OUKA
The University of Osaka
Institutional Knowledge Archive

Title	人文学研究者必見! テキストデータとTEIで描く新たな研究ビジョン
Author(s)	吉賀, 夏子; 田畑, 智司; 甲斐, 尚人 他
Citation	
Version Type	VoR
URL	https://doi.org/10.18910/101978
rights	This content is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

OSAKA UNIVERSITY

グローバル日本学教育研究拠点「拠点形成プロジェクト」

ー人文科学分野向け研究データ管理を促進するデジタル・ヒューマニティーズ学習教材開発ー

人文学研究者必見!
テキストデータとTEIで描く新たな研究ビジョン

TEIとは何か？

Text Encoding Initiative

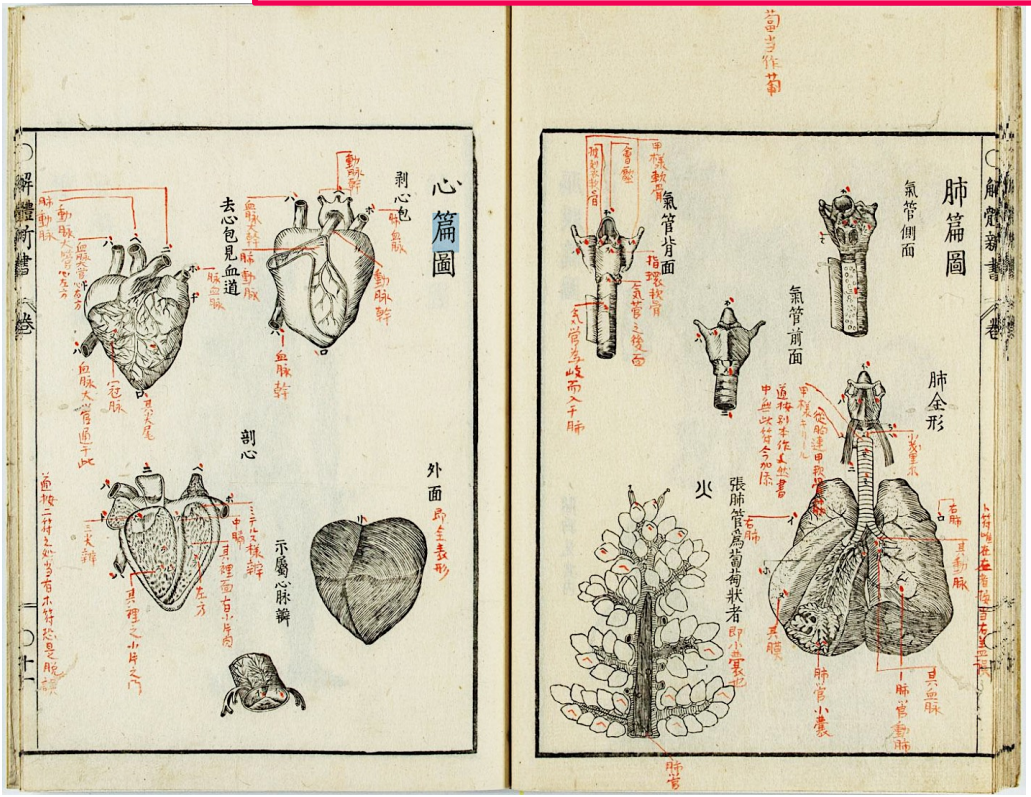
TEI(Text Encoding Initiative)

