Database Framework for Small-Scale Datasets: Case Studies on Field Survey Data in Southeast Asia

Kimiya Kitani[1], Toshiro Kamiya[2]

1. Abstract

The data collected in field research of area study surveys by a single researcher is usually a few hundred to a few thousand. This amount is too small to conduct analytical research using the latest digital technologies, such as big data. On the other hand, it is too costly to build and maintain a database for such a small amount of data. Even if one tries to extract data from a database developed in the past and reuse it or migrate to a new database, what has been built will quickly become unusable unless it is sustainable to use in the first place. Thus, many challenges must be overcome before valuable small-scale data can be fully utilized and publicly available. In this presentation, we illustrate some case studies of building a database of small data sets by using a general cloud system.

2. Introduction

The Center for Southeast Asian Studies, Kyoto University, has a set of academic databases called “MyDatabase” (https://kyoto.cseas.kyoto-u.ac.jp/database/). This platform project was led by Prof. Shoichiro Hara, a professor emeritus of CSEAS [1]. “MyDatabase” is an outstanding academic achievement. However, the aim is to integrate various databases and to share information resources from different fields. So, a dataset must be modified or mapped according to specific rules from the original format on which the researcher depends. Moreover, the system maintenance costs are expensive, and organizing data is not easy, taking time and effort. On the other hand, even if research materials are not fully arranged under strict rules, they may get a chance for new collaboration or new projects in some research by making them publicly available.

Eventually, it would be desirable to share the data with academic databases. However, as a first step, we consider constructing a small-scale, easy-to-develop database system worthwhile, so we started this project.

During the research project, we faced two issues:

i. **The difficulty of conducting quantitative analysis in case the data is small**: In area study fieldwork, a researcher collects a maximum of several hundred or several thousand pieces of data, which is too small to conduct analytical surveys using the latest digital technologies such as Big Data. The small amount of data makes it difficult to decide what to analyze by

[1] Center for Southeast Asian Studies, Kyoto University. email: kitani@cseas.kyoto-u.ac.jp
[2] Kyoto Sangyo University. email: kamiya.tosirou.82w@st.kyoto-u.ac.jp

mechanical processing. In addition, morphological analysis is difficult when the data is multilingual.

ii. **Difficulties in constructing a sustainable database system:** Databases constructed for research projects require a specificity that does not exist elsewhere. Therefore, maintenance and management are expensive, as they require consideration of various aspects such as servers, operating systems, security measures, etc.

As in the above issues, even if the data is valuable as a research resource, raising the database development cost may not be possible. Also, if the project cannot be continued, the system will not be updated, there will be no one to maintain and manage the database, and it will eventually become a digital burden.

We have considered two approaches to building databases relatively easily to resolve these issues.

3. Approaches for the database framework for small-scale datasets

A "small-scale dataset" means the meta-information (item) is small. In our project, we are developing four small databases (Fig. 1). We will introduce the details of each of them later, but first, we will explain why we need them.

When developing a database using a large-scale system in the integrated platform, the rules for datasets become strict, partly because the goal is to share information resources. However, it is not easy to modify and map these to the platform's rules, especially in the case of unorganized field research datasets. As a result, it is likely that the datasets may not be made public after the publication of their own research findings.

