterogeneous Data [Technical Report]. LIX, Ecole polytechnique.

[2] Several writeups about these can be found at
https://jvmg.iuk.hdm-stuttgart.de/category/tiny-use-case/

Corman, J., Reutter J.L., & Savković, O. (2018). Semantics and Validation of Recursive SHACL. In Vrandečić D. et al. (eds) *The Semantic Web – ISWC 2018. ISWC 2018.* Lecture Notes in Computer Science, 11136, 318-336.

Cruz, I.F. and Xiao, H. (2005). The role of ontologies in data integration. *Engineering Intelligent Systems for Electrical Engineering and Communications*, 13(4), 245-252.

Das, S., Sundara, S., & Cyganiak, R. (2012). R2RML: RDB to RDF Mapping Language.

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration.* Elsevier.

IFLA Study Group on the Functional Requirements for Bibliographic Records. (2009). Functional Requirements for Bibliographic Records, from https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

Kiryakos, S. & Sugimoto, S. (2019). Building a bibliographic hierarchy for manga through the aggregation of institutional and hobbyist descriptions. *Journal of Documentation*, 75(2), 287-313.

Kiryakos, S., Sugimoto, S., Nagamori, M., & Mihara, T. (2017). Aggregating Metadata from Heterogeneous Pop Culture Resources on the Web. *International Conference on Dublin Core and Metadata Applications* 2016, 65-74.

Lee, J.H., Jett J., Cho, H., Windleharth, T., Disher, T., Kiryakos, S., & Sugimoto, S. (2018). Reconceptualizing superwork for improved access to popular cultural objects. *Proceedings of the Association for Information Science and Technology* 55, 274–281.

Munir, K. & Anjum, M.S. (2018). The use of ontologies for effective knowledge modelling and information retrieval. *Applied Computing and Informatics*, 14(2), 116–126.

Pfeffer, M. and Roth, M. (2019). Japanese Visual Media Graph: Providing researchers with data from enthusiast communities. *International Conference on Dublin Core and Metadata Applications*, 136–141.

Prud'hommeaux, E., Gayo, J.E.L., & Solbrig, H. (2014). Shape expressions: an RDF validation and transformation language. *Proceedings of the 10th International Conference on Semantic Systems*, 32-40.

Singhal, A. (2012): Introducing the Knowledge Graph: things, not strings. In Google Official Blog, from https://blog.google/products/search/introducing-knowledge-graph-things-not/.