# Aggregating Manga Metadata Across Institutions: Lessons Learned in the Application of EDM

Senan Kiryakos[1], Shigeo Sugimoto[1]
[1]University of Tsukuba

## Abstract

Many different institutions create bibliographic data for manga, a style of Japanese comic. These institutions typically describe the same resources, but in different ways. The exchange of data would result in a more complete bibliographic data landscape for manga, however the majority exists in isolation from one another. In seeking to connect this data, this poster discusses a study that created a Linked Data model for manga, based on the Europeana Data Model and using Dublin Core and BIBFRAME vocabularies for bibliographic description. Data was collected and aggregated from Monash University's JSC Manga Library, the Media Arts Database from the Japanese Agency for Cultural Affairs, and Wikipedia. This poster outlines the issues encountered in the creation of the model, the lessons learned from these issues, and possible future extensions of the study.

**Keywords:** Metadata; metadata aggregation; bibliographic data for manga; europeana data model; BIBFRAME
**doi:** 10.9776/16482

**Acknowledgements**: The authors express our thanks to Mitsuharu Nagamori and other members of our laboratory for their input. We thank Monash JSC, Agency for Cultural Affairs, and Toppan Printing Co. Ltd. for contributing data. This research is supported in part by JSPS KAKENHI Grant Number Grant No. 25240012 and collaborative research project with Infocom Corporation.
**Contact**: senank@gmail.com, sugimoto@slis.tsukuba.ac.jp

## 1 Background

Between 2013 and 2015, the authors conducted a study that utilized Linked Data (LD) technologies in the aggregation of bibliographic data for manga, a form of Japanese comic. The data that was being aggregated came from multiple institutions, namely academic and special libraries, corporate databases, and the Web. The purpose of aggregating this bibliographic data was to improve the granularity of bibliographic description for manga at these respective institutions by aggregating their data in the LD cloud, allowing institutions to use existing data from other sources to build on their own.

The specific institutions used in the case study were Monash University's Japanese Studies Centre (JSC) Manga Library (MML), the Media Art Database for Manga (MAD-M) for the Japanese Agency for Cultural Affairs, created by Toppan Printing Co. Ltd., and Web data from Wikipedia. Initially, data was also gathered from US Academic libraries, though a lack of a large amount of data and the similarity to the MML data resulted in its exclusion. The case study was successful in improving granularity through aggregation, both because these different institutions describe different bibliographic properties for manga, and do so in different languages. The process for data aggregation involved transforming the data from the MML and MAD-M databases into RDF-based vocabularies, with Works using Dublin Core (DC) and Items using BIBFRAME. The MML and MAD-M databases contain records for individual volumes of manga, or in FRBR entity terms, Items. The Web data from Wikipedia, however, typically describes manga as an entire series, agnostic of specific publication instances or media formats, or, in FRBR terms, Works. An aggregation model based on the Europeana Data Model (EDM) was developed that not only allowed for the aggregation of data from multiple institutions for individual manga resources, but also the establishment of relationships between manga Items (volumes) and the Work entity to which they belong. In other words, the aggregation model allowed for greater levels of bibliographic description for manga by aggregating bibliographic properties from various institutions for volumes, while also connecting them to the respective Work level resources to which they belong.

This poster will detail the lessons learned in utilizing of an EDM-based model for use with institutional data for manga, as well as outline future work and extensions of the model.

## 2 Applying EDM to Serialized Manga: Issues and Lessons

This section will discuss issues encountered and lessons learned in the application of an EDM aggregation model to serialized manga resources.

## 2.1 Practical Implementation of the Model

While the model was successful in its goal at the level of the case study, the problems of practical implementation on a wider scale are apparent.

EDM was chosen for the model as it is perhaps the most prominent and successful LD aggregation model, but it's current implementation is mainly to aggregate data for use in the Europeana portal rather than a data model used for sharing amongst individual institutions. Therefore, while the model created would be useful in the creation of a bibliographic manga portal, it may not be the best choice for individual institutions that wish to include external bibliographic data in their records.

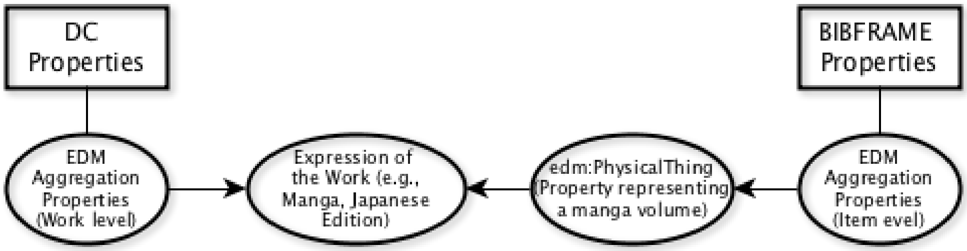
DC Properties
BIBFRAME Properties
EDM Aggregation Properties (Work level)
Expression of the Work (e.g., Manga, Japanese Edition)
edm:PhysicalThing (Property representing a manga volume)
EDM Aggregation Properties (Item evel)

Figure 1. Aggregation Model. A simplified overview of the full aggregation model.
Arrows flow to the conceptual Work entity.