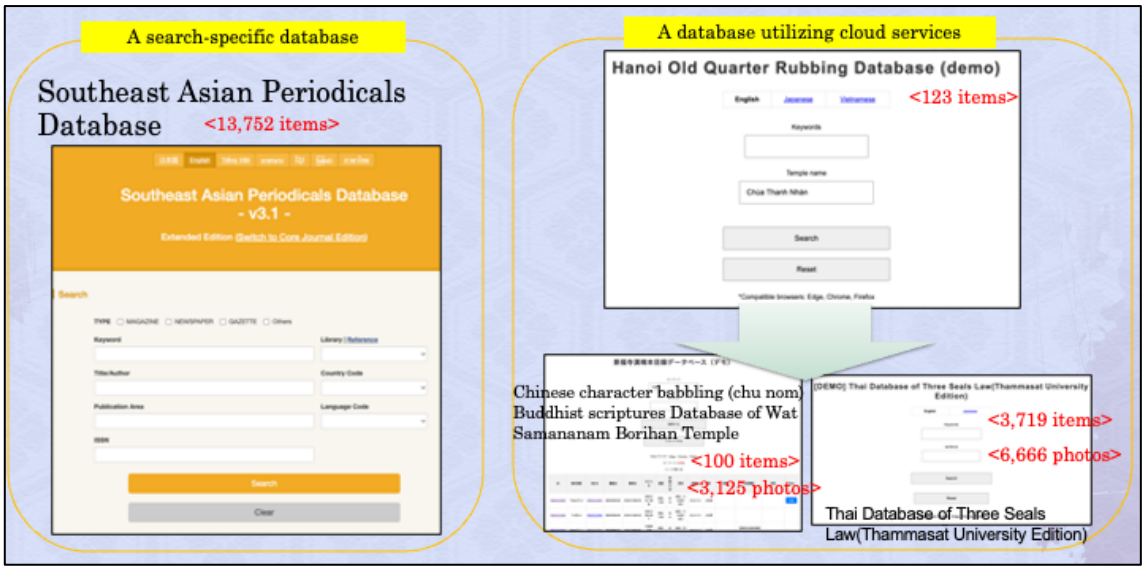

Fig. 1 Interface of the four small-scale databases we construct

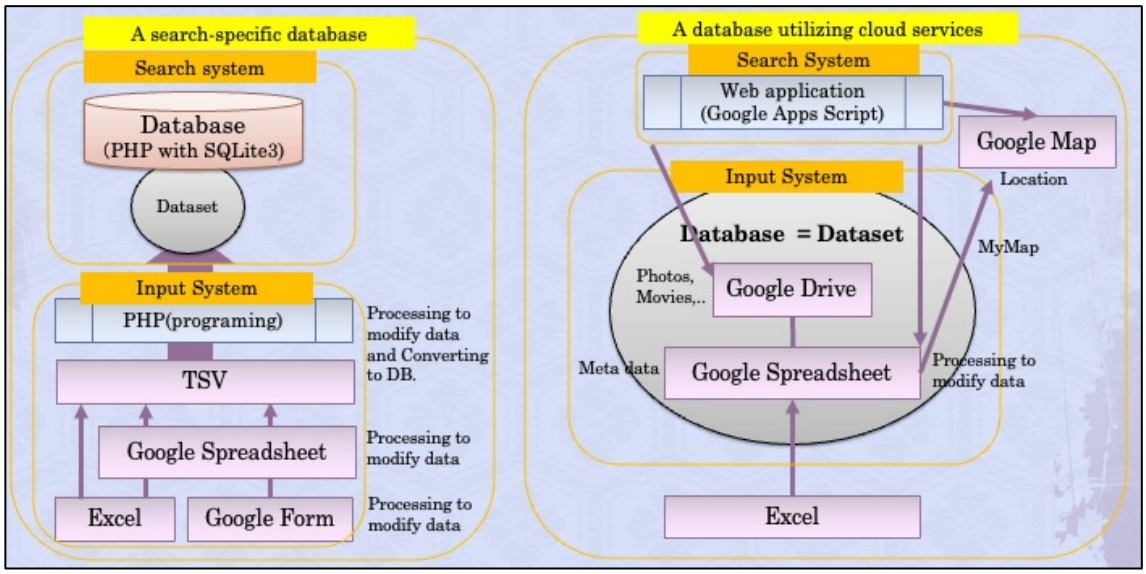

Fig. 2 Some approaches to the database framework

On the other hand, even if the research data is only partially organized, making it public may be useful for some studies, as mentioned earlier. Not only is it desirable to make data publicly available, but it may be mandatory to do so under an appropriate data management plan (DMP). Databases can be helpful in these cases, but it is difficult for an individual to build and manage a large-scale database, and it may be hard to use a large-scale database architecture for a data set containing only a few thousand data, which is also undesirable. The value here is building a specialized database system for small data sets.

1. **A Search-Specific Database:** Separating the input part from a general database can reduce the costs of security measures, user registration, permissions, and data checks (Fig. 3). To accommodate a variety of data sets, we provide flexibility by preparing configuration files for

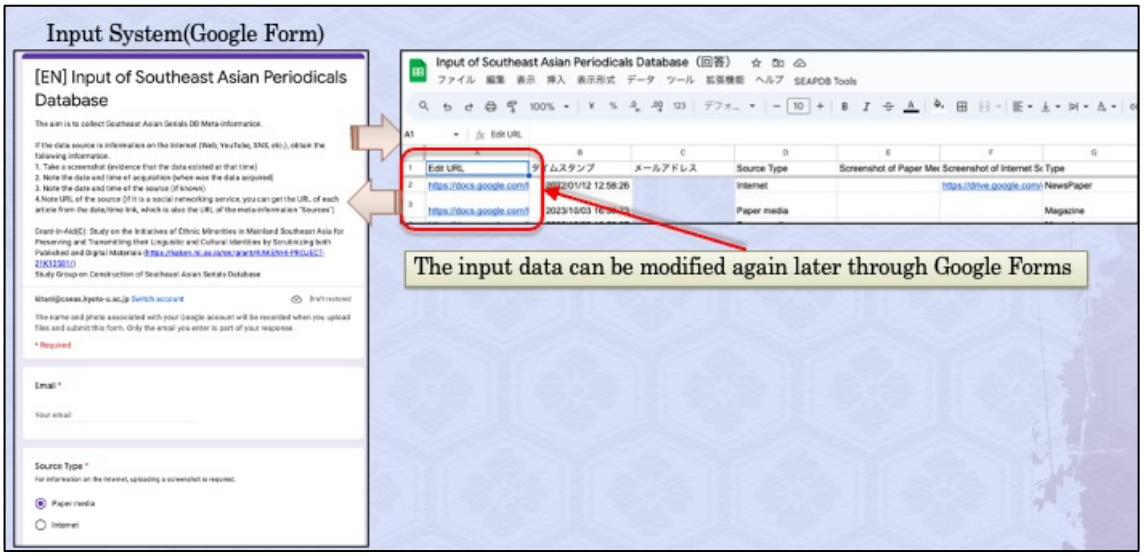

Fig. 3 Input tool using Google Form

each search target and item. However, it is inconvenient that the user can import the data only in bulk and cannot freely enter or modify them. When using Google Forms, the user can use the re-edit function. Still, there is another problem with directly editing data in Google Spreadsheets, which will cause it to become inconsistent with Google Forms.

2. **Database Utilizing Cloud Services:** The approach of databases utilizing cloud services is a system whereby meta-information is made into a dataset that utilizes Google Spreadsheet (Fig. 4). This approach of making databases utilizing cloud services is a system whereby meta-information is made into a dataset that utilizes Google Spreadsheet, and image and video file material is made into a dataset that utilizes Google Drive. An application is developed that allows direct searching of the Google Spreadsheet. Data updates can be done by rewriting the

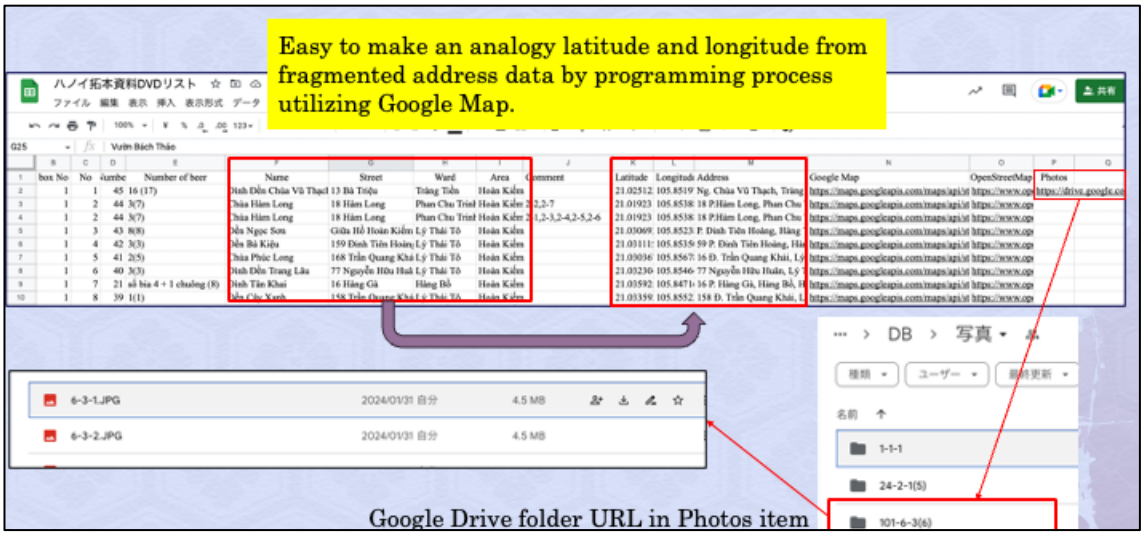

Fig. 4 Composition for a dataset utilizing Google cloud service

Google Spreadsheet or adding or changing data in Google Drive, and the processing power for images and videos can be left to the cloud side. Furthermore, a wide range of utilization is possible, such as links to Google Maps from meta-information, and the data set itself can be published externally or in a limited manner. It is easily possible to make an analogy of latitude and longitude from fragmented address data by programming processes utilizing Google Maps. In addition, photos, videos, and other documents are stored on Google Drive so they can be easily added or deleted. Of course, inferring GPS information from incomplete information will only result in inaccurate data, so in the end, an on-site investigation will be necessary. However, it can still provide a visual clue to the approximate location.

