

FIAS REGISTRY OF RESEARCH SPECIFICATIONS

# TEI-Pivot Microformat for Source-Interpreter-Translation Alignment in Small Training Corpora

REGISTRY IDENTIFIER

FOAR-CON2X-49S4MF19J

OBJECT TYPE

Data model or interchange format

DOCUMENT MATURITY

Stable

PUBLISHED

15 December 2025, 12:00 UTC

AUTHORITATIVE VERSION

Registration records the outcome of the Registry review for this version: the submission was sufficiently identified, documented, and internally determinate for the stated procedure. Scope, adoption, and competing conventions are documented separately where relevant.

VERSION

1.0

SUBJECT AREA

Linguistics and philology

SUBMITTED

08 September 2025, 12:00 UTC

LICENCE

Creative Commons Attribution 4.0

REGISTRY STATUS

**Registered**

15 December 2025, 12:00 UTC

# Core description

---

**SUBMISSION KIND**

New registry concept

**FULL TITLE**

TEI-Pivot Microformat for Source-Interpreter-Translation Alignment in Small Training Corpora

**SHORT TITLE OR ACRONYM**

TEI-SIT

**OBJECT TYPE**

Data model or interchange format

**STATUS CLAIMED BY THE SUBMITTER**

Project standard

**MATURITY OF THIS VERSION**

Stable

**VERSION**

1.0

**PRINCIPAL SUBJECT AREA**

Linguistics and philology

**KEYWORDS**

- interpreting studies
- translation studies
- TEI P5
- oral corpus
- temporal alignment
- stand-off annotation
- interpreter training

**PROBLEM, SCOPE, AND EXCLUSIONS**

Small interpreter-training studies often retain source speech, an oral interpreted rendition, a later written translation and pedagogical notes, and need exchange mechanisms beyond project-specific spreadsheets. TEI-SIT prescribes a constrained TEI P5 representation for sessions, speakers, timed utterance segments, two kinds of rendition, many-to-many alignment, interpreter self-repair and analyst notes. It covers consenting adult classroom or

public-domain material in pairs of languages. Its scope is limited to exchange representation; transcription orthography, translation-quality scores, automatic alignment, biometric speaker data, professional certification, and restricted-audio storage remain outside the XML profile.

#### INTENDED USERS AND CONTEXTS

- Corpus-based interpreting and translation researchers
- Interpreter-training seminars exchanging small consented datasets
- Digital-humanities data stewards converting ELAN or spreadsheet exports to TEI
- Developers validating source-rendition alignment before analysis

#### RELATