This PDF is a simplified version of the original article published in Internet Archaeology. Enlarged images which support this publication can be found in the original version online. All links also go to the online version.

Please cite this as: Takata, Y. and Yanase, P. 2023 Following the Thread: Integrating SORAN's Japanese Dataset into ARIADNE, Internet Archaeology 64. https://doi.org/10.11141/ia.64.15

Following the Thread: Integrating SORAN's Japanese Dataset into ARIADNE

Yuichi Takata and Peter Yanase

The Comprehensive Database of Archaeological Site Reports in Japan (SORAN) is an online index of domestic archaeological excavations operated by the Nara National Research Institute for Cultural Properties (NABUNKEN). SORAN emerged as a response to the need to improve the findability and accessibility of Japanese archaeological grey literature and the information contained therein. NABUNKEN joined the ARIADNEplus project in 2019 and finished integrating SORAN's metadata into the ARIADNE Catalogue in 2022. In this article we give a short overview of how archaeological data, especially fieldwork reports, are produced in Japan. Next, we summarise the history of SORAN and the nature of its dataset. Finally, we explain the steps taken to transform the Japanese dataset to allow its integration in the ARIADNE Catalogue.

1. Introduction

The Nara National Research Institute for Cultural Properties (NABUNKEN) was established in 1952 as an auxiliary organisation under the National Commission for Protection of Cultural Properties (later the Agency for Cultural Affairs) for the *in situ* study of movable and immovable cultural properties in the Nara region. It is a member of the National Institutes for Cultural Heritage, an umbrella entity formed in 2007 comprising four museums and three research institutes. NABUNKEN is in charge of leading and educating the Japanese community of cultural heritage experts, especially, but not limited to, the archaeological community. The institute itself is an active producer of archaeological data as it has been involved in countless excavations in and outside of Japan since its inception. It also manages three museums showcasing its finds and constantly strives to improve current methods of preservation, conservation, restoration, presentation, and maintenance of cultural properties (Nara Bunkazai Kenkyujō 2022).

ARIADNEplus was a Horizon 2020 project running between 2019 and 2022, funded by the European Commission. It was the successor to ARIADNE (Archaeological Research Infrastructure for Archaeological Data Networking in Europe), whose goal was 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' (Niccolucci and Richards 2019, 7). 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 (see Introduction, Richards *et al*. 2023).

The Comprehensive Database of Archaeological Site Reports in Japan (SORAN) is an online index of domestic archaeological excavations operated by NABUNKEN. SORAN emerged as a response to the need to improve the findability and accessibility of Japanese archaeological grey literature and the information contained therein. In 2008, five national university libraries located on the western edge of Japan, with Shimane University Library taking the lead, formed an alliance to publish the full texts of local archaeological fieldwork reports on the Internet. This endeavour eventually morphed into a nationwide project with twenty-one national university libraries participating, called the *Zenkoku iseki shiryō ripojitori purojekuto* [Nationwide Archaeological Site Information Repository Project]. During the lifetime of this project, nearly twelve thousand reports were digitised and made freely accessible online. Although the project was a joint effort, the prefectural datasets were managed separately. It was not until NABUNKEN took over management duties in June 2015 that the accumulated data was merged into SORAN's monolithic database (Maizō Bunkazai Hakkutsu Chōsa Taisei Tō No Seibi Jūjitsu Ni Kansuru Chōsa Kenkyū Iinkai 2017). In the following years SORAN grew considerably, with the help of an increasing number of data providers from across Japan. In November 2022, the catalogue contained information on about 140,000 archaeological interventions and 125,000 fieldwork reports, of which more than 32,000 were downloadable in PDF format (Takata 2023).

