Following the Thread

Integrating SORAN's Dataset Into ARIADNEplus

Yuichi TAKATA, Nara National Research Institute for Cultural Properties, Japan
Peter YANASE, Nara National Research Institute for Cultural Properties, Japan

Keywords: *Japan — fieldwork reports — thesauri mapping*

CHNT Reference: Takata, Y., Yanase, P. (2022). 'Following the Thread: Integrating SORAN's Dataset Into ARIADNEplus', in CHNT Editorial board. *Proceedings of the 27th International Conference on Cultural Heritage and New Technologies, November 2022.* Heidelberg: Propylaeum.
DOI: xxxxxxx.

The Comprehensive Database of Archaeological Site Reports in Japan (SORAN) (see fig 1) 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 state-level research institutes in the country 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 as full-text PDFs. The metadata stored comes from various sources, of which the datasheets attached to fieldwork reports published post-1994 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.

Fig. 1. The top page of SORAN.

In 2017, NABUNKEN was invited to join the ARIADNEplus data infrastructure, the successor of ARIADNE, a project aiming to overcome the fragmentation of archaeological data repositories. NABUNKEN accepted the offer and decided to integrate a significant part of SORAN's data into ARIADNEplus to improve its data's findability and enrich the international dataset.

The aggregated metadata of the ARIADNEplus project partners is stored in the ARIADNE Catalogue. The Catalogue is searchable, via the ARIADNE Portal, according to the three facets of "where" (space), "when" (time), "what" (object), as well as keywords drawn from controlled vocabularies. While SORAN supports information retrieval in a similar manner, the way relevant information is implemented and presented differs greatly from ARIADNEplus. Therefore, NABUNKEN and ARIADNEplus had to collaborate closely in a lengthy integration process involving data cleansing, schema transforming, and concept mapping.

Mapping SORAN's internal data schema to ARIADNE's ontology was a largely technical step. Although the two schemas differ in concept and format, the mapping could be done in a few weeks. Mapping the Japanese data to the aforementioned facets proved to be 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. These problems were solved with a combination of scripts and manual intervention.

The second facet required temporal information to be linked to definitions stored on PeriodO (a multilingual gazetteer of temporal information). As a first step, a controlled vocabulary for the time periods was established. Next, all past entries in the database were converted to conform with the new vocabulary. This was a very resource-heavy task as most entries had to be manually disambiguated. After the conversion of past data was completed, SORAN's interface was altered to only accept entries from the controlled vocabulary moving forward. However, there was a further obstacle in the way of integration: no single authoritative source covered all the time periods used by SORAN. To solve this, NABUNKEN arranged an extended discussion of the possible definitions among its interdisciplinary team of experts. The results were published in 2022 and then registered in PeriodO.

The final facet of objects required the most work as it involved mapping culture- and discipline-bound terms to the Getty Art & Architecture Thesaurus. Similar to the temporal entries, the data entered in the relevant field in the database for excavated materials were eclectic and contained many typos. However, because of the internal policy of the SORAN, these entries could not be cleansed in the same way as the temporal information and a different approach was taken instead. First, a list of terms was generated based on the uploaded information. Next, these terms were sorted and mapped to the AAT manually. Because most Japanese archaeological terms are compound, in order to keep as much information as possible, the terms were divided into main terms and sub-terms. Main terms define the function of objects, while the sub-terms define their properties. In cases where even this proved lacking, further terms were linked to the Japaneseterms. This approach was inspired by the mapping process of multilingual thesauri and was peculiar to the Japanese dataset. In a further step, the mappings were analyzed and harmonized to make them as consistent as possible to ease the way for future additions to the list of terms.

The collaboration with ARIADNEplus proved to be more than just a simple integration of data: it provided NABUNKEN with an opportunity to learn about good practices of data stewardship and reflect on the nature and structure of the data stored in SORAN. This presentation aimed to present this long but educational road leading to the final aggregation of SORAN's data into ARIADNEplus.

Funding

ARIADNEplus is a project funded by the European Commission under the H2020 Programme, contract no. H2020-INFRAIA-2018-1-823914.

Author Contributions

Writing – original draft: Yuichi TAKATA, Peter YANASE
Writing – review & editing: Yuichi TAKATA, Peter YANASE

References

Comprehensive Database of Archaeological Site Reports Japan. Available at https://sitereports.nabunken.go.jp/en (Accessed: 25 June 2022).

Niccolucci, F. and Richards, J. (2019). ARIADNE and ARIADNEplus, *The ARIADNE Impact*, Budapest, ARCHAEOLINGUA FOUNDATION pp. 7–26. DOI: 10.5281/zenodo.3476711

PeriodO – periods, organized. Available at https://perio.do/en/ (Accessed: 25 June 2022).