**GOTO — Digitalization**

<3 images>

**Digitalization of Shoso-in Monjo and Extraction of Knowledge**

Goto, M. and Naito, M.

We have been conducting research on digitalization methods for the restoration of Shoso-in Monjo since 2002. Shoso-in Monjo is a collection of historical documents from 8th-century Japan, comprising approximately 15,000 documents in total.1 It is the largest group of historical documents from ancient Japan or East Asia, and is highly valuable data. However, because of the historical background of Shoso-in Monjo, the documents have become disorganized compared to their original, 8th-century state. A “restoration” to this original state is necessary in order to conduct research using Shoso-in Monjo. To resolve this issue using computers, we reconstructed a virtual Shoso-in Monjo and created a comprehensive, all-encompassing database of the scattered Shoso-in Monjo information. We called this database the Shoso-in Monjo Database: ‘SOMODA’.2

We conducted a markup of the Shoso-in Monjo textual data using XML (eXtensible Markup Language), thus structuring the database, and linked image data and index information. Thereby, we created a system through which any information from the Shoso-in Monjo can be acquired.



Figure 1. The main page of SOMODA.

SOMODA does not just display complete texts or images; it is a system containing various information derived from data that was restored, and a system through which this information can be viewed simultaneously. We developed the latter database in 2007.

After considering how SOMODA ought to be developed from now on—that is, approximately seven years after the database was first created—we concluded that it is necessary to simplify the data search and information discovery model of Shoso-in Monjo along with enabling cooperation with many databases, thereby ensuring effective information discovery based on knowledge data. We therefore adopted an ISO standard Topic Maps as an ontological method and applied it to the SOMODA data. In the next section, we report our findings.

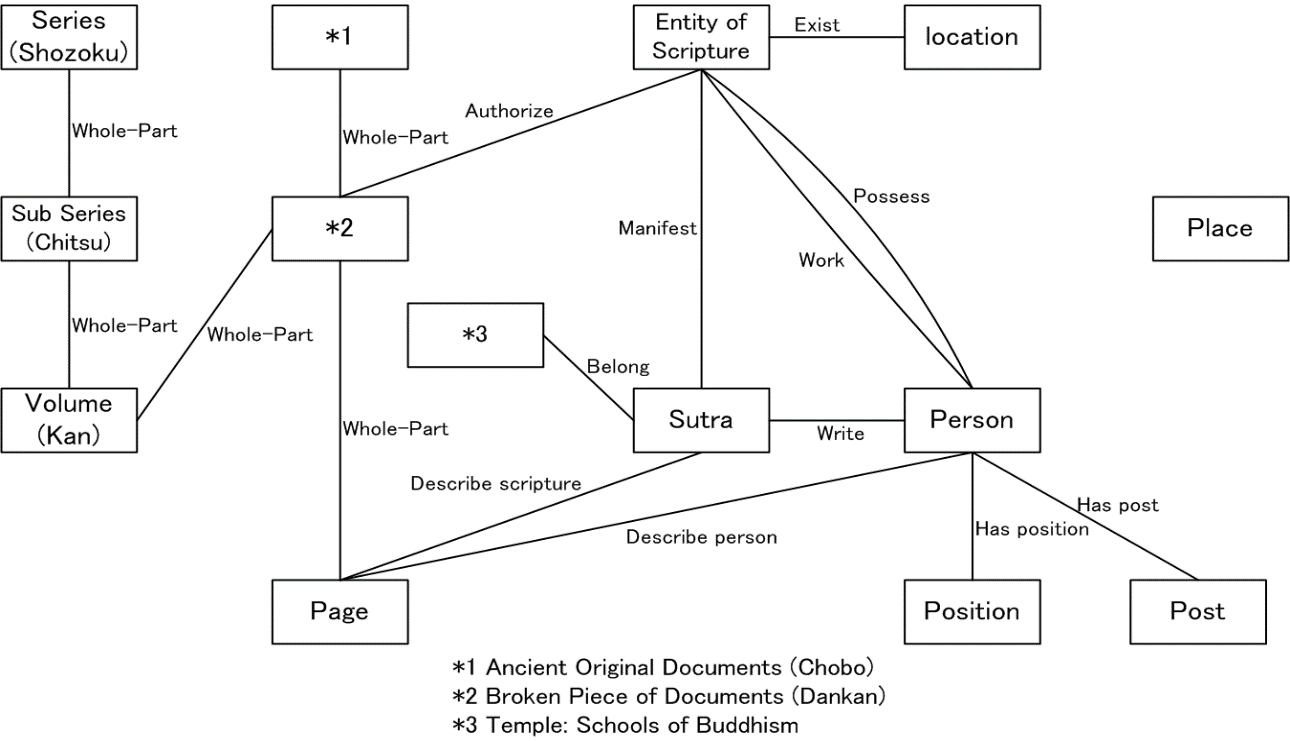
**Characteristics of the Topic Maps**

Topic maps are defined as follows by ISO 13250 Topic Maps Data Model. ‘Topic maps is a technology for encoding knowledge and connecting this encoded knowledge to relevant information resources. Topic maps are organized around topics, which represent subjects of discourse; associations, representing relationships between the subjects; and occurrences, which connect the subjects to pertinent information resources’.3