Historically, SORAN was built to satisfy the needs of Japanese researchers; however, two brief meetings in 2017 and 2018 between the core team managing the database and Julian Richards, Director of the Archaeology Data Service, brought profound changes to the project's direction. During these meetings, NABUNKEN learnt much about international best practices for handling archaeological data, especially those described in the ADS Guides to Good Practice. The conversations led to NABUNKEN joining the communities built around the ARIADNEplus and SEADDA (Saving European Archaeology from the Digital Dark Ages) projects and the eventual aggregation of SORAN's metadata into the ARIADNE infrastructure in 2022. NABUNKEN also adopted several international best practices of data stewardship promoted by these projects and published a Japanese translation of the ADS Guides to Good Practice in 2022.

In this article, we will first give a short overview of how archaeological data, especially fieldwork reports, are produced in Japan. Next, we will summarise the history of SORAN and the nature of the dataset. Finally, we will explain the steps we have taken to transform the Japanese dataset to allow its integration into the ARIADNE Catalogue.

2. Cataloguing Grey Literature

Archaeological information produced in Japan derives mainly from rescue excavations undertaken in advance of development projects for highways, railways, or housing areas. The number of annual rescue excavations over the last twenty years has fluctuated between 7000 and 9000. The number is this high because two-thirds of Japan's landmass is covered with forests and mountains, resulting in past and present settlements concentrating on the remaining flatland and coastal areas. There are around 460,000 registered archaeological sites, meaning that, on average, every populated square kilometre in Japan contains more than three sites.

The government in Japan is decentralised; local municipalities are in charge of local matters, including archaeological investigations. The details of the central governmental policy defining the requirements for rescue excavations are laid out in article nos 93–95 of the Law for the Protection of Cultural Properties (enacted in 1950, last amended in 2022). The law states that local governments should locate and register archaeologically relevant sites and strive to preserve them in their original form. However, if unavoidable, sites can be destroyed or altered. In such cases, the authorities should request thorough pre-construction excavations for the retrieval of all relevant artefacts and information. This act is often referred to with the oxymoron 'preservation by record' (*kiroku hozon*).

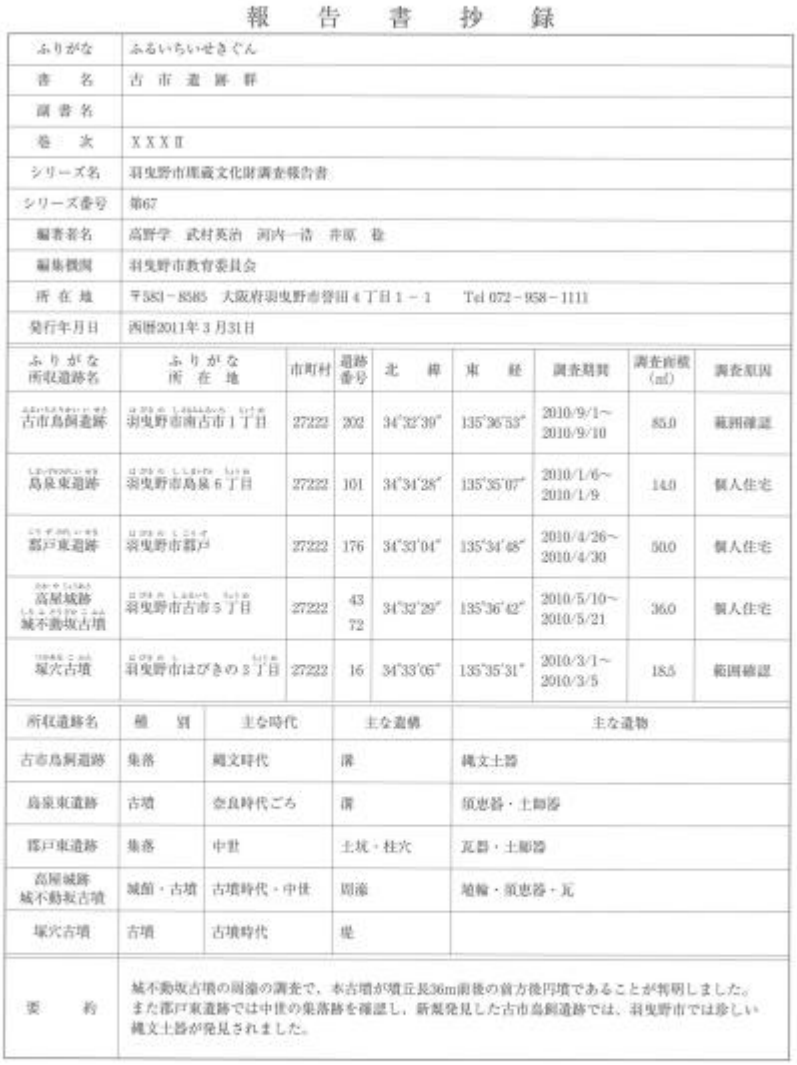
Two types of physical data emerge from excavations: 1) raw data, such as illustrations, photographs, measurements, and notes, and 2) the fieldwork report. The preservation and dissemination of raw data still needs to be solved in Japan. Fieldwork reports are better managed as they are considered the culmination of excavations. This is most clearly seen in the fact that developers' financial responsibilities are considered finished once the reports are published. As such, the reports are usually handled with care. Currently, the Japanese government advises that 300 hard copies should be dispersed throughout Japan to prevent the loss of information and promote the reports' contents.

Because of the decentralised nature of the Japanese administration, the production and dissemination of archaeological information are not managed coherently. As a result, neither the extent of archaeological grey literature nor the extent to which it has been published is easily identifiable in Japan. To improve this situation, the Agency for Cultural Affairs released a governmental report in 2004 in which they implicitly asked the local municipalities to provide bibliographical information to SORAN (Takata and Yanase 2021). Lately, the Agency is sending out more explicitly phrased annual notices reminding governments that it expects them to provide the necessary data. Based on the provided information and NABUNKEN's extensive collection of grey literature, NABUNKEN managed to gather bibliographical information for more than 100,000 reports. Catalogues are planned to be released through 2023 and 2024.

3. SORAN's Source of Metadata

Bibliographical data in itself is not very useful when it comes to archaeological fieldwork reports because it does not convey much beyond the names and types of the sites. To solve this problem, the Agency of Cultural Affairs started requesting the local governments to attach semi-structured datasheets to the reports in 1994 (Morimoto 2017). These sheets should contain information on every archaeological intervention covered in a given fieldwork report and record the names, addresses, coordinates, sizes, types, and ages of the sites excavated, the dates and reasons for the excavations, and lists of the types of structural remains and artefacts found (Figure 1). NABUNKEN started aggregating the information from these datasheets into a public electronic database in 2003. This database was merged into SORAN in 2019 and now provides the core metadata of the dataset.

From 2019 onward, new information is directly uploaded to SORAN by its users via a WEB interface. To generate metadata systematically for pre-1994 reports, the Agency of Cultural Affairs sends an annual notice asking all relevant parties to prepare data for a specific year. For example, in 2023, the Agency requested metadata be created for reports published in 1973. Newly prepared information for pre-1994 reports is not attached to the hard copies but exists only inside SORAN.



報　告　書　抄　録

ふりがな	ふるいちいせきぐん
書　名	古 市 遺 跡 群
副 書 名	
巻　次	ＸＸＸⅡ
シリーズ名	羽曳野市埋蔵文化財調査報告書
シリーズ番号	第67
編著者名	高野学　武村英治　河内一浩　井原　稔
編集機関	羽曳野市教育委員会
所 在 地	〒583－8585　大阪府羽曳野市誉田４丁目１－１　Tel 072－958－1111
発行年月日	西暦2011年３月31日

