DH ・ 技術要素 ・ TEI / XML

TEIとXML入門

人文学テキストを「データ」にする

DH入門 ／ 技術要素シリーズ

中村

※実験的な取り組みです（構成・図・AI音声合成を含む）。内容はご確認・ご注意のうえご利用ください

この動画について

- ✓ **クリエイティブ・コモンズ**のオープン教材を参照し、独自に構成した解説です
- ✓ スライド・図は新規作成、ナレーションは**本人声のAI音声合成**
- ✓ これは**実験的な取り組み**です。内容は**ご確認・ご注意のうえ**ご利用ください
- ✓ 誤りに気づいたら概要欄からご指摘ください。出典・ライセンスは末尾と概要欄に記載しています

この回のゴール

テキストを「構造をもつデータ」として扱う考え方をつかむ

- ✓ TEIとXMLが、テキストの**何を**記述する仕組みかを説明できる
- ✓ 要素・属性・入れ子という**基本構造**を読める
- ✓ ヘッダ（メタデータ）と本文符号化の**役割の違い**がわかる
- ✓ 符号化が**解釈をともなう行為**だと説明できる

前提知識は特にありません。コードもほとんど書きません。

今日の流れ

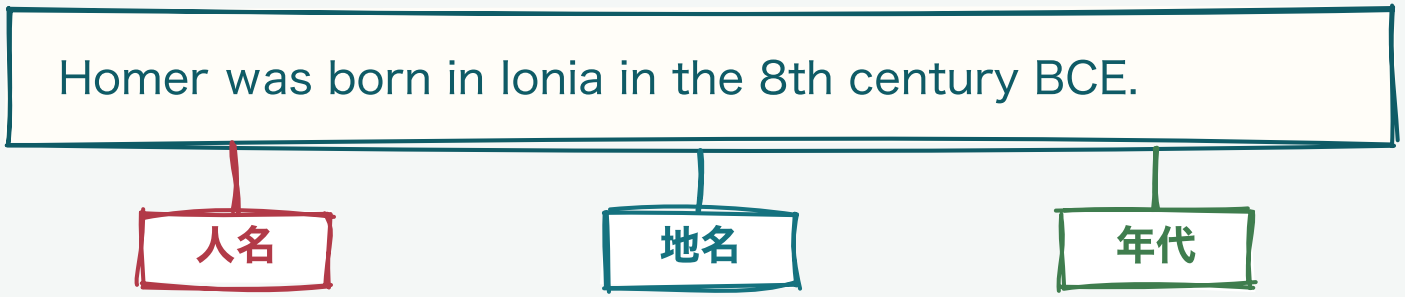
- ✓ テキストを「データ」にするとは（XMLの考え方）
- ✓ TEIとは何か
- ✓ なぜ「標準」を使うのか、どこで使われているか
- ✓ 符号化は「解釈」である
- ✓ 始め方・学ぶには

