人文学向けにテキスト資料を構造化する TEIガイドライン ～そろそろ日本でも面白くなってきました～

永崎研宣（一般財団法人人文情報学研究所）

TEI (Text Encoding Initiative)

- 「TEI協会」が作成公開する「TEI ガイドライン」
- 人文学の電子テキスト資料を構造化すべく1987年に始まった
  - 最初はSGML、現在はXMLで書かれるようになっている。
  - 欧米ではデファクトスタンダード的な感じに
  - 図書館の人もかなり関わっている
    - 理事にはオーストリア国立図書館の人、技術委員にはノースイースタン大学図書館の人、主力メンバーにも色々な図書館に
  - TEI協会
    - https://tei-c.org/
  - TEIリポジトリ
    - https://tapasproject.org/
  - 日本では最近ようやく取り組みが本格化しつつある

デジタル知識情報基盤の中でのTEIの良さ

- 始まり：ルール不在の電子テキスト界で世界の関係者が集まってテキスト作成の共通ルールを作ることになった。
  - 当初は「電子テキストにヘッダ情報を同梱する共通ルール」が重視された模様

- 現在：
  - リニアなテキストデータから知識を抽出した後、抽出した知識の出所に戻れる（構造化された知識の検証可能性を担保）
    - そこに至るための様々な便利なルールやツール類

TEIの（割と有名な）事例：

- ゴッホの手紙
  - http://vangoghletters.org/vg/
  - 地名・人名等を手紙のマークアップテキストから抽出
  - マークアップテキストも丸ごと公開
- シェイクスピアの戯曲
  - シェイクスピアの戯曲を詳細にマークアップして様々な抽出ができるように
  - マークアップテキストだけでなくAPIも提供:
    - https://www.folgerdigitaltexts.org/api?_ga=2.113712903.1070924881.1592716473-1458558134.1592383887
    - APIの例：ハムレットで誰がいつ登場しているかを抽出：
      - https://www.folgerdigitaltexts.org/Ham/charChart

TEI協会東アジア／日本語分科会

- TEI-EAJ
- https://github.com/TEI-EAJ/
- 各種活用ツール類（主に東京大学中村覚氏が作成）
- 「青空文庫テキストをより便利にする（機械可読性を高める）ためのプロジェクト」
  - https://github.com/TEI-EAJ/aozora_tei
- 「日本語資料のための日本語によるガイドライン」
  - https://github.com/TEI-EAJ/jp_guidelines/wiki
- 7/2（木） 18:00～ 「**TEIで使えるツールを調べる会**」@Zoom
  - 参加者募集中！
  - https://www.facebook.com/events/2879859458806533/