1. Aim of the study
Examine the possibility to convert classic documents with TEI-compliant XML.

An old wood-printed block book of Sharebon’s "Keisei-kai futasuji-michi" (1798) as a model case.

2. Significance of encoding Sharebon
1) Published in the broad age from the 18th to the first half of the 19th century.
2) Uses colloquial words and expressions among conversations between characters.
3) Abundantly describes things in Edo (former name of Tokyo) language and Kamigata (Kyoto-Osaka region) language.

Describing Sharebon in a machine-readable manner has profound significance for:
- bringing numerical analysis of unexplained colloquial expressions of the modern period into reality.
- facilitating humanities research in language history and descriptive bibliography.
- offering a common format for archiving coeval literature, since many texts that have a similar structure to Sharebon were published over the same period of time.

3. Text encoding
To analyze the manuscript from a corpus linguistic viewpoint, outward mark-ups and internal mark-ups of both document structure and linguistic structure are required.

3.1 Entire text
\[
\text{<back> <div><div><div>, Chapter 2<br>Chapter 1</div><div>back matter<br>narrative body<br>front matter</div></div><div>...</div><div>...<div></div></div></div>
\]

3.2 Below <p> (paragraph) level
- descriptive texts
  - colloquial expressions
  - warigaki interlinear notes
  - extended characters
  - ruby annotation
  - column break
  - line break
  - page break

Example: Conversation between two characters Suma (組野) and Goro (五郎)

3.3 Below <q> (sentence unit) level
- extended characters
- ruby annotation
- column break
- line break
- page break

Example: Conversation between two characters Suma (組野) and Goro (五郎)

4. Hierarchy within a group of elements

5. Conclusion
5.1 Issues in encoding Sharebon
The problems incurred in the process of encoding can be broadly classified into two matters:

(a) Text formatting on a CDATA (character data)
- Voice marking (dakuten, 促音)
- Phonetic correction (kana, 仮名)
- Iteration symbols (odoriji, 通り字)
  - Difficult to conduct morphological analysis properly.

(b) Structuring ruby annotations
- Text and base text are not in one-to-one correspondence
- Carries damaged text, typographical errors and omissions
- Ruby annotation on left-hand side simultaneously
  - Difficult to mark up both outward information and linguistic structure simultaneously.

5.2 Future challenge
- Devised a tag-set based on the TEI element, and structured Sharebon’s “Keisei-kai futasuji-no-michi” as a model case.
- Developing a corpus for historical documents in a comprehensive and versatile way was considered the ultimate goal.
- Our future challenge is to examine a compromise that satisfies both structures simultaneously while working out the problems with the TEI Council.