テキストを「データ」にするとは

まずは XML の考え方から

一文の中には、いくつもの情報がある

同じ一文に、いくつもの種類の情報が重なっている



人が読むときは、人名・地名・年代を**自然に区別**している

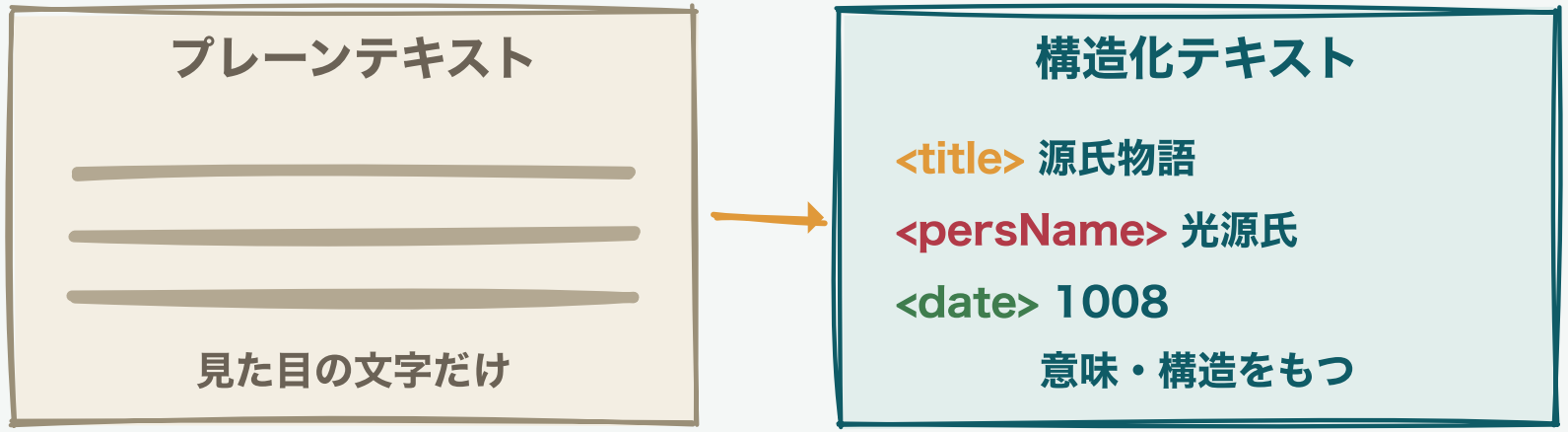
人は読んで区別できる。では機械は？

✓ わたしたちは「ホメロス」を**人名**、「イオニア」を**地名**と、無意識に見分ける

✓ けれど、ただの文字の並びを渡された機械には、その区別は**見えていない**

✓ 「ここは人名」「ここは年代」と**印をつけて**あげる必要がある

プレーンテキストの限界



見た目の文字だけでは、**意味や構造**が残らない

マークアップ＝意味に「タグ」を付ける

語を「ここは人名」と印づける

<persName> Homer </persName>

開始タグ と 終了タグ で内容をはさむ

「この範囲は人名です」と、**目印**を文章に書き込む

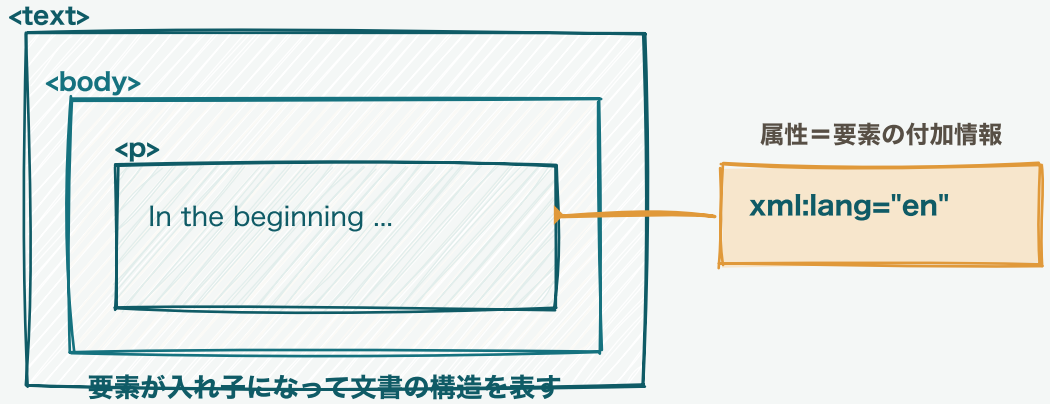
XMLの基本①　要素とタグ

開始タグと終了タグで内容をはさんだ、ひとまとまりを**要素**と呼ぶ

```
<persName>Homer</persName>
```

<persName> 開始タグ　／　**</persName>** 終了タグ　／　はさまれた Homer が中身

XMLの基本②　入れ子と属性



- ✓ 要素の中に要素を入れる＝**入れ子**で構造を表す
- ✓ タグに **属性** を足して、付加情報を書ける
- ✓ 例：その段落が何語か（xml:lang）

