Unit-based Scheme Connection Between TEI and Original Scheme To Promote Data Sharing Beyond Cultual Diversities

Kazushi Ohya
Tsurumi University
2014-04-28
extended abstract

1 Introduction

This paper is a report of a comparative study between the TEI scheme and the dominant bibliographic data format for old Japanese books. This study is a part of the preparation to make a support system to be used in a field investigation of old books. In order to apply the TEI scheme to a Japanese bibliographic(JB) format or other cultural categorization that was not presumed at the beginning of TEI, the TEI scheme is to be re-estimated by comparison on at least two levels: (a) data unit level and (b) data structure level. On a unit level, we confirmed that there are six types of correspondences between TEI and JB items: (a1) one to many, (a2) many to one, (a3) one to nothing, (a4) nothing to one, (a5) clear one to one, and (a6) ambiguous one to one. On a structural level, we confirmed that there are two types of changes in the TEI scheme to be applied to the JB format: (b1) adding a new unit as a sub-unit of an existing TEI element, and (b2) adding a new unit as an upper-unit of an existing TEI element.

In the former case (b1), instead of making original elements, we can use the element fs in order to expand a data structure downwards. However, in the latter case (b2), changing an internal structure by adding non-terminal nodes into an existing data structure influences the data handling in many ways. Therefore, we have to decide whether to stop using TEI and make original data scheme, to keep using TEI in disassociation with a traditional data format having been used in our culture, or to keep using heavily customized TEI without warranty of data conversion to the other data in the normal TEI scheme. Up to the present time, making original data formats has been taken for granted in a study of Japanese bibliography on old books, and this situation possibly might not be easily changed. In this paper, we propose using the TEI scheme not as a rigid data format ensuring data conversion but as a referential scheme to cover cultural diversities with loose connectivity. This stance on the TEI scheme stems from a strategy of unit-based data sharing, which is different from the hitherto adopted structure-based data sharing strategy. As a benefit of this approach, we can make it relatively easy to define a data format, such as with the data of which is usually hard to see that a document structure exists, e.g. metadata of non-book materials, data of graphical novels, text data from maps, and so on.

2 Background

Books known to be published in Japan are recorded in the Catalogue of National Books, called "Kokusyo Soh-Mokuroku" in Japanese, which stores about 500,000 titles published from the first record in the history of Japan to before the Meiji era(circa 1900). The data has been updated and now we can retrieve the whole database through a web service from the National Institute of Japanese Literature(NIJL), which is known as the General Catalogue of Japanese Books called "Kokusho Sohgoh Mokuroku."[1] This database is a result of the full-out investigation by bibliographers, who use the same data sheet for the investigation(Fig.1 A data sheet of NIJL).

This investigation is often done as fieldwork. In fieldwork we have to do a lot of post work and the same in this investigation. So, we have a plan to make a support system to be used in this field work, which consists of a tablet as an input device and a server to assist in entering the data and storing it for post work. In preparation of making this system, we analyzed a data format of the data sheet and made an XML scheme for them.

---

[1] http://base1.nijl.ac.jp/~tkoten/about-en.html

3 Comparison between JB Format and TEI Scheme

The data sheet used in NIJL has 34 data units, and each of them corresponds to an element of TEI as Fig.2(A table of comparison). In preparation for this comparison, the JB format in a data sheet was analyzed and the result was defined by using documentation elements in TEI(Fig.3 A part of JB format definition). From this list, we found that there are six types of correspondence on a data unit level and two types of it on a structural level.

The six types of correspondence on a data unit level are as follows. (a1)1:n relation; an element in TEI has multiple correspondent elements in the JB data format: e.g. binding, condition, etc. (a2) n:1 relation; an element in the JB data format has multiple elements in TEI: e.g. "youji"(kinds of usage of characters), "jo"(kinds of introduction), etc. (a3) 1: relation; no element in the JB data that corresponds to an element in TEI: e.g. heraldy. (a4)   :1 relation; no element in TEI that corresponds to an element in the JB data format: e.g. "koukoku"(kinds of advertising), "kiwamehuda"(kinds of certification), etc. (a5) 1:1 clear relation; there is a clear one to one relationship between the TEI and the JB data format: e.g. repository, idno, etc. (a6) 1:1 unclear relation; there is an ambiguous relationship between the TEI and the JB data format: title and "gedai"(kinds of titles on a cover), etc.

In the case of (a3) and (a5), there is nothing for us to do. In the case of (a4) and (a6), there is nothing for us to do but may be something for TEI to do, e.g. making new elements or revising definitions. In the case of (a1), we can keep using an original TEI element that corresponds to multiple original elements. However, it seems to be better to prepare sub-units suitable for JB data which would realize a simple and clear scheme for Japanese reseachers. In the case of (a2), we have to reconsider not only data units but also data structure or scheme. For example, if two corresponding elements of TEI exist in separate sub-trees that do not have the same parent, we cannot set a new abstract element that directly governs the two elements without link functions. To some extent, in order to handle this case, we have to revise a way to use the TEI scheme.

The former cases of (a1) and (a2) engender general types of structure changing: (b1) adding a new unit as a sub-unit, and (b2) adding a new unit as an upper-unit. The new units possibly have sub-structures. In the case of (b1) resulting from (a1), we can use the element fs as a sub-unit if the parent element permits, and keep using the TEI scheme. In the case of (b2) resulting from (a2), we have to customize the TEI scheme. However, when the new added upper-unit has an original structure, the customization is difficult to implement. In fact, making an original data scheme instead of using the TEI has been a natural selection for Japanese researchers. Of course, it is unrealistic to keep using the TEI scheme in disassociation with a traditional data format having been used in our culture.

4 Ideas for cultural diversities

As a way to solve the problem in (b2), it is possible to make a new module for the JB format. However, we do not think it is a good idea to set a new alternative module for metadata instead of the TEI header. A scheme for metadata is a common data format for information retrieval, thus is inappropriate to be a target of module selection.

As a way to cooperate with the TEI scheme in this project, we plan to use the TEI scheme just as a referential scheme. This means that the TEI scheme does not work as a rigid data format to ensure data exchange. The TEI scheme and the JB format is loosely connected with a referential relationship. This strategy abandons structure-based data sharing, but it supports unit-based data sharing or scheme connection. This approach would make it relatively easy to define a data format, such as with the data of which is hard to clearly see that a document structure exists, e.g. metadata of non-book materials, data of graphical novels, and text data in maps.

5 Conclusion

Using the TEI scheme as a referential scheme is not to deny the availability of a scheme for data sharing. This is a way to cover cultural diversities that were not presumed at the beginning of TEI activities. To tell the truth, we still now hesitate to adopt this stance into our project. However, as a theoretical conclusion, we have not found a viable alternative. As a next step, we have a plan to make a support system for the field work of bibliographic studies of Japanese old books by using an original scheme based on this strategy.