ふりがな 所収遺跡名	ふりがな 所　在　地	市町村	遺跡番号	北　緯	東　経	調査期間	調査面積（㎡）	調査原因
古市鳥飼遺跡	羽曳野市南古市１丁目	27222	202	34°32'39"	135°36'53"	2010/9/1～2010/9/10	85.0	範囲確認
島泉東遺跡	羽曳野市島泉６丁目	27222	101	34°34'28"	135°35'07"	2010/1/6～2010/1/9	14.0	個人住宅
郡戸東遺跡	羽曳野市郡戸	27222	176	34°33'04"	135°34'48"	2010/4/26～2010/4/30	50.0	個人住宅
高屋城跡 城不動坂古墳	羽曳野市古市５丁目	27222	43 72	34°32'29"	135°36'42"	2010/5/10～2010/5/21	36.0	個人住宅
塚穴古墳	羽曳野市はびきの３丁目	27222	16	34°33'05"	135°35'31"	2010/3/1～2010/3/5	18.5	範囲確認

所収遺跡名	種　別	主な時代	主な遺構	主な遺物
古市鳥飼遺跡	集落	縄文時代	溝	縄文土器
島泉東遺跡	古墳	奈良時代ごろ	溝	須恵器・土師器
郡戸東遺跡	集落	中世	土坑・柱穴	瓦器・土師器
高屋城跡 城不動坂古墳	城館・古墳	古墳時代・中世	周濠	埴輪・須恵器・瓦
塚穴古墳	古墳	古墳時代	墓	

要　約	城不動坂古墳の周濠の調査で，本古墳が墳丘長36m前後の前方後円墳であることが判明しました。また郡戸東遺跡では中世の集落跡を確認し，新規発見した古市鳥飼遺跡では，羽曳野市では珍しい縄文土器が発見されました。

Figure 1: A typical datasheet attached to fieldwork reports (Source: Habikino Shi Kyōiku Iinkai 2011)

4. The Impact of SORAN

In 2022, SORAN had over 117 million pageviews, and visitors downloaded over two million PDF files (internal statistics, for public numbers up until 2021, see Takata 2023). The service's popularity is evident, but it is not easy to assess its exact impact on the scholarly community because authors of academic publications

rarely clearly state how they got their resources. On the other hand, there are many reports on the positive effects of SORAN used in non-academic contexts.

In a recent online round-table discussion held by NABUNKEN, university students reported that being able to access full-text versions of grey literature significantly lessened their financial burdens. Fieldwork reports cannot be bought; therefore, the only way to access them is to either visit a place having a copy or borrow it via interlibrary loan. Both of these solutions are costly, require applying for permissions, and limit the time one can spend reading a report. On top of that, without having access to the full text, it is difficult to judge whether one truly needs a given report, which can result in unnecessary expense. Students also report that having been able to access the reports online during the height of the COVID-19 pandemic, when physical libraries and reading rooms were closed, allowed them to continue their research without significant setbacks (Hayashi *et al*. 2023).

Grey literature from SORAN is also being used as reference material in Wikipedia articles, thus helping the spread of freely accessible knowledge. There are reports of local governments experimenting with Wikipedia as a means to promote local heritage by producing relevant articles as well (Ichikawa *et al*. 2022; Miyoshi *et al*. 2022). To promote such usage, SORAN has a built-in help function that generates citations and links in the format used by Wikipedia.

Fieldwork reports are increasingly used in compulsory education too.

Previously, the limited circulation of the reports prevented schools from using reports in such a manner. However, with universal access, it is now possible for all the students to examine the same text simultaneously or access the necessary literature from home.

Experiments show that, given proper instructions, even elementary school children are able to extract information from the seemingly inaccessible, highly academic literature (Miyazawa 2022; Figure 2).

Figure 2: Elementary school children using SORAN to learn about local heritage (Source: Miyazawa 2022)

5. 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. To make this possible, project partners were asked to make their metadata available so it could be collected, transformed into the ARIADNEplus data model (AO-Cat), stored on a triple store based on GraphDB technology, and enriched with links to PeriodO and the Getty Art & Architecture Thesaurus (AAT) (Richards 2023). Practically, this meant we had to prepare our data in XML, provide a mapping of the local schema to the AO-Cat, prepare mappings between the local vocabulary and the Getty AAT, and create a PeriodO dataset. Unfortunately, the metadata in SORAN was neither based on controlled vocabularies nor available in XML format. Therefore, to integrate SORAN's data into the ARIADNE Catalogue, we first had to normalise and cleanse the dataset.