ここまでのポイント

- ✓ テキストには、見た目の裏に**人名・構造などの情報**がある
- ✓ それを機械にも分かる形にするのが**マークアップ**
- ✓ XMLは、要素・入れ子・属性で**構造を書き表す**書き方

では、人文学のテキストには「どんなタグ」を使えばよい？　→　そこで TEI

TEIとは

TEI＝人文学テキスト符号化の国際標準

- ✓ TEI は **Text Encoding Initiative**（テキスト・エンコーディング・イニシアティブ）の略
- ✓ 人文学のテキストに、**どんなタグをどう使うか**を定めた共通の約束ごと
- ✓ 規格であると同時に、それを支える**国際的なコミュニティ**でもある

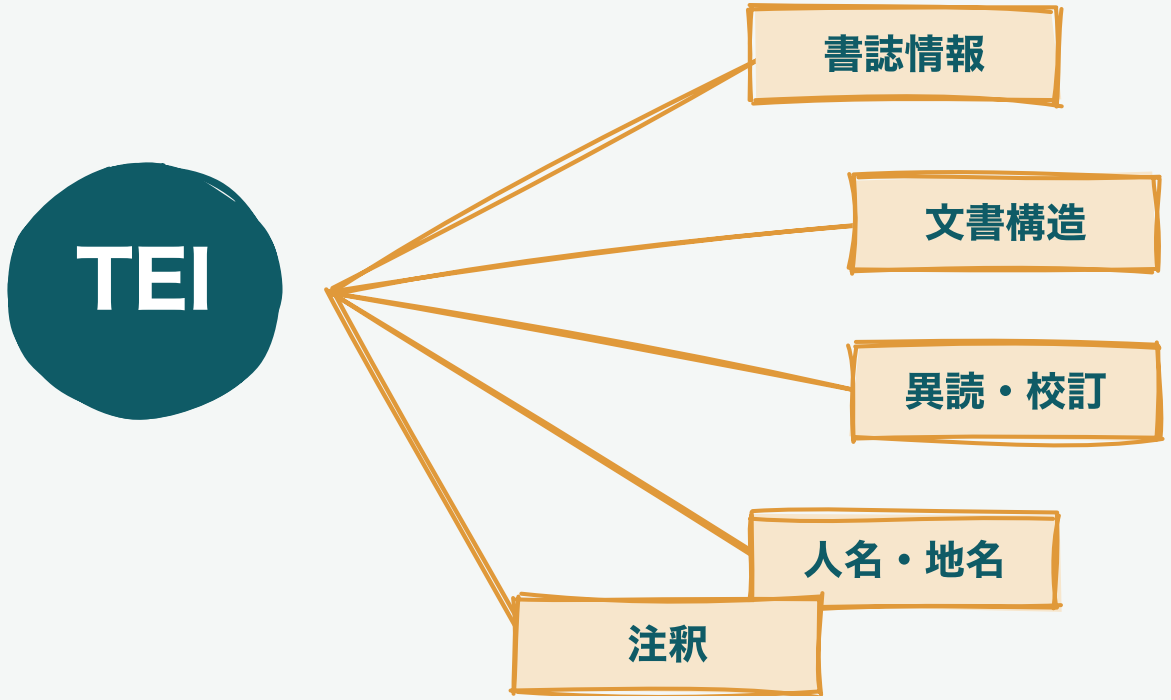
背景：研究者たちが育ててきた標準

✓ TEI は **1987年**に、国際的な共同の取り組みとして始まった

✓ 以来、世界中の研究者や図書館が議論を重ね、**改訂を続けている**

✓ いま広く使われているのは、**第五版（P5）**と呼ばれる版

TEIで記述できること



書誌から本文の細部まで、**幅広い対象**を一貫した形で書ける

TEIヘッダ：本文の「説明書き」

- ✓ TEI文書は**ヘッダ**と**本文**の2層でできている
- ✓ ヘッダ＝**メタデータ**。誰が・何を・どの底本から作ったか
- ✓ 本文＝説明書きではなく、**テキスト本体**そのものを収める層

本文を符号化してみる

```
<p>
  <persName>Homer</persName> was born in
  <placeName>Ionia</placeName>.
</p>
```