4. Performance comparison of the approaches

Comparing these two approaches (Fig. 5), the ‘database utilizing cloud services’ isn’t easy to

	A search-specific database	A database utilizing cloud services
Processing Speed	High (PHP)	Slow (Google Apps Script)
Use Restriction	Not particular	Script runtime: 6 minutes (*1)
Display images using <img> tag	No restrictions	Required: signed in to a Google account.
Import the data to the database	Only developer can be imported.	No need
Ease of applying to other datasets	Bit difficult	Easy

Fig. 5 Performance comparison of the approaches

search and modify data by program processing on a large scale, and implementation must be done with processing time in mind because Google Apps Script, based on JavaScript, is slow and has a 6-minute execution time limit. Also, the disadvantage of this approach is that it is susceptible to changes in specifications, for example. It became required for users to sign in with a Google account in January 2024 when images on Google Drive are displayed using <img> tag in the web application. As such, the specifications are subject to change at any time.

As for application to other data sets, since it is a database that references Google Spreadsheets, it is easier to access databases using cloud services. In addition, when it comes to processing and modifying data, there is the advantage that large amounts of data can be processed quickly by making full use of SQL, a database language that can be used in Google Spreadsheets.

In summary, the approach of using a search-specific database has the advantage of allowing detailed modification and high-speed searching. For how far the database approach utilizing the cloud services can withstand data, we investigated the use of the “Historical Place Name Data Dictionary” (Historical Place Name Data Dictionary), the dataset of 298,913 rows, provided by GeoNLP (https://geonlp.ex.nii.ac.jp/dictionary/)[2]. As a result, a full-text search took about 5 seconds for 10,000 records, about 30 seconds for 100,000 records, and about 40 seconds for about 298,913 records. So, we consider that the number of datasets should be less than 10,000.

5. Four Case Studies of Small-scale Databases

Below, we will introduce four concrete examples of how these techniques are used to build small databases.

■ **Case 1: Southeast Asian Periodicals Database** (Fig. 6, 7): The first database dataset released in 2011 contained 3,045 bibliographic data items. During several projects since then, data on Indochina countries and data on ethnic minorities in mainland Southeast Asia were gradually added. For a long time, the dataset remained small, but in September 2023, the incorporation of all periodicals held in the Library of Center for Southeast Asian Studies, Kyoto University and the Institute of Developing Economies Library was completed, and the meta-information (bibliographic information) contained in this database reached 13,752 items, which is no longer considered small. The meta-data items consisted of the NACSIS-CAT coding manual based on Anglo-American Cataloguing Rules II. NACSIS-CAT is a system for constructing union catalog databases designed to provide at-a-glance information on academic information archived at university libraries across Japan. As for the Title and Author, it adapted just a name that is easy for the public to understand, which means the Title

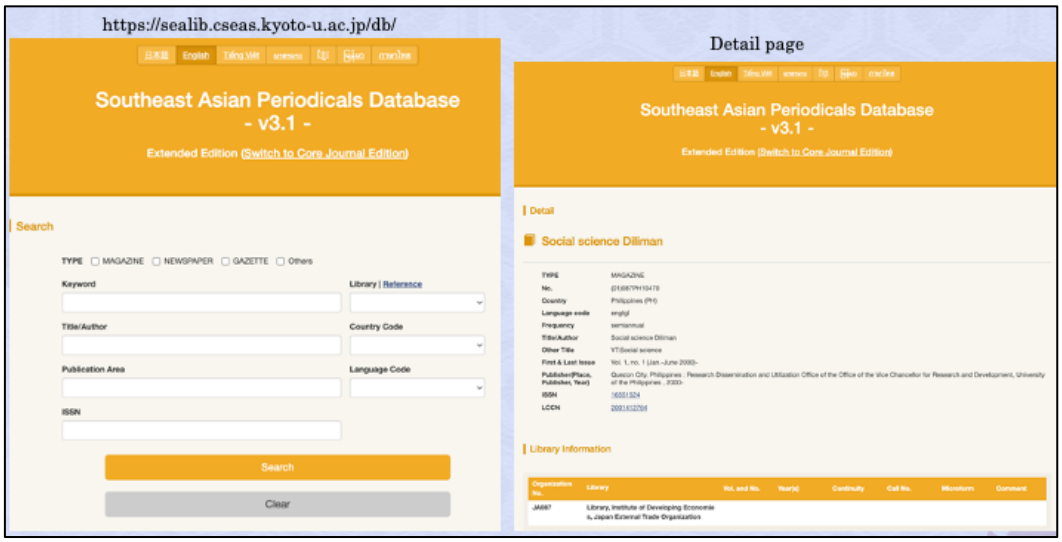

Fig. 6 User Interface of Southeast Asian Periodical Database

Specification		Status	Detail
Bibliographic information	**Type**	**Required**	Magazine, NewsPaper, Gazette, and so on.
	Country code	**Required**	Please comply with MARC21 Standard Code(https://www.loc.gov/marc/countries/countries_code.html).
	State/Region/Provice code	Optional	Not yet developed(Free style now) * To be developed in the future (to be unified at the state, provincial,
	Langauge code	**Required**	Please comply with the LC code (https://www.loc.gov/marc/languages/language_code.html). If there is
	Frequency	**Required**	Please follow the National Institute of Informatics (NII) Cataloging System Coding Manual
	Title/Author	**Required**	For non-Latin titles, use the LC romanization form. For languages not included in the LC romanization
	Title (Original Script)	Optional	In the case of non-Latin titles, the original spelling should be written here.
	Parallel Title	Optional	A title in another language that is annexed to the information source.
	Other Title	Optional	Titles not mentioned in the sources are included.Romanization methods other than the LC method,
	First & Last Issue	Optional	Adapt to NC labeling laws. *NC: http://catdoc.nii.ac.jp/MAN2/CM/7_2_7.html
	Title Change	Optional	Adapt to NC labeling laws. *NC: http://catdoc.nii.ac.jp/MAN2/CM/7_2_7.html
	Publication Area (Place of Publication,	**Required**	For non-Latin scripts, use the LC romanized form. *LC romanization tables:
	Publication Area (Original Script)	**Optional**	**In the case of non-Latin titles (*Required),** the original spelling should be recorded here. In the
	NOTE	Optional	Changes in publishers, frequency of publication, etc., will follow the description method of NC's
	Summary	Optional	
	Ethnicity	Optional	
	URL	Optional	Journal or editor/publisher's website.
	ISSN	Optional	
	NCID	Optional	National Institute of Informatics of bibliography number (Japan), if it exists.
	NDLID	Optional	National Diet Library of bibliography number (Japan), if it exists.
	LCCN	Optional	Library of Congress Control Number (LCCN), if it exists.
	Source	**Required**	Information on materials used to create metadata (XX library collection / personal items / web URL /
	Comment	Optional	Notes other than bibliographic items. Items for sharing information among authors, such as font
	Create Date	**Required**	Date of data creation
	Renewal Date	**Required**	If you create a new file, make sure that the update date is the creation date.
	Creator	**Required**	Name of the creator (private on the DB, for administration)