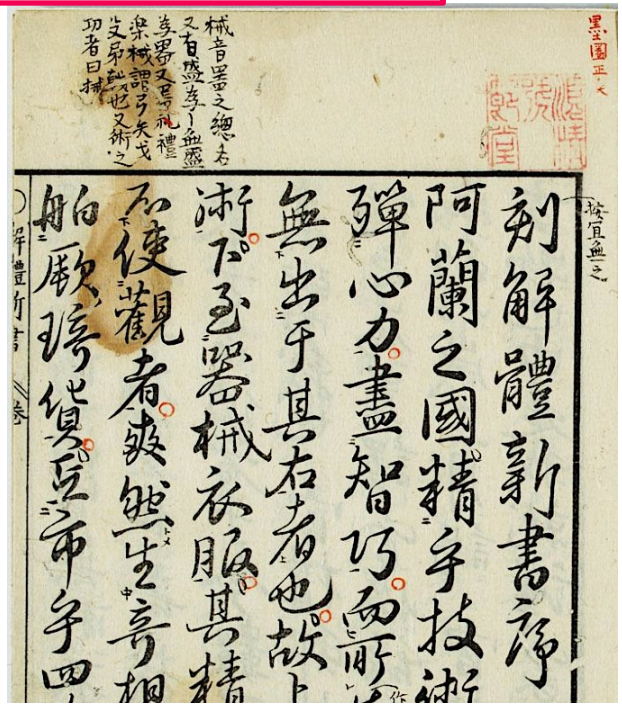
https://tei-c.org/

文学作品や歴史文書、書簡などのテキストに込められた意味や構造をデジタル形式で保存し、表現するための国際的な指針をまとめる国際団体

TEIの目的と利点(1)

本文以外の周辺情報(書誌情報、書き込み、レイアウト、印など)も知見





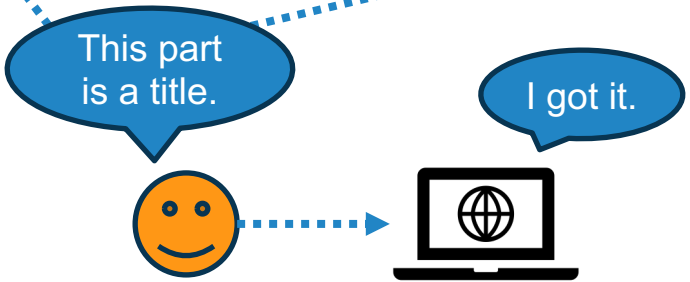
佐賀大学地域学歴史文化研究センター蔵「解体新書」

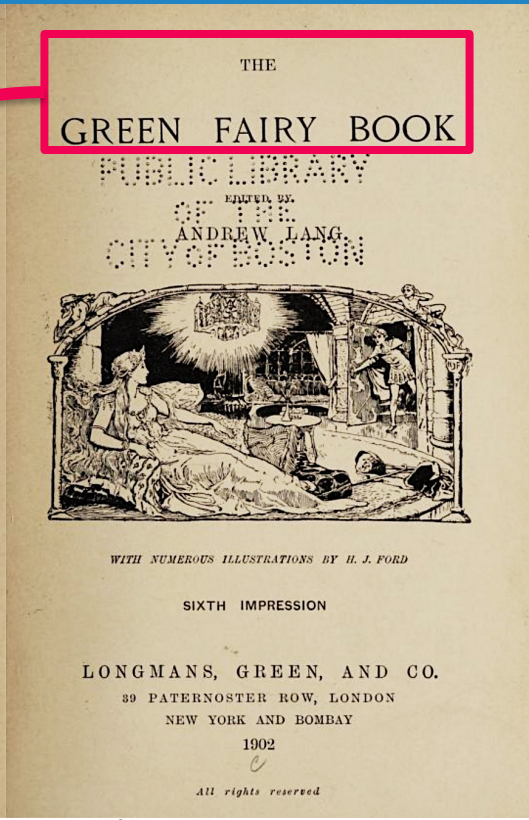
TEIの目的と利点(2)

XML(Extensible Markup Language)

- テキストの内容を整理し、構造化する記法のひとつ
  タグ(tag)

