Integrating SORAN's Dataset into ARIADNEplus

Peter Yanase

The following is based on the paper titled "Integrating the Japanese Archaeological Dataset into the ARIADNEplus Data Infrastructure," delivered at the Digital Humanities 2022 conference (July 25–29, 2022) by Franco Niccolucci, Yuichi Takata, and Peter Yanase.

Ⅰ Introduction

The Comprehensive Database of Archaeological Site Reports in Japan (SORAN) is Japan's largest repository and aggregator of archaeological data and information. It is operated by the Nara National Research Institute for Cultural Properties (NABUNKEN), one of the two national research institutes focusing on cultural heritage. SORAN primarily functions as an index of domestic archaeological excavations. Its catalog currently contains information on roughly 140 thousand archaeological interventions and 110 thousand publications—of which circa 30 thousand are available in PDF. SORAN is an immensely popular service that in 2020 had over 13.5 million visits and 78.5 million page views. However, because it was originally built for the domestic market, its spatial coverage is delimited by national borders and its user base by a language barrier.

To overcome the limitations of SORAN, NABUNKEN decided to integrate a part of its data into the Archaeological Research Infrastructure for Archaeological Data Networking in Europe (ARIADNEplus).

Ⅱ What is ARIADNEplus?

ARIADNEplus is a project funded by the European Commission "to provide open access to Europe's archaeological heritage and overcome the fragmentation of digital repositories, placed in different countries and compiled in different languages."[1]

The most readily visible part of the project is the ARIADNE Portal, a website providing access to the ARIADNE Catalogue containing the aggregated metadata of the project partners.

Ⅲ SORAN's Source for Meta Data

The metadata stored in SORAN comes from various sources, of which the datasheets attached to the fieldwork reports are the most important. These sheets contain information on every archaeological intervention covered in a given fieldwork report and record the name, location (address), position (latitude and longitude), size, type, age(s) of the sites excavated, the date and reason for the excavations, and the most significant structural remains and materials found. The information from the datasheets is uploaded to SORAN by local governments, museums, universities, and academic societies through a Web interface. The main problem with the uploaded data is that there is no strict regulation on what exactly should be written in the cells of the datasheets that provide its basis. Therefore, to integrate SORAN's data into the ARIADNE Catalogue, NABUNKEN and ARIADNEplus had to collaborate closely in a long integration process involving data cleansing, schema transforming, and concept mapping.

Ⅳ Transforming the Data

The ARIADNE Catalogue is searchable according to the three facets of "where" (space), "when" (time), and "what" (object) based on controlled vocabularies. Mapping SORAN's internal data schema to ARIADNE's ontology was a largely technical step we could finish in a few weeks. Mapping the Japanese data to the three facets was more challenging.

The first facet required spatial coordinates to be converted to comply with the WGS84 (World Geodetic System 1984), which a significant amount of the original data did not follow. On top of that, many of the manually entered coordinates had typos. We solved these problems with a combination of scripts and manual intervention. For example, we flagged all the sites that had coordinates in the sea but were not underwater sites.

The second facet required temporal information to be linked to definitions stored on PeriodO[2], a multilingual gazetteer of temporal information. As a first step, we established a controlled vocabulary for the periods. Next, we converted all past entries in the database to conform with the new vocabulary. Finally, we altered SORAN's interface to only accept entries from the controlled vocabulary moving forward. However, there was a further obstacle to be cleared: no single authoritative source covered all the periods used by SORAN. To solve this, NABUNKEN arranged an extended discussion of the possible definitions among its interdisciplinary team of experts. We compiled the results of the discussions in an internal document and then registered the new definitions in PeriodO[3].

The final facet of objects required the most work as it involved mapping culture- and discipline-bound terms to the Getty Art & Architecture Thesaurus[4] (AAT). Similar to the temporal entries, the data entered in the database was eclectic and contained many typos. However, we opted not to cleanse the data to keep the integrity of the uploaded information. Instead, we generated a list of strings from the uploaded data. Then, our team sorted and mapped the strings to the AAT manually. Finally, we used a script to look up the URIs of the AAT terms and generate the JSON file necessary for ARIADNEplus.

ARIADNEplus originally focused on loose one-on-one mappings for objects[5], but because the extracted Japanese archaeological terms were mostly compound terms, we chose to employ one-to-many mappings instead. First, we broke down the terms into simpler concepts and mapped those to the AAT. Next, we mapped the results of the simpler concepts to the compound terms. This approach largely follows the usual mapping process of multilingual thesauri as outlined in the Guidelines for Multilingual Thesauri published by the IFLA[6].

One difficulty in this approach was that the integration pipeline required declaring the SKOS mapping property[7] between each link. We solved this by semi-automatically generating the properties depending on two simple criteria: the length of a term to be matched and the placements of the simple terms inside a given compound term. Essentially, if a simpler term, e.g., *kagami* (mirror), is inside a longer term, e.g., *dōkyō* (bronze mirror), then the simpler concept is either a broader or a related concept to the longer term. Whether it is a broader or a related concept can be safely judged based on the place of the simple term inside the compound term, i.e., if the simple term comes at the end of the compound term, it is a broader term. If it is located anywhere else, then it is a related term. For example, in the case of *dōkyō, kagami* is a broader (more general) concept, while *dō* (bronze) is a related one. This is because in Japanese, similarly to English, the last component in a compound word or term identifies the general concept to which the whole word refers to.

In cases where this approach proved lacking, we manually linked further terms to the Japanese ones. For example, we have augmented *sekka* (stone replicas of ceremonial bronze halberds) with “ritual objects” after mapping it to “rock (inorganic material)” and “ge (ceremonial knives).”

A further challenge we had faced in the aggregation process was how to generate meaningful names for each archaeological intervention in bulk. Our solution was to create new titles by combining the romanized names of the sites with descriptive English terms and dates referring to the time of excavations. For example, we generated names like “Nambori Shell Midden: 19840801-19850325” or “Shimotsuke Provincial Temple: 19850701-19851101.”

V Conclusion

Integrating SORAN’s data into the ARIADNE Catalogue was time-consuming and difficult. However, now a large part of Japan’s archaeological dataset is included in an extensive international dataset searchable and processable via a common user interface through the ARIADNE Portal. The integration not only improved the findability of SORAN’s data, but because of the transformations the data went through, it also made it easier to manipulate it inside an ever-growing global dataset.

Notes

[1] Franco Niccolucci and Julian Richards, “ARIADNE and ARIADNEplus,” *in The ARIADNE Impact,* ed. Franco Niccolucci and Julian Richards (Budapest: Archaeolingua Foundation, 2019), 7. https://doi.org/10.5281/zenodo.3476711.

[2] PeriodO (website), accessed November 25, 2021, https://perio.do/en/.

[3] The definitions are available at http://n2t.net/ark:/99152/p0fbfth.

[4] Art & Architecture Thesaurus (website), accessed July 28, 2022, https://www.getty.edu/research/tools/vocabularies/aat/.

[5] Ceri Binding and Douglas Tudhope, “Multilingual Vocabulary Mapping in ARIADNEplus” (PowerPoint presentation, 19th European Networked Knowledge Organization Systems [NKOS] Workshop, Oslo, September 12, 2019.) https://nkos-eu.github.io/2019/content/NKOS2019-presentation-tudhope.pdf

[6] IFLA Working Group on Guidelines for Multilingual Thesauri, *Guidelines for Multilingual Thesauri* (The Hague: IFLA Classification and Indexing Section, 2009).

[7] For details, see SKOS Simple Knowledge Organization System
Reference (website), https://www.w3.org/TR/skos-reference/#mapping.