Fig. 7 Meta data of Bibliographic

and the statements of responsibility. We have been developing through repeated trial and error, such as establishing a new “ethnicity” item to keep notes on ethnic minorities and creating a “source” item to incorporate information from the Internet.

We found that some publications were disseminated only through Facebook and other social networking services. After discussions among librarians, researchers, and information technologists involved in the project, we agreed that the four items listed here are necessary.

1. Screenshot (which proves it existed once at a time)
2. Date and time of acquisition (when the data was obtained)
3. Date and time of the information source (if known)
4. URL of the information source

   * If it is a social networking service, the URL of each article is available from the date/time link, which is also the URL of the meta information “information source.”

■ **Case 2: Database of Rubbings of Inscription in Hanoi Old Quarter** (Fig. 8, 9): This database comprises the rubbings from stone monuments and bells in Vietnamese Buddhist temples. It contains 1,312 materials with 123 digitized images. Based on this dataset, we have been developing the system using Google Cloud Services. In creating the design, we used ChatGPT Plus, one of the generative AIs. The Google Maps API is difficult to use because it requires a fee when the free quota is used up. As a result, we have tried to find out how to implement this within the scope of free services; we used a method of asynchronously passing the map and marker information obtained with the StaticMap class in Google Apps Script. This has allowed us to paste Google Maps and their locations onto our site without incurring any fees. Another great feature is that if you later add or delete images on Google Drive, they will be reflected immediately in the database.

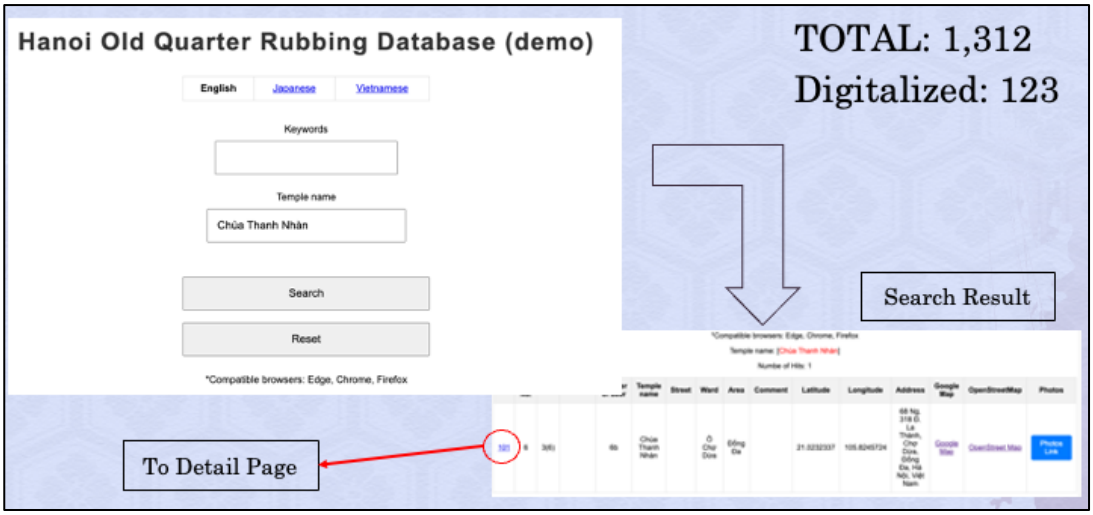

Fig. 7 User Interface of Hanoi Old Quarter Rubbing Database (demo)

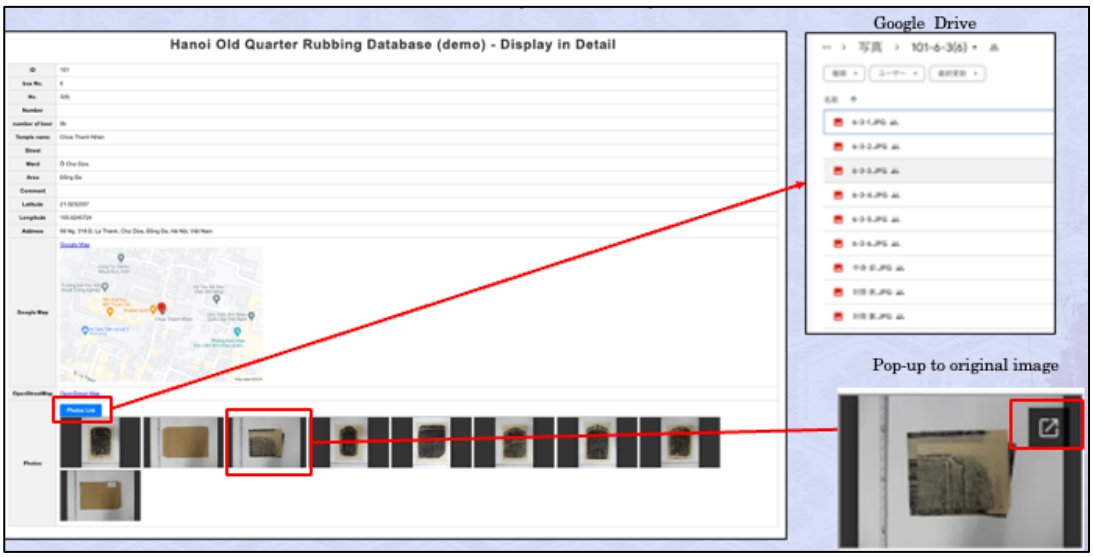

Fig 7 Detail Page of Search Result

■ **Case 3: Vietnamese Chinese Character (Chu Nom) Buddhist Scriptures Database (**Fig. 10, 11**):** This is a cloned version of the database in Case 2, with some features disabled, such as Google Maps and multilingual UI. This database consists of meta-information and materials of the rubbings printed from stone monuments and temple bells in Vietnamese Buddhist temples. The database contains about 100 items, 3,125 images and these basic data are stored in the Kyoto University Area Materials Digital Archive. The reason for necessity of a database, even though it is already available as a database, is that there are requests to also store data linking related research materials. We confirmed that the processing speed becomes very slow when assigning a Google Drive folder for each meta-information and checking for the presence or absence of data. Therefore, the checking functions was turn off.