段落は <p>、人名は <persName>、地名は <placeName>

例を読み解く

- ✓ ** … ** が、ひとつの段落の範囲
- ✓ その中の **<persName>** が人名、**<placeName>** が地名を指す
- ✓ こうしておくと「人名だけ集める」「地名を地図に載せる」が**機械的にできる**

考えてみよう

あなたの手元の資料なら、**何にタグを付けたい**ですか？

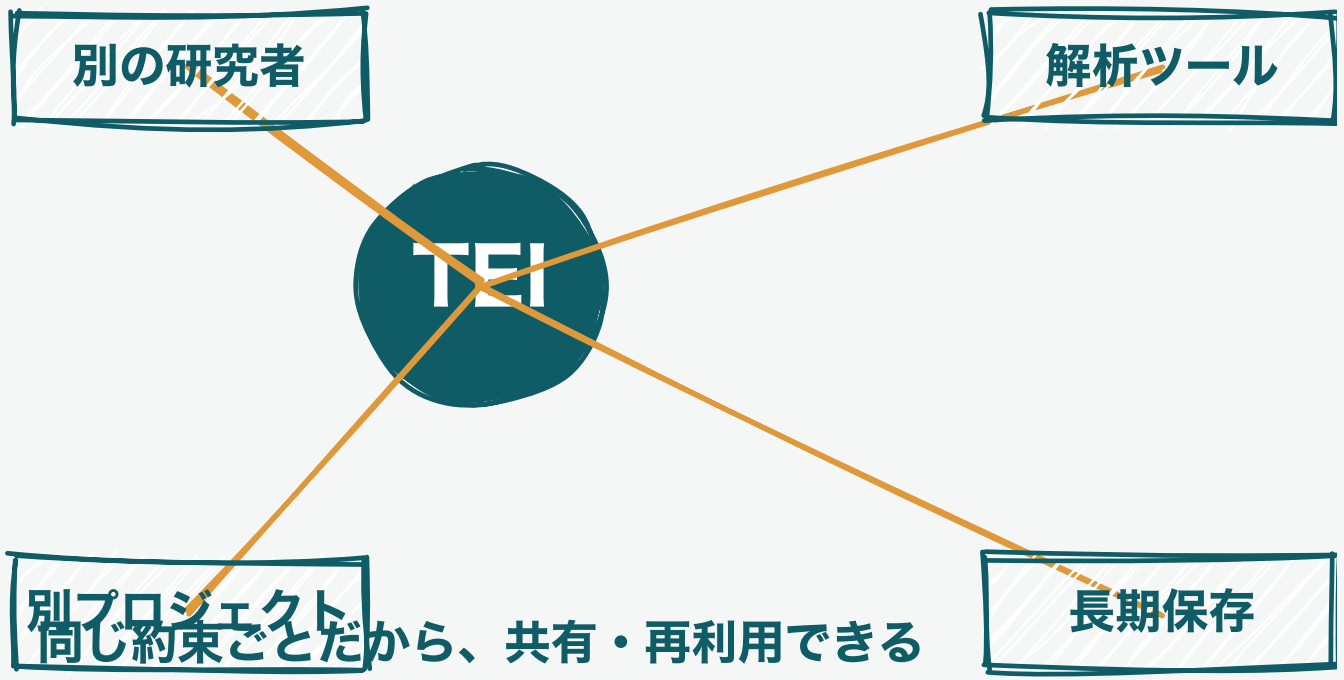
人名／日付／地名／引用 …　ここで少し、動画を止めて考えてみてください。

ここまでのポイント

- ✓ TEI は、人文学テキストのための**タグの共通規格**
- ✓ 文書は**ヘッダ（説明書き）**と**本文（タグ付き）**の2層
- ✓ 段落・人名・地名などを印づけると、**機械が扱える**ようになる

