Unit-based Scheme Connection Between TEI and Original Scheme To Promote Data Sharing Beyond Cultural 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 reestimated 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 nonterminal 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).

This investigation is often done as fieldwork. In fieldwork we have to do a lot of post work

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

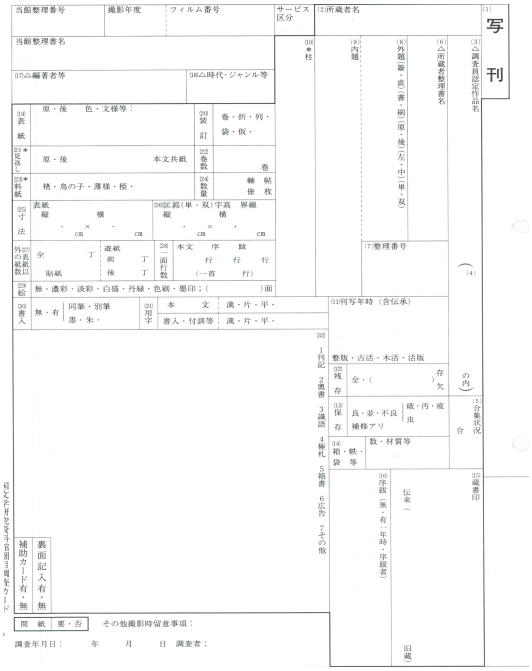

Figure 1 A Data Sheet of NIJL

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.

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.

JB Items	Desc.	TEI
jabib1	manuscripts or printed documents	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/@type
jabib2	a name of the reading material	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/repository
jabib3	a name defined by the investigator	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/msName
jabib4	additional information on a name defined by the investigator	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/msName
jabib5	a condition of biding	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib6	a name defined by the owner	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/altIdentifier
jabib7	ID defined by the owner	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/idno
jabib8	a title on the cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title
jabib8@base	a base paper for a title on the cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title/fs
jabib8@originalOrNot	original or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title/fs
jabib8@positionOfTitlePage	a position of the title on the cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title/fs
jabib8@typeOfBox	a type of a base paper for a title on the cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title/fs
jabib9	a title indicated in the content pages	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/titlePage
jabib10	a space in the center of a folio	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/foliation
jabib11	the year when the book or manuscript was made or appeared, which may be an estimated year	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/history
jabib11@typeOfPrinting	a type of printing	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/history/fs
jabib12	a condition of the book or manuscript; whether it contains a whole contents or not.	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib12@conditionOfContentParts	a condition of the part of content pages	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13	a condition of the preservation	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@totalCondition	a total condition of the preservation	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@broken	whether being broken or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@stained	whether being stained or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@shabby	whether being shabby or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@vermiculate	whether being vermiculate or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@otherDamage	other damages	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib13@mending	whether being mended or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/condition
jabib14	a description on an outer of the printed documents or manuscripts	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/p
jabib15	a description about a seal made by owners who may be the present or previous owners	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/addition/stamp
jabib15@provenance	provenance information	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/addition/stamp/note
jabib16	a description about contents in a front matter	there are multiple elements in TEI (1:n relation)
jabib16#JO	whether there is a description about this reading materials in a front matter or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/incipit
jabib16#BATU	whether there is a description about this reading materials in a back matter or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/colophon
jabib16@JO	a content of JO	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/incipit
jabib16@BATU	a content of BATU	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/colophon
jabib17	a description about editors or people who took part in this publication	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/bibl/editor
jabib18	a description about ages or a text type of body text to help identify this reading material	there are multiple elements in TEI (1:n relation)
jabib18#DATE	a content of DATE	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/bibl/date
jabib18#TEXTCLASS	a content of TEXTCLASS	/TEI/teiHeader/profileDesc/textClass
jabib19	a description about a cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/p/title
jabib19@typeOfCover	a type of a cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/adminInfo/custodialHist
jabib20	information on binding	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/bindingDesc
jabib20@	a type of binding	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/bindingDesc/condition/fs
jabib21	a flyleaf of a title page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/condition/p
jabib21@condition	a condition of the flyleaf	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/condition/fs
jabib21@typeOfPaper	whether the paper is the same as one used in content pages or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/bindingDesc/condition/fs
jabib22	the number of book forms of this title in terms of works, not physical forms	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/bibl/series
jabib23	a description about paper used	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/support/material
jabib23@typeOfPaper	a type of paper	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/support/material/fs
jabib24	the number of sub-parts to make this binding	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc
jabib24@typeOfUnit	a type of sub-parts or units	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/condition/fs
jabib25	information about size in a unit of cm.	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/extent
jabib25@width	width	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/extent/fs
jabib25@length	length	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/supportDesc/extent/fs
jabib26	information about lines on a page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib26@frameLines	a type of the frame lines	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib26@linesBetweenTexts	information about lines between text lines	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib26@width	width	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib26@length	length	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib27	the number of paper excluding a cover page	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib27@totalNumberOfPaper	the number of paper in this reading material	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib27@numberOfFrontFlyleaf	the number of a fly leaf in the front part of this reading material	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib27@numberOfBackFlyleaf	the number of a fly leaf in the back part of this reading material	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding/condition/fs
jabib28-1	the average number of lines in a page of a body part	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib28-2	the average number of lines in a page of a front matter	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib28-3	the average number of lines in a page of a back matter	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib28-4	the average number of lines in a verse	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/bindingDesc/binding
jabib29	information about pictures included	?
jabib29@existanceOfPicture	whether there are pictures or not	?
jabib29@typeOfPicture	a type of the pictures	?
jabib29@numberOfPicture	the number of the pictures	?
jabib30	information about additional writings besides or within the reading material	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc
jabib30@existanceOfNotes	whether there are additional writings or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc/summary/fs
jabib30@sameAsWriterOfBodyText	whether the writer of the additional writings is the same as one of the body text or not	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc/summary/fs
jabib30@color	a type of color	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/handDesc/summary/fs
jabib31-1	a type of scripts, letters, or characters in a body text	there is a n:1 relation between JABIL and TEI
jabib31-1#		/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/scriptDesc
jabib31-1#		/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/typeDesc
jabib31-2	a type of scripts, letters, or characters in other parts than a body text	there is a n:1 relation between JABIL and TEI
jabib31-2#		/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/scriptDesc
jabib31-2#		/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/typeDesc
jabib32-1	information about publication	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/imprimatur
jabib32-2	colophon	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/colophon
jabib32-3	information about this title in a back matter	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msContents/msItem/docTitle
jabib32-4	information about certification given by some authority such as appraisers	?
jabib32-5	information written on the surface of outer box that contains this reading material	?
jabib32-6	advertisement	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/physDesc/objectDesc/p
jabib32-7	other type of information	?
jabib33	supplemental information that cannot be described fully in data items	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/msIdentifier/altIdentifier/note
jabib34	any other information figured out by the investigator	/TEI/teiHeader/fileDesc/sourceDesc/msDesc/note
		/TEI/teiHeader/fileDesc/sourceDesc/msDesc/note