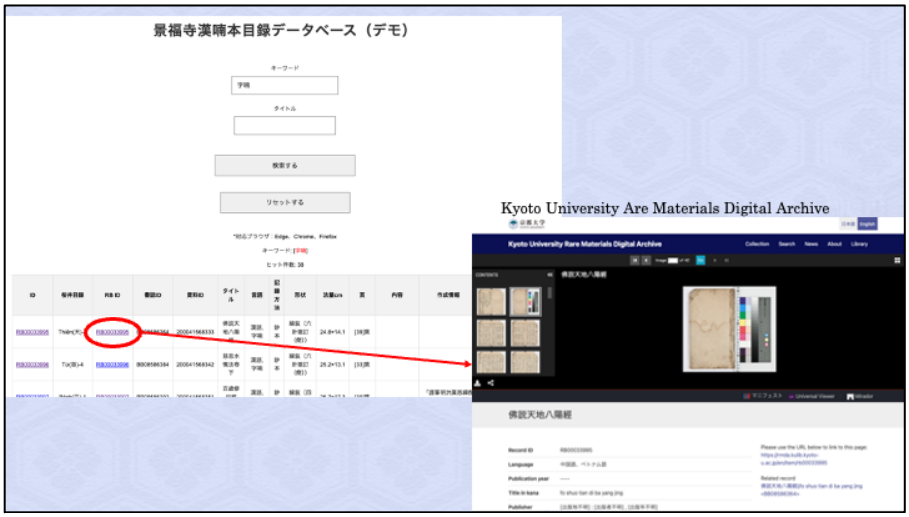

Fig. 8 User Interface of the Chu Nom Database

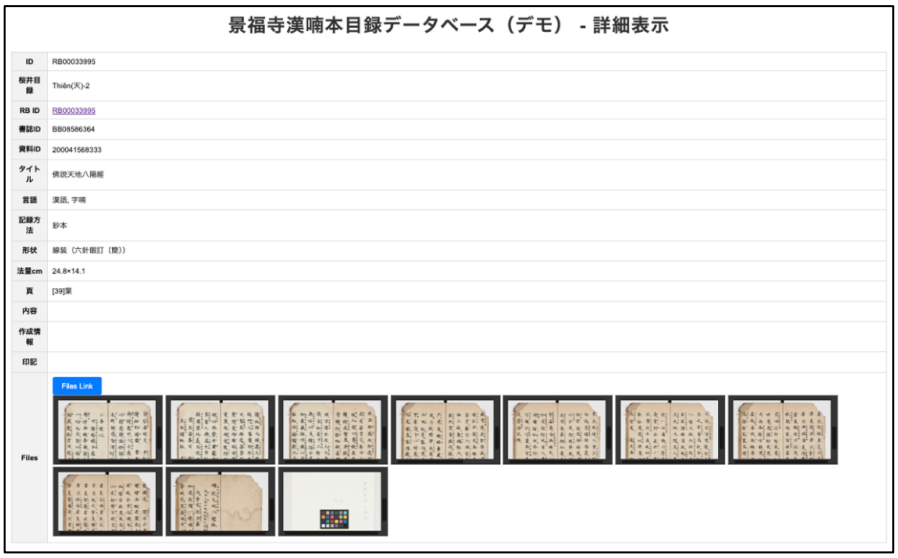

Fig. 11 Detail Page of Search Result

The implementation was challenged using TIFF, about 40 MB per image file and the total volume amounts to 142 GB, as the high-resolution image format. As a result that it has not slowed down the processing speed regarding loading images from Google Drive.

- **Case 4: Thai Database of Three Seals Law (Thammasat University Edition)** (Fig. 12, 13): The database has 3,719 items and 6,666 (1GB) of photo data. This one also uses the system from Case 2, but as close as possible to the UI of the existing database without using a complicated system. We can achieve to implement basic searching using our approach. Of course, we must remember that the "My database", which is an integrated search system, has a variety of search options. However, even in case the integrated database can no longer continue to be maintained, datasets can be migrated to this framework when the size is small enough.

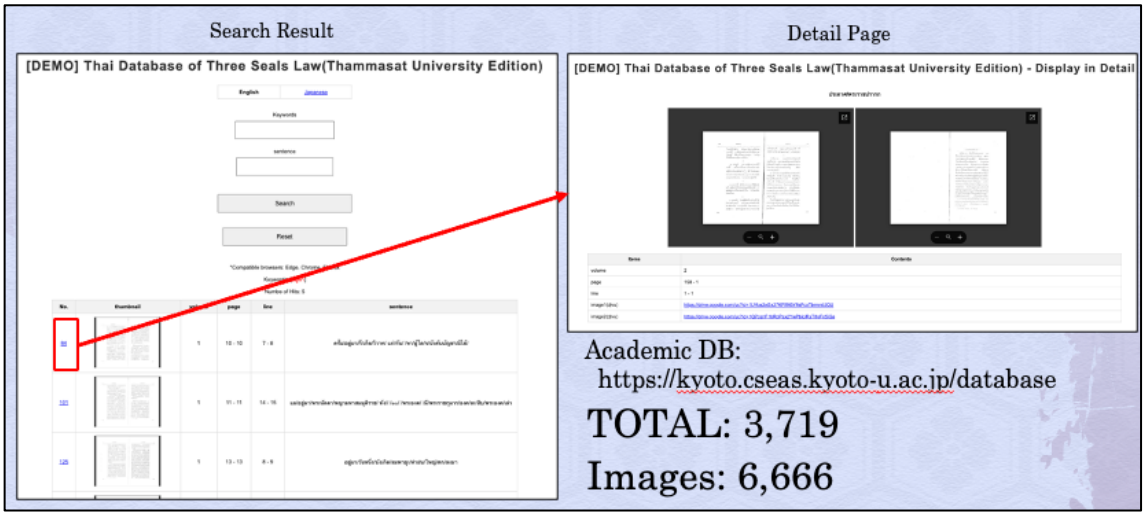

Fig. 12 User Interface of Thai Database of Three Seals Law (Thammasat University Edition)