<title>The Green Fairy Book</title>



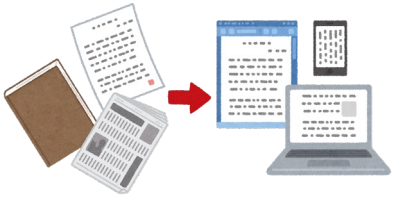


"The green fairy book" by Lang, Andrew, 1844-1912

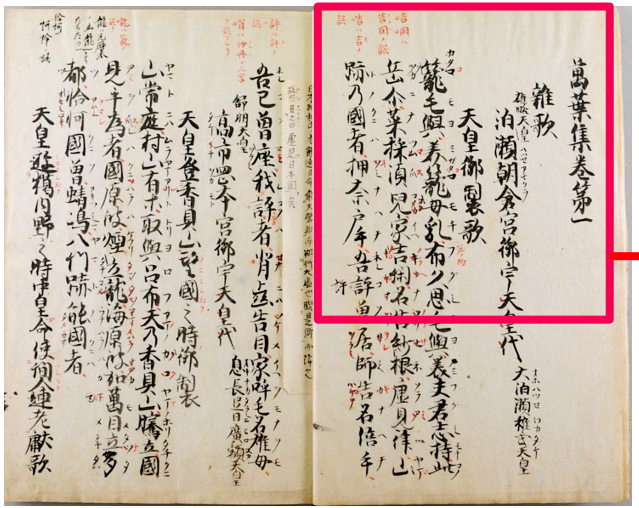
TEIの目的と利点（3）

利点

1. **互換性の確保**　研究者同士で簡単にデータ共有
2. **長期保存**　将来の技術変化にも対応しやすい
3. **検索と分析の効率化**　後から特定の情報を探したり、特定の項目に基づいてデータを分析するのが簡単に

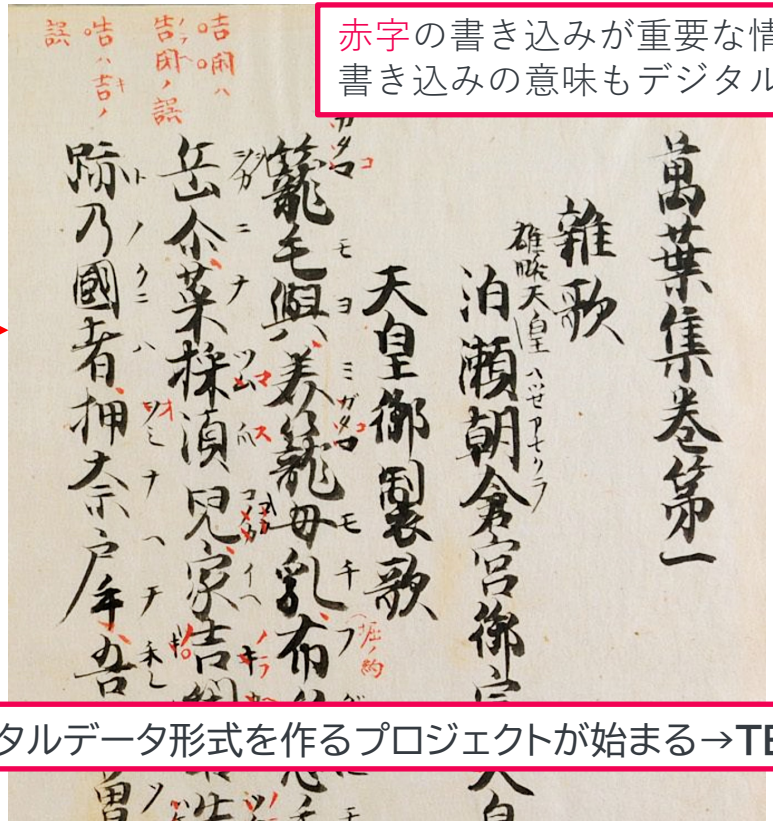


TEIの歴史と背景(1)



「萬葉集 20巻」関西大学アジア・オープン・リサーチセンター[KU-ORCAS]