Spatial coordinates required the most work. Coordinates registered in SORAN are based on four different datums: a) Tokyo Datum, used between the end of the 19th century up until 2002; b) JGD2000, officially used between 2002 and 2011; c) JGD2011, the latest datum adjusted to reflect the geographical changes brought by the 2011 Tōhoku earthquake; and d) WGS84. Coordinates fed into SORAN are automatically converted to WGS84. However, many of the coordinates were mistyped because most of the metadata were created without the now-available restrictive WEB interface. As preparation for the aggregation, we corrected the wrong coordinates with a combination of scripts and manual intervention. First, we flagged all the sites that a) had coordinates in the sea but were not underwater sites or b) were located outside the jurisdiction of the entity performing the excavation.

After that, our team deduced the sites' coordinates based on the datasheets, descriptions, and maps found within the reports (Figure 3).

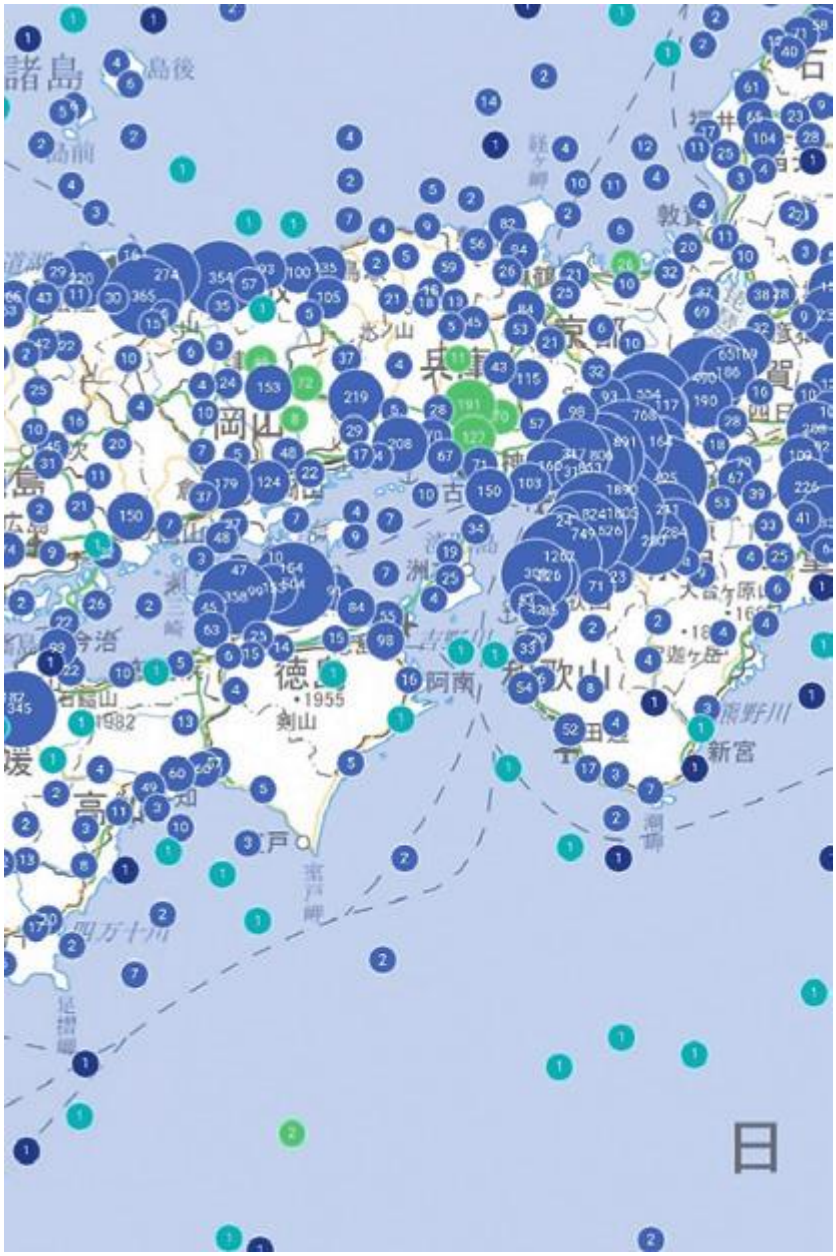


Figure 3: Coordinates registered in SORAN before normalisation shown on a map (Source: Takata 2020)

The datasheets attached to the reports place no restriction on what can be written in a given field. This is true for the cell containing the dates of the sites as well. However, in the ARIADNE Catalogue, temporal information must be linked to definitions published in PeriodO, a multilingual gazetteer of temporal information. SORAN originally had no controlled vocabulary for named periods, but in 2019 we established one. After that, 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 were further obstacles to be surmounted before we could prepare a PeriodO dataset: the named periods had no definitions in absolute dates, and there was not a single authoritative

source that covered all the required periods. To solve this, we arranged an extended discussion of the possible definitions among NABUNKEN's interdisciplinary team of experts. Afterwards, we compiled the results of the discussions in an internal document and then registered the new definitions in PeriodO.

Mapping the culture- and discipline-bound terms found in SORAN to the Getty AAT was the most challenging part of the metadata integration process. As with the other fields, information entered into the system about excavated artefacts was eclectic and contained many typographic errors. However, we opted not to cleanse the data for this field in order 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 Python 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 (Binding and Tudhope 2019). However, because the extracted Japanese archaeological terms were mostly compound terms and the Japanese terminology was underrepresented in the AAT, after consulting with our colleagues at ARIADNE, 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 IFLA Working Group on Guidelines for Multilingual Thesauri (2009).

One difficulty in this approach was that the integration pipeline required declaring the SKOS mapping property 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, in common with English, the last component in a compound word or term identifies the general concept to which the whole word refers.

In cases where this approach proved insufficient, 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)'.

We are glad we were allowed to take such a pragmatic approach because it was important to us to make our mapping as reproducible as possible so we could revise and reuse it later. Some redundancy and slight inaccuracy in the mapping were sacrifices we were willing to pay.

A further challenge we faced in the aggregation process was generating meaningful names for each archaeological intervention en masse. Our solution was to create

new titles by combining the Romanised 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'.

6. Conclusion

Integrating SORAN's metadata into the ARIADNE Catalogue was a lengthy and difficult process. This was primarily because the ARIADNEplus project's lifetime overlapped with the maturing phase of SORAN. SORAN began in 2015 but only reached its maturity in 2019 when several legacy databases were merged with it, after which a long phase of data normalisation followed. Therefore, many of the guidelines we had to follow to allow the integration of SORAN's metadata into the ARIADNE Catalogue directly affected how we thought about data. This made our involvement in the project a learning process.

Integrating SORAN's metadata into ARIADNE not only improved the findability and accessibility of SORAN's data internationally, but because of the transformations the metadata went through, it has also made it significantly easier to manipulate the data both in and outside the ARIADNE Portal. However, integrating the metadata of the Japanese fieldwork reports into ARIADNE was only the first step. SORAN will keep on evolving. For historical reasons, Japanese archaeology has been focusing on producing printed fieldwork reports. However, for a truly data-driven archaeology, we need to break out from the confines of this format. To further improve the usability of the results of Japanese archaeology, SORAN's repository needs to be able to extract and serve more granular data in the future. We also need to learn how to accommodate new and emerging data formats used in archaeology, such as GIS and 3D data.