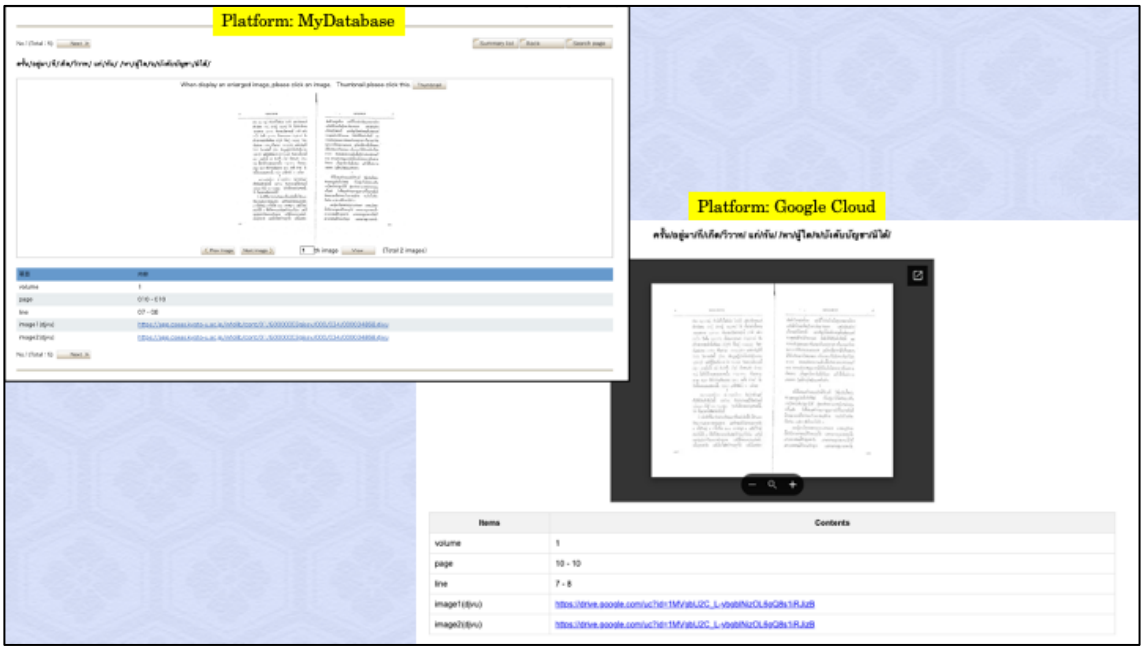

Fig. 9 Comparison of Detail Pages

6. Publishing trend analysis of multilingual periodicals.

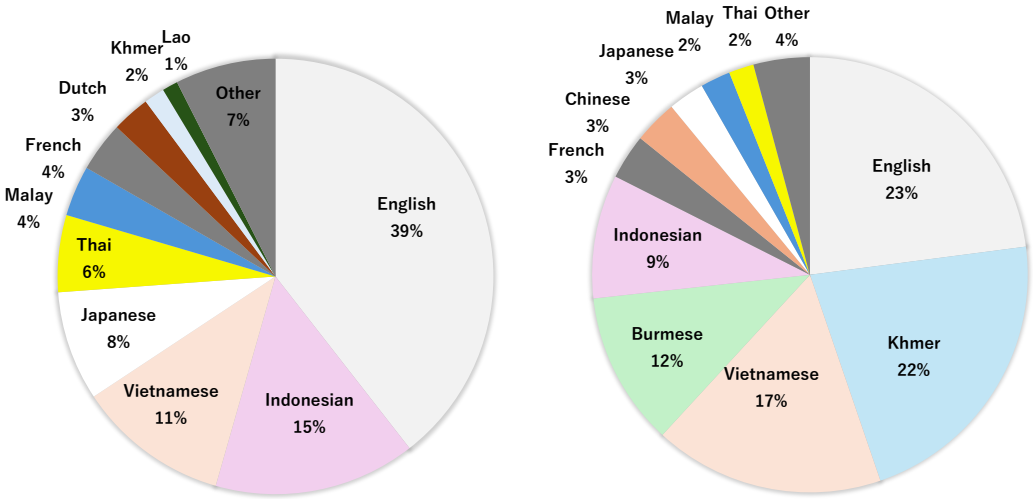

Fig. 104 Language used in periodicals; academic journals and magazines (left), and newspapers (right)

Although the primary purpose of the database is to store information that users will want to find, the data can also be used for analysis. Another objective of this research is to explore its potential uses for academic study. This presentation will show examples of simple sociolinguistic analyses using the database.

The pie chart on the left shows the top 10 languages used by journals published in Southeast Asian countries and for Southeast Asian people. The ranking is based on the language codes registered in the database's metadata. Here, "journals" stands for both academic journals and general magazines. All numbers refer to the number of publication titles, and those using multiple languages are counted multiple times.

As it shows, English is the most widely used language but does not account for the majority. Indonesian is second, followed by Vietnamese, Thai, and Malay, which are the official languages of each country. These languages are ranked higher than the languages of the former colonial powers, such as French and Dutch, and this shows that the variations of printed media communication in the people's languages in each country are relatively strong.

The pie chart on the right summarizes the top 10 languages used in "newspapers". Although English is still the primary language in each country, it accounts for only 23% of the total, a lower percentage than that of journals. Khmer is only 2% of the total number of journals, but here in newspapers, it is prominent at 22%. According to the database analysis, this is because many Khmer-language newspapers were published in the 1990s, many of which were discontinued within a few years.