赤字の書き込みが重要な情報
書き込みの意味もデジタル化したい



1987年に人文学のための統一的なデジタルデータ形式を作るプロジェクトが始まる→TEIへ

TEIの歴史と背景(2)

SGML(Standard Generalized Markup Language

- 主に出版業界などで使用されていた、複雑な文書を構造化するための技術
- 広く一般に普及できなかった

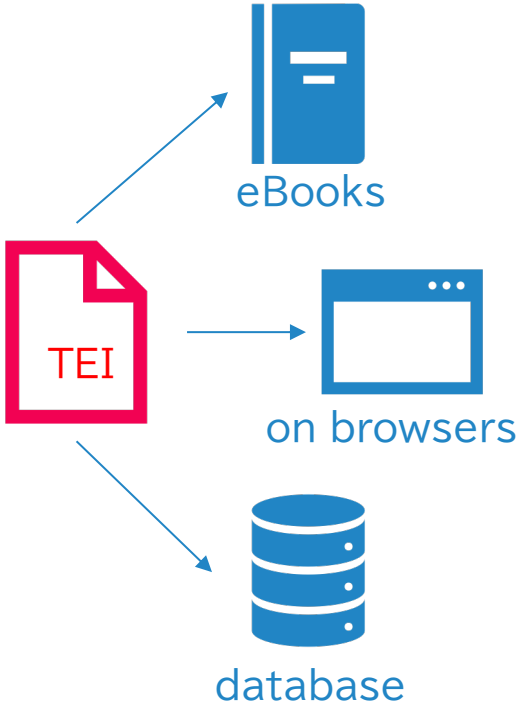
XMLの登場

- 2000年代により多くの人にとって使いやすいXML（Extensible Markup Language）が登場
- XMLに基づくTEIガイドライン(利用案内)に再構築

TEIの歴史と背景(3)

XMLの特徴

1. **タグを自由に決められる**
   使うタグ(ラベル)を自分で自由に決めることができる
2. **拡張性が高い**
   データの構成や内容に合わせて自由にタグを定義できるため、様々な種類のデータや用途に対応できる
3. **構造の変換できる**
   XMLはデータの構造を別の形式に変換できる
   ＝同じデータをさまざまな目的で活用できる



TEIの歴史と背景(4)

The TEI's conference and annual general meetings https://members.tei-c.org/Events/meetings

ENCODING CULTURES

JOINT MEC AND TEI CONFERENCE 2023

TEIの構成(1)