Furthermore, it is possible to assign Subject Identifiers to topics to identify/differentiate subjects within a topic map. Internationalized Resource Identifiers (IRI) are used as Subject Identifiers. Subject Identifiers that have been made public are called Published Subject Identifiers (PSI). By using PSIs, it is possible to accurately identify/differentiate subjects based on unique IRIs, without being troubled by issues such as synonyms, homonyms, and polysemy, which are difficult to resolve on the Web today.

Based on this, it will be possible to assign academic IDs one by one to the Shoso-in Monjo on the Web while enabling a high level of cooperation with other digital archives (see Figure 2).4

Figure2. The ontology of Shoso-in Monjo topic map



**The Generated Shoso-in Monjo Topic Map**

Based on these characteristics, we created a topic map for the Shoso-in Monjo as described below.

First, based on the historical particulars of, and changes to, Shoso-in Monjo, we noted the relationship between the state of the Shoso-in Monjo today and the structure of the 8th-century original. At present, Shoso-in Monjo has a layered structure of association: Series (*Shozoku*), Sub Series (*Chitsu*), Volume (*Kan*), *dankan* (the smallest unit of paper remaining from the original 8th-century documents), page. We created a representation of this layered structure. Next, we noted the original structure as being the *Chobo* (the original Documents)—*dankan* (the broken piece of Documents)—page. We then assigned PSIs to each page and *dankan*, making higher-level use of the Shoso-in Monjo digital data possible. This will become the foundational information for the Shoso-in Monjo topic map as a whole.

We next extracted information on personal names in ancient Japan from existing dictionaries. Additionally, we affixed associations regarding the social status, occupation, and location to the personal names based on existing research articles on Japanese history. Then we affixed associations with the ‘pages’ in the Shoso-in Monjo in which these personal names appear. This allows one to understand what kinds of people are described in the Shoso-in Monjo and in what kinds of historical records they appear.

Third, we extracted information relating to East Asian scriptures and sutras from existing dictionaries. We then noted the version, anthology, and author of the sutra. Further, we affixed associations between the titles of these sutra and the places in which they actually appear within historical records. Through this, the relationship between book titles and publications as entities can be understood.

**The Usefulness of the Shoso-in Monjo Topic Map**

We anticipate that the Shoso-in Monjo topic map will be useful in the following ways.

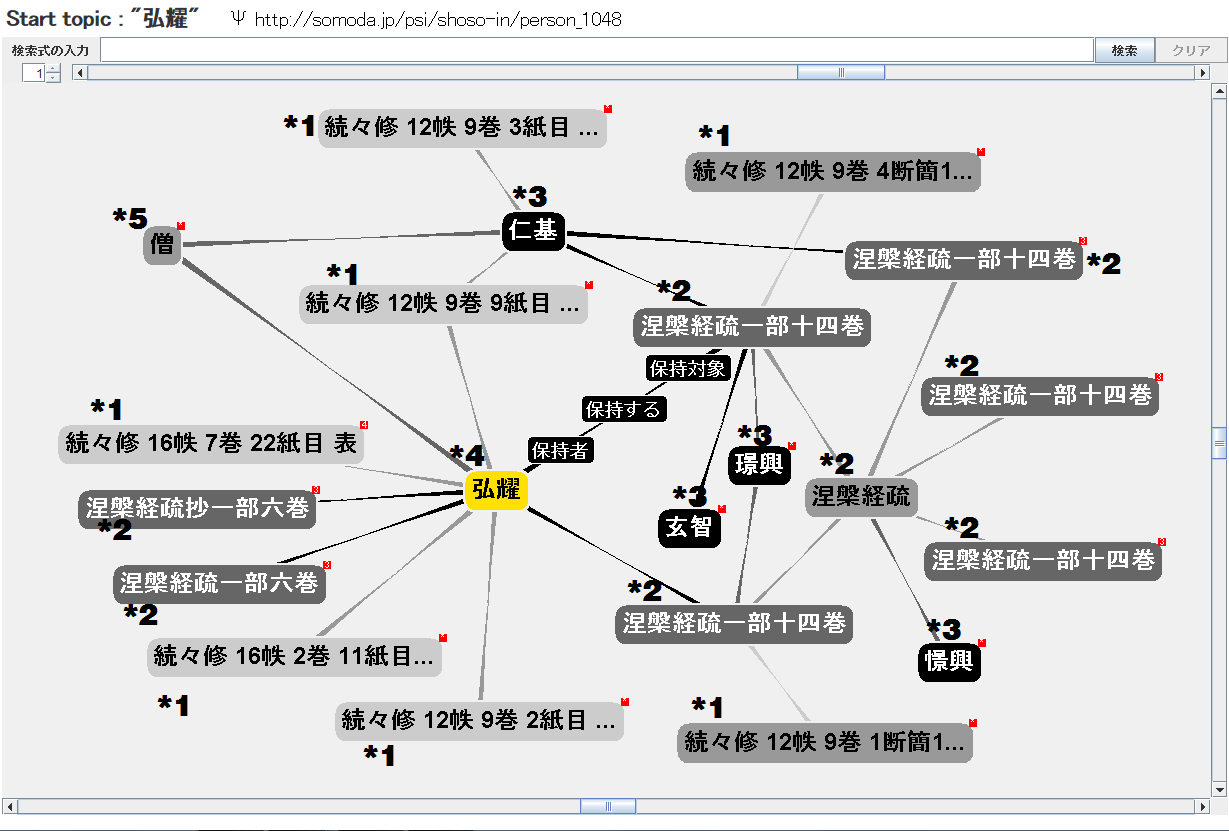
• Connecting personal names in the Shoso-in Monjo with sutra titles.