Figure 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). 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.

```
<specGrp>
  <elementSpec ident="jabbl1">
    <desc xml:lang="ja">写本・版本の別</desc>
    <desc xml:lang="en">manuscripts or printed documents</desc>
    <attList>
      <attDef ident="val" usage="req">
        <valList type="closed">
          <valItem ident="写本">
            <desc xml:lang="ja">写本</desc>
            <desc xml:lang="en">manuscripts</desc>
          </valItem>
          <valItem ident="刊本">
            <desc xml:lang="ja">刊本・版本</desc>
            <desc xml:lang="en">printed documents</desc>
          </valItem>
        </valList>
      </attDef>
    </attList>
  </elementSpec>
  <elementSpec ident="jabbl2">
    <desc xml:lang="ja">所蔵者名</desc>
    <desc xml:lang="en">owner name</desc>
    <content>
      <rng:text/>
    </content>
  </elementSpec>
  <elementSpec ident="jabbl3">
    <desc xml:lang="ja">調査員認定作品名</desc>
    <desc xml:lang="en">a title name defined by an investigator</desc>
    <content>
      <rng:text/>
      <rng:optional>
        <rng:ref name="jabbl4"/>
      </rng:optional>
    </content>
  </elementSpec>
  <elementSpec ident="jabbl4">
    <desc xml:lang="ja">調査員認定作品名の補足</desc>
```

Figure 3 A Par of JB Format Definition

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:φ relation; no element in the JB data that corresponds to an element in TEI: e.g. heraldry. (a4) φ: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 researchers. 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.