家（間取り図）

- リビングルーム

--ソファ

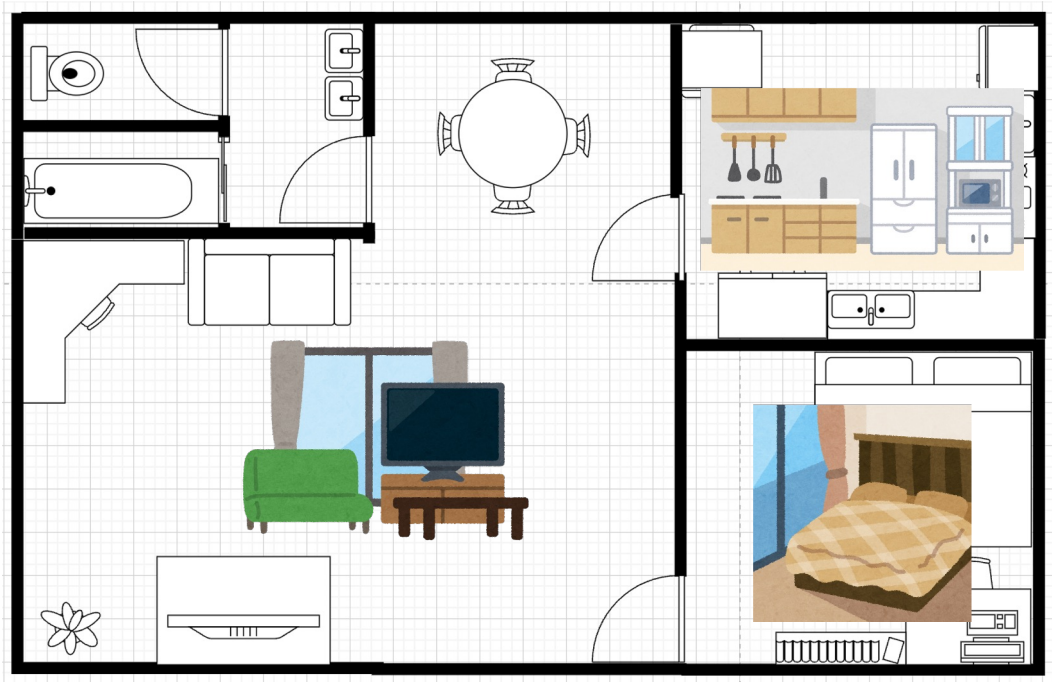
- キッチン

-- 冷蔵庫

- 寝室

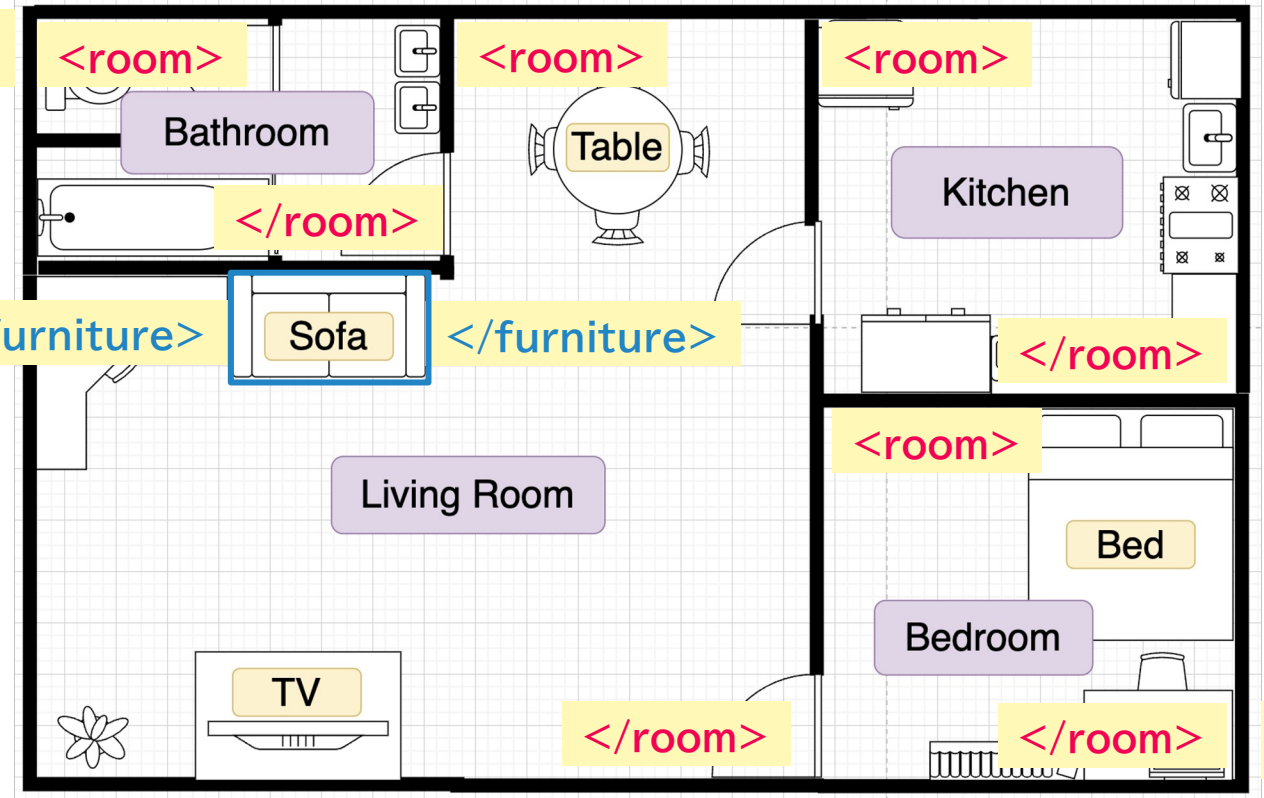
-- ベッド

各部屋＝章、段落などのセクション



TEIの構成（2）

<house>



</house>

TEIの構成（3）

1. メタデータタグ
   - テキスト全体に関する情報を示すタグ（誰が書いたのか、いつ書かれたのか、どこで発行されたのかなど）
2. 構造タグ
   - テキストの大きな構成要素を示すタグ（章や段落、セクションなど）
3. 内容タグ
   - テキストの中身を説明するタグ（見出し、名前、日付、引用など、具体的なテキストの内容

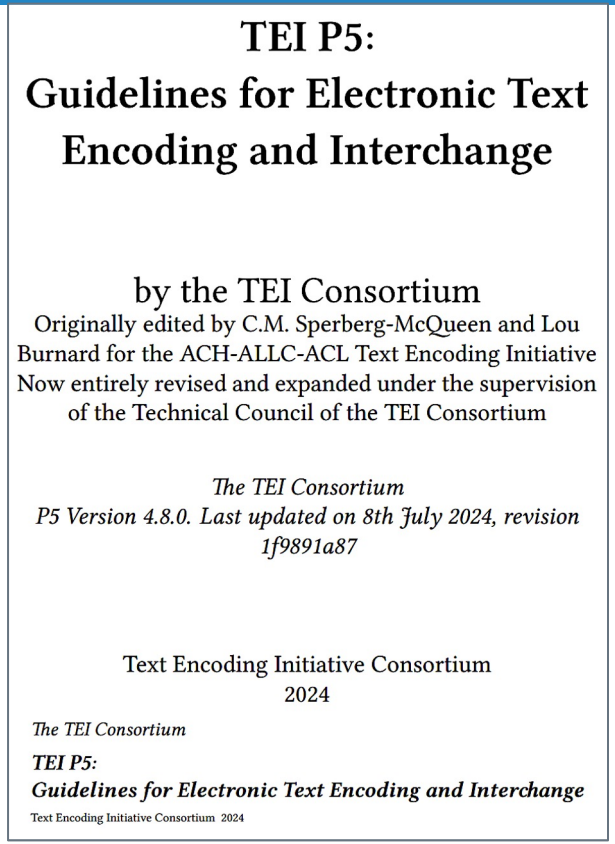
TEIガイドライン(1)

「TEIガイドライン」のページ

テキストとその構造、注釈などの周辺情報へのタグ付けに関するガイドライン

https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

(“TEI guideline” で検索してください)



TEI P5:
Guidelines for Electronic Text Encoding and Interchange

by the TEI Consortium
Originally edited by C.M. Sperberg-McQueen and Lou Burnard for the ACH-ALLC-ACL Text Encoding Initiative
Now entirely revised and expanded under the supervision of the Technical Council of the TEI Consortium

The TEI Consortium
P5 Version 4.8.0. Last updated on 8th July 2024, revision 1f9891a87

Text Encoding Initiative Consortium
2024

The TEI Consortium
TEI P5:
Guidelines for Electronic Text Encoding and Interchange
Text Encoding Initiative Consortium 2024

TEIガイドライン(2)

目的

TEIデータの標準化手法を公開する

TEIガイドライン(3)

```
<text>
  <body>
    <div type="poem">
      <lg type="stanza">
        <l>静かな森の中、</l>
        <l>月明かりが照らす道を歩く。</l>
        <l>冷たい風が吹き抜け、</l>
        <l>心の中に思い出が蘇る。</l>
      </lg>
    </div>
  </body>
</text>
```