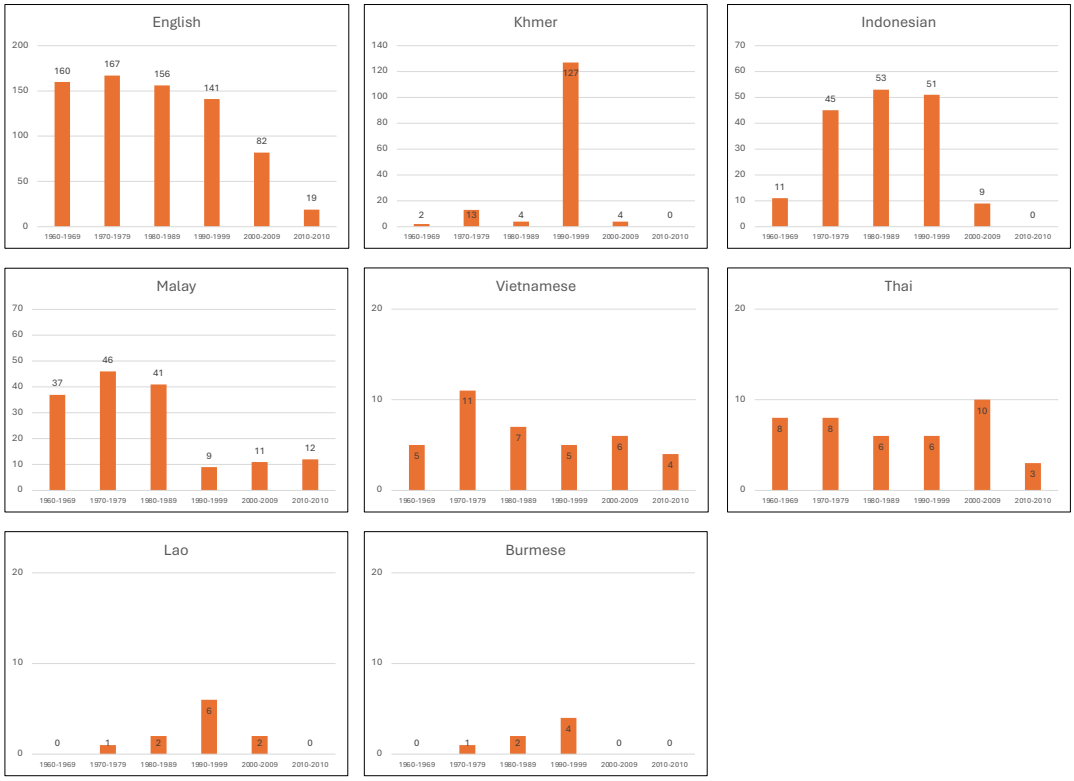

Fig. 15 The year (decade) the first issue of a periodical was published.

This bar graph shows the number of first issues of serials published per decade, based on the metadata "Year of Publication." There are many English, Indonesian, and Malay publications, but we can see that they have stagnated since the beginning of the 21st century. On the other hand, Vietnamese and Thai publications are fewer overall, but new publications are constantly being produced.

The place of publication can also provide clues about ethnic dynamics. The Shan is an ethnic minority residing primarily in Shan State, Burma. Ethnologue.com estimates there are 4.59 million native speakers of this Tai-Kadai language, making them the second-largest ethnic

Title	Medium	Language	Place of Publication
Hsenpai Media	WEB (Youtube)	Shan	BR
ယွတ်ၼုင်း yaut*khūñˈ″ (Market)	WEB (archive)	Shan	BR
ပီးမ့်,တီး ၇၁၁၄ဢီး pīˈ″ mvˈ,taiˈ″ 2114 nî (Shan New Year 2114)	WEB (archive)	Shan / English / Thai	BR
ပပ်.ယိလုၵ်ငဵတီး papˈ′ yiñˈ ñeˈ taiˈ″ (Shan Culture Magazine)	WEB (archive)	Shan	BR
ပိုၼ်ၼ်.ဂူ,ပတေ,သၢ puinˈ′ hūˈ′ pateşā,(General Knowledge)	WEB (archive)	Shan	BR
လိူင်းလ်ဝ်း cūiñˈ″ lèvˈ″ (Freedom Way)	WEB (archive)	Shan / English	BR
လုမ်းဓၢဝ်,ၽူးတွံ,ဂူဂ်း Cumˈ″ khāvˈ,phǔ toi,hauk* (Shan Herald Agency for News / S.H.A.N.)	WEB	Shan	BR
The Journal of Tai Studies	PRINTED	English	BR
ဥၼ်းဓေၢ Kaunˈ″ kho (Independence)	PRINTED	Shan	BR
ၸိုၼ်ၼ်လိဂ်တီး မုဂ်ထုံးယုဂ်ယွင်လိဂ်တီး တဂ်ဂသုဝ် ဝိင်းရၼ်ဂုင် Phuinˈ likˈ Taiˈ″ mukˈ cuɱˈ″ yukˈ yauñˈ likˈ Taǐ takˈ ka suvˈ viñˈ′ ″ Ranˈ kuñˈ (Shan Literary Society University of Rangoon)	PRINTED	Shan / Burmese / English	BR
ရန်ကုန်တက္ကသိုလ် တိုင်းလူငယ် စာစော Ranˈ kunˈ takkasuilˈ Rhamˈ′ ″ cā pe mrhañˈ′ tañˈ reˈ″ ˈa sañˈ′ ″ (The Rangoon university Tai youth magazine / Shan literary society university of Rangoon)	PRINTED	English / Burmese	BR
ၸုၼ်းတွံ,ဂူဂ်း Phǔ toi,hauk*(Reporter)	PRINTED	Shan / Burmese / English	BR
ယွတ်ၼုင်း Yaut*khūñˈ′ ″	PRINTED	Burmese	BR
လွင် ယိလုၵ် ၼၼ် တီး Lauñˈ yiñˈ ñeˈ Taǐ (Speak in Shan)	PRINTED	Shan / Burmese	BR
သိၼ် တီး Señˈ Taiˈ″ (Voice of Shan)	PRINTED	Shan	BR
ၸိုၼ်ၼ်လိဂ်း သၢ်ဇ်ုတီး Phuinˈ lik*Sāicvˈ Taiˈ″ (Shan Life Book)	N.A.	Shan / English / Burmese	BR
ရှမ်းပြည်နယ် မဂ္ဂဇင်း Rhamˈ′ ″praññˈ nayˈ Maggajañˈ′ ″ (Shan State Magazine)	N.A.	Burmese / English	BR

Fig. 16 The periodicals for the Shan People

minority in Burma after the Burmese and, therefore, the most significant ethnic minority in Burma. Outside Burma, there are also native speakers in Yunnan, China, and northwestern Thailand. There are 16 periodical publications registered in the database for the Shan people (Fig. 16). All of them are published in Burma (BR), which shows that the Shan language occupies a relatively strong position as an ethnic minority in Burma.

The contrasting case is Hmong. The Hmong people are a minority group living in southern China (Miao Autonomous Prefecture), the mountainous regions of Vietnam, Thailand, and northern Laos in Southeast Asia. Their total population is estimated at 4-5 million.

Title	Medium	Language	Place of Publication
18XEEM : Cultural Hmong Magazine	PRINTED	English, Hmong	Wisconsin, USA
Hmong Times	PRINTED	English	California, USA
Báo ảnh Dân tộc và Miền núi	WEB	Vietnamese, Hmong, and other minority languages	Hanoi, VNM
Hmong Daily News	WEB	English	California, USA
Hmong Today	WEB	English	Minnesota, USA
Hmong Star	WEB	English, Hmong	Minnesota, USA
Hmong Newsletter	WEB	English	Brussel, BEL (Unrepresented Nations and Peoples Organization /UNPO)

Fig. 17 The periodicals for the Shan People

Seven media outlets aimed at the Hmong people are registered in the database, of which five are published in the United States, one in Belgium, and one in Vietnam (Fig. 17). The Hmong

people also live in China, Laos, and Thailand, but these are not published or disseminated in those countries (or at least not included in the database).