• Discovery of semantic information that may not be found using a conventional search model, such as whether a given person had anything to do with a unique scripture at a sutra-copying place.

For instance, the relationships between the holders of annotated editions of sutras at a given time during the 8th century and their locations have been visualized in Figure 3, using the topic map we generated as the source. Based on this, information such as what kinds of texts a set of annotated editions belonging to a given monk are and whether there is any bias regarding the authors of these annotated editions can be understood. Further, by adding information on the location and time, information can be obtained on whether the holder of an annotated edition or the location of the edition changes with time. This will become an important source of information for analysing how the monks of the 8th century studied sutras.

Figure 3　Affixing associations between the whereabouts of sutras and the holders of records using the topic map

\*1: PageID \*2: Sutras Name \*3: Name of Monk \*4: Current Node \*5: Position



*Connecting registers and sutras/personal names*. The Shoso-in Monjo had a history of causing problems for research due to its exceedingly complicated structure, which made it difficult to see the big picture. The Shoso-in Monjo database created up to 2007 is what resolved this issue. However, it was not suited for viewing multiple materials relating to a specific keyword at once. That has been resolved and it has become possible to mechanically analyze whether, for instance, a given sutra appears in a specific register (document).

*Connecting personal names with each other.* Many personal names can be seen in the Shoso-in Monjo. Accordingly, it will be possible to analyze whether particular connections exist between personal names or between occupations. Based on this, it is possible that the personal relationships that have hitherto been only vaguely identified within the Todaiji sutra-copying center—the focus of the writings in the Shoso-in Monjo—will become clear.

In the future, we can expect that large amounts of research information/knowledge beyond what has already been described will be organically connected to the historical records of the Shoso-in Monjo within the Shoso-in Monjo topic maps.

**Challenges and Prospects**

The development of the database created in this study has only just begun, and there are the following challenges.

1. Processing complex time transitions. While the Shoso-in Monjo is a collection of documents from a relatively short period—that is, the mid-8th century—there are still some temporal changes; within those temporal changes, there are also changes to the ontological structure. How these changes are processed will be investigated from now on. Further, while the contents of the Shoso-in Monjo texts are from the 8th century, it is necessary to take up to the Meiji Period into consideration in order to include subsequent changes in complex historical record structures. How to respond to this kind of circumstance will be a challenge hereafter.

2. Establishing appropriate topics. While the first step is inputting information on personal names and scriptural titles based on historical records, inputting more abstract research knowledge/information is desirable if one is to achieve the original goals of the topic map. It will be necessary to constantly consider the extent to which establishing these topics conforms to the goals of Shoso-in Monjo research.

At present, we have obtained permission to make secondary use of the index appended to the research papers of the Shoso-in Monjo, and our next plan is to put it to practical use.

In the future, we believe it will be necessary to extend the research of the Shoso-in Monjo outward by further brushing up on ontology and creating links to historical records from the same period as the Shoso-in Monjo, such as the Shoku Nihongi, and to databases on Buddhist sutras.

Using this system, it will be possible to develop materials for an even higher level of research in history and humanities in Japan.

**Notes**

1. A detailed explanation about Shoso-in Monjo in English appears at Bryan Lowe and Chris Mayo, Guide to Shoso-in Research, https://my.vanderbilt.edu/shosoin/.

2. See Shoso-in Monjo Database SOMODA, http://somoda.media.osaka-cu.ac.jp/.

3. ISO 13250-2: Topic Maps—Data Model, http://www.isotopicmaps.org/sam/.

4. Shoso-in Monjo Topic Map, http://www.somoda.jp (scheduled to open in 2015).