Apart from EDM, the use of BIBFRAME may also hinder adoption. The BIBFRAME model and vocabulary was chosen as it supported the varying levels of descriptive granularity coming from different institutions, as well as provided a LD-capable bibliographic description vocabulary for the data, a requirement for the data to work within the EDM-based model. It was also an interesting opportunity to examine how the new BIBFRAME model would work with non-library data sources. Overall, BIBFRAME was successful in performing LD capable bibliographic description, but its future is still rather unclear. At present, the wide implementation of BIBFRAME by non-library institutions is unlikely, though it does not appear to be as much of a hurdle as was expected when the study began. The usage of EDM, however, will most likely remain problematic in regards to implementation of the model, though it does enable the exploration of manga resource portal, discussed in Section 3.

Note that while BIBFRAME was used for the MML and MAD-M databases, both of which deal with individual volumes of manga, DC was used to describe Web data from Wikipedia, as bibliographic data at the Work level is generally broad enough to be sufficiently described by DC terms rather than requiring specific properties such as those found in BIBFRAME.

## 2.2 Identifying Matching Manga Resources

In order to aggregate bibliographic data for specific manga volumes, matching resources across databases have to be identified. Issues encountered during this portion of the study are as follows.

First, when aggregating data at the Item level, that is, for specific publication instances of manga volumes, one has to match multiple bibliographic properties. As a single volume can be published multiple times, one must not only match series and volume title, but edition-specific information such as publisher information. In addition, matching data must also be found across languages. Again, simple title matching would not be sufficient if translations were not already available.

Similar to a study by He, et al. (2013), a solution was to leverage LD resources and obtain URLs that could represent a manga Work, regardless of language used. This was done using the Reconciliation function of the OpenRefine software, which allows one to match tabular data against various Web resources. In the study, MML and MAD-M data were reconciled against DBpedia to obtain identical URLs for use across languages where title matching would not work. For example, reconciling MML entries labeled "Astro Boy" and on MAD-M entries labeled with the Japanese title for Astro boy (鉄腕アトム) produces the same URL, http://dbpedia.org/page/Astro_Boy.

While this solution aids in aggregating data at the Work level, it does not solve matching Item level data about specific publication instances. An automated method of matching data at this level has yet to be discovered by the authors, so much of the Item level aggregation was done manually. Any future extensions of this study that require Item level aggregation should therefore be preceded by an investigation to automate this process.

### 2.3 Usefulness of Higher Granularity

While aggregating data from multiple sources increased granularity, the level to which the increase is of significant benefit to users remains to be investigated.

A user-focused study to gauge the quality of improvement may be needed to determine whether the level of granularity is satisfactory. In *Where Is the Justice… League?: Graphic Novel Cataloging and Classification,* Fee (2013) points out that fans of comics and manga typically have an interest in minutiae, though specific bibliographic properties have yet to be identified. Some interesting granular information, such as chapter titles and full volume summaries, were not incorporated in the study due to a limitation in data gathering; the information is available on Wikipedia, but not accessible via unique properties in DBpedia. This means that the data is accessible if one uses methods such as HTML scraping, but not through the LD methods used to access other properties. If users are interested in this information, a method to include it should be identified.

## 3 Future Extensions

While the study was successful in achieving its modest goal, it laid the groundwork for future work, mainly in the related domain of pop-culture metadata.

As mentioned in Section 2.1, the use of EDM invites the possibility of a portal that gathers and makes available bibliographic data for manga. This would require some substantial technical work, but the model demonstrates that the aggregation of data, similar to the Europeana Portal, is a possibility for manga resources. After the conclusion of the study, the MAD-M database was made available online at http://mediaarts-db.jp/mg/. While not a data portal and containing data only from MAD-M, it serves as a tangible and visual example of how a bibliographic manga database may operate, and allows one to easily point out where aggregation from other sources would improve the available data.

Section 2.3 mentioned the possibility of a survey to discover specifically what minutiae users of manga would be interested in have described. With this in mind, any future attempts at model implementation should first look at conducting such a survey to identify what specific data should be aggregated.

As a long-term extension of the study, the authors are also interested in incorporating non-manga resources. The aggregation model was designed with this possibility in mind, and it allows for the future inclusion of such resources; the Expression of the Work node in Figure 1 represents the model portion that enables this. Often, a manga will spawn related works in other media formats, such as Anime, films, video games, etc. An obstacle in including these resources will be finding appropriate models and ontologies that sufficiently describe each medium, while also working within a LD model. Still, pursuing this would result in a more complete bibliographic landscape for these unique pop-culture resources and will hopefully be examined in the future.

## 4 References

Fee, W. T. (2013). Where Is the Justice… League?: Graphic Novel Cataloging and Classification. *Serials Review*, 39(1), 37–46. doi:10.1016/j.serrev.2013.02.004

He, W., Mihara, T., Nagamori, M., & Sugimoto, S. (2013). Identification of works of manga using LOD resources. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL '13*. doi:10.1145/2467696.2467731