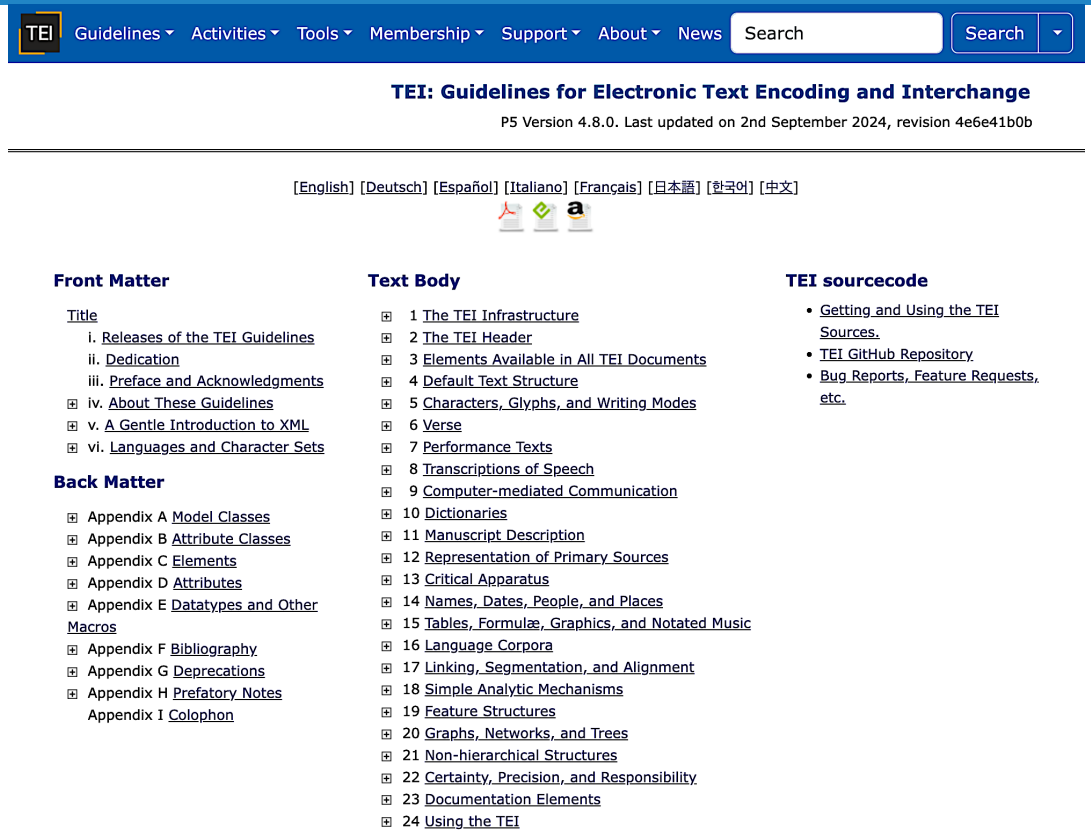
↑詩

手紙→

```
<text>
  <body>
    <div type="letter">
      <opener>
        <salute>親愛なる友へ、</salute>
      </opener>
      <p>長い間、手紙を出せずにすみません。最近は忙しい日々が続いていますが、
      あなたのことをよく思い出しています。</p>
      <closer>
        <salute>敬具、</salute>
        <signed>太郎</signed>
      </closer>
    </div>
  </body>
</text>
```

日記→

```
<text>
  <body>
    <div type="diary">
      <div type="entry" xml:id="entry1">
        <head>2024年9月25日</head>
        <p>今日はとても忙しい一日だった。朝から仕事に追われ、昼食も取れずに過ごした。</p>
      </div>
      <div type="entry" xml:id="entry2">
        <head>2024年9月26日</head>
        <p>今日は少しゆっくりと過ごせた。天気も良く、散歩を楽しんだ。</p>
      </div>
    </div>
  </body>
</text>
```

TEIガイドライン（4）

TEI Guideline トップページ→