During the Vietnam War (1955–1975), Hmong living in Laos, mainly in Long Tieng in the central region, cooperated with the CIA. After the US withdrew from Vietnam in 1975, many Hmong emigrated as refugees to the US, France, Australia, etc., via Thailand. The number of Hmong who emigrated to the US has now grown to a community of 360,000, and they have established the most significant Asian community, mainly in Minnesota and California. It is for these reasons that many periodicals aimed at the Hmong people are published in the US.

7. History and Future of the Projects

We have introduced four databases using small-scale datasets. We believe that the database approach utilizing cloud services has great potential. Therefore, we would like to establish this as one of the database frameworks while applying it to other datasets in the future, exploring the extent to which cloud services, including APIs and external collaboration, can be used more easily. We also hope that our challenge will lead to new projects, such as the analysis of multilingual publication trends and the challenge of formulating evaluation indicators, as described here.

This April, we entered the next stage of our project. This stage concerns evaluating and selecting the "core journal" in the Southeast Asian Periodicals Database (Fig. 18). In phase 4, we will reconsider the definition of the core journals from Southeast Asian periodicals and re-select them according to our definition. We will collect information on holdings of core journals in libraries that would cooperate with our project and try to create a system of DDS among them.

Currently, we are challenging to see if we can build an evaluation system that replaces the impact factor, focusing on the "core journals" of Southeast Asian periodical publications. For the

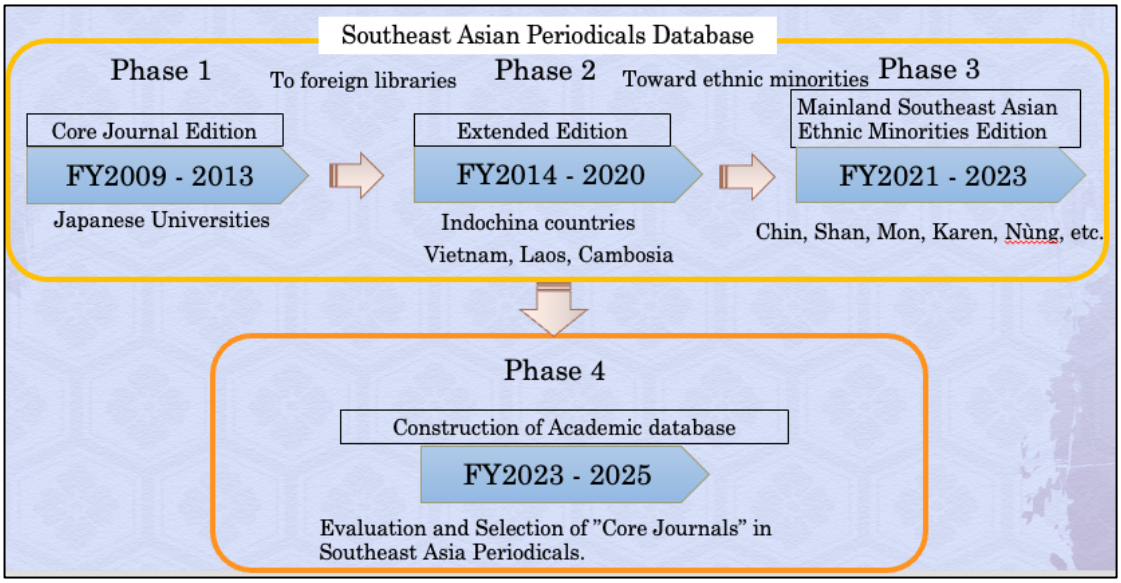

Fig. 18 History and Future of the Project

field of area studies, we can improve the accessibility of research information resources in Japan and Southeast Asian countries by valuing such publications because articles in 'local' journals published in the region (language) concerned may be of far greater academic value than articles in international journals with high impact factors.

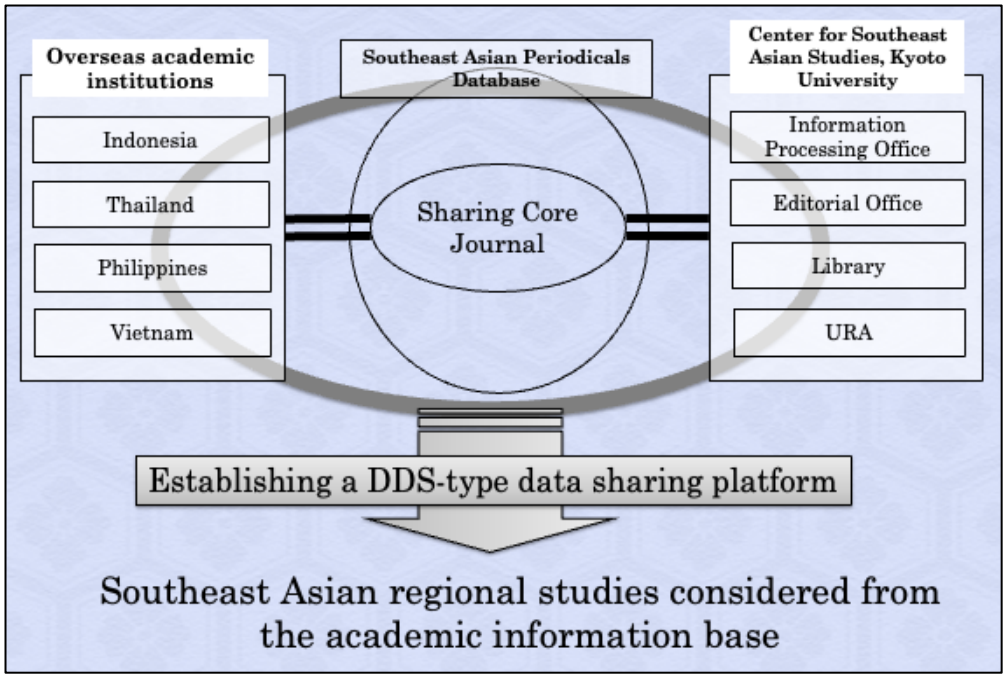

Fig. 19 Image of further collaborations

References

[1] Shoichiro Hara. 2016. "Open Platform for Academic Humanities Data." *International Symposium on Grids & Clouds 2016 (ISGC 2016),* March 2016, Academia Sinica.

[2] GeoNLP Project, 2022, "Historical Place Name Data Dictionary"(Rekishi chimei deta jisho), March 26, 2022, (Retrieved July 3, 2024, https://geonlp.ex.nii.ac.jp/dictionary/nihu-placename/). Japanese.