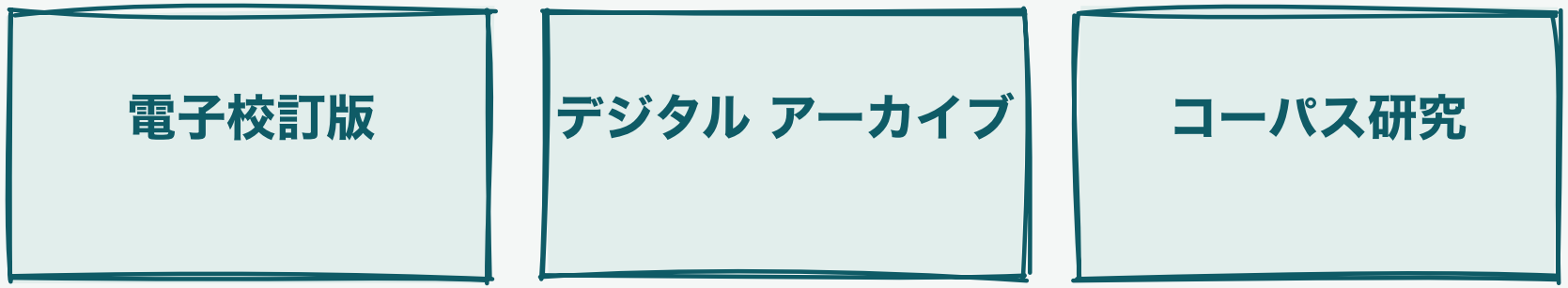
なぜ「標準」を使うのか

共通の約束ごとだから、つながる



他の人・別のツール・将来の自分とも**共有や再利用**がしやすい

どこで使われているか



研究の電子校訂版から、図書館の資料公開まで

たとえば、こんな使われ方

- ✓ **電子校訂版**：複数の写本の異読を重ね、画面で読み比べる
- ✓ **デジタルアーカイブ**：古典籍や文書を、構造ごと検索できる形で公開
- ✓ **コーパス研究**：大量のテキストを、語や構造の単位で分析する

符号化は「解釈」である

何にタグを付けるかを「選んでいる」

A <persName>Homer</persName> … in Ionia

B <persName>Homer</persName> … in <placeName>Ionia</placeName>

同じ文でも、**着目点によって符号化は変わる**

だから、唯一の正解は決まらない

- ✓ 「どこを・何として印づけるか」には、**研究上の判断**が入る
- ✓ これは欠点ではなく、**解釈を明示できる**という長所でもある
- ✓ 符号化は、機械作業のようでいて、すぐれて**人文学的な営み**と言える

始め方・学ぶには

- ✓ 文書を編集する：**oXygen**（商用のXMLエディタ）
- ✓ 自分用に仕立てる：**Roma**（カスタマイズ＝ODDを作る無料のWebツール）
- ✓ 例で学ぶ：**TEI by Example**（モジュール別のチュートリアル）
- ✓ 体系的に：**DARIAH-Campus** の「Text Encoding and the Text Encoding Initiative」
- ✓ まずは小さなテキストを、自分でタグ付けしてみるのが近道

まとめ

- ✓ テキストを**構造をもつデータ**にするのが、マークアップとXML
- ✓ **TEI** は、人文学テキストのためのタグの国際標準
- ✓ 文書は**ヘッダ**と**本文**の2層。だから広く共有・再利用できる
- ✓ 符号化は「何を印づけるか」を選ぶ、**解釈をともなう行為**

テキストを「読む」だけでなく「構造化する」視点を、ひとつ手に入れた、と言えます

出典・ライセンス

本動画は、以下のオープンライセンス教材を参照して作成しました。

- ✓ Text Encoding and the Text Encoding Initiative ／ S. Schreibman・R. Bleier（DARIAH-Campus）— CC BY 4.0
- ✓ TEI Guidelines (P5) ／ TEI Consortium — CC BY 3.0 ＋ BSD 2-Clause（デュアル）
- ✓ DARIAH-DE TEI Tutorial ／ DARIAH-DE — CC BY 4.0

スライド・図は中村による新規作成（概念を参照し、表現は新たに構成）。

ご清聴ありがとうございました