Producing digital data becomes easier by the day. Creating highly accurate 3D scans of artefacts with a smartphone has recently become a reality. With lowered technical barriers, the amount of data will keep on growing exponentially. With each passing day, there will be more data and more kinds of data in archaeology. Such challenges cannot be effectively solved alone. The ARIADNE Portal might be the most visible part of the ARIADNEplus project, but its community-building and knowledge-sharing activities were just as important, if not more so. The integration of SORAN's metadata into the ARIADNE Catalogue was a Herculean effort made possible only because of the ARIADNE community; always ready to provide advice and help to its members. We look forward to the next phase of ARIADNE, in which we can continue to tackle the problems of archaeological data production and stewardship with returning and new members of the community.

Bibliography

Binding, C. and Tudhope, D. 2019 '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

Habikino Shi Kyōiku Iinkai 2011 *Furuichi isekigun* **32**, Habikino: Habikino Shi Kyōiku Iinkai. https://doi.org/10.24484/sitereports.17334

Hayashi, R. *et al*. 2023 'Gakusei zadankai: Korona wazawai wa, gakusei no bunken shūshū katsudō ni dō eikyō wo ataetaka? Jisedai no chōsa kenkyū kankyō no arikata wo kangaeru', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **5**. https://doi.org/10.24484/sitereports.130529

Ichikawa, H. *et al*. 2022 'Shizuoka ken Namazu shi ni okeru Wikipedia Town no jissenrei', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **4**. https://doi.org/10.24484/sitereports.115736

IFLA Working Group on Guidelines for Multilingual Thesauri 2009 *Guidelines for Multilingual Thesauri*, The Hague: IFLA Classification and Indexing Section.

Maizō Bunkazai Hakkutsu Chōsa Taisei Tō No Seibi Jūjitsu Ni Kansuru Chōsa Kenkyū Iinkai 2017 *Maizō bunkazai hogo gyōsei ni okeru dejitaru gijutsu no dōnyū ni tsuite 2 (hōkoku)*, Tokyo: Agency for Cultural Affairs. https://doi.org/10.24484/sitereports.71613

Miyazawa, Y. 2022 'Gakkō toshokan kakeru GIGA sukūru kakeru chiiki bunkazai shiryō', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **4**. https://doi.org/10.24484/sitereports.115736

Miyoshi, S. *et al*. 2022 'Bunkazai kakeru Wikipedia, wakugumi to jissen', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **4**. https://doi.org/10.24484/sitereports.115736

Morimoto, S. 2017 *Iseki Jōhō Kōkan Hyōjun no Kenkyū 4*, Nara: Nara National Research Institute for Cultural Properties.

Nara Bunkazai Kenkyujō 2022 *Nara Bunkazai Kenkyujō gaiyō 2022*, Nara: Nara Bunkazai Kenkyujō. https://doi.org/10.24484/sitereports.130391

Niccolucci, F. and Richards, J. 2019 'ARIADNE and ARIADNEplus' in J. Richards and F. Niccolucci (eds) *The ARIADNE Impact*, Budapest: Archaeolingua Foundation. 7–25. https://zenodo.org/record/4319058

Richards, J.D. 2023 'Joined up Thinking: Aggregating archaeological datasets at an international scale', *Internet Archaeology* **64**. https://doi.org/10.11141/ia.64.3

Richards, J.D., Aspöck, E. and Niccolucci, F. 2023 'Introduction', Internet Archaeology 64. https://doi.org/10.11141/ia.64.1

Takata, Y. 2020 'Iseki shōroku no genjō to chūiten', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **2**. https://doi.org/10.24484/sitereports.69974

Takata, Y. 2023 '2022 nendo sūji de miru zenkoku iseki hōkoku sōran', *Dejitaru gijutsu ni yoru bunkazai jōhō no kiroku to rikatsuyō* **5**. https://doi.org/10.24484/sitereports.130529

Takata, Y. and Yanase, P. 2021 'The production, preservation and dissemination of archaeological data in Japan', Internet Archaeology **58**. https://doi.org/10.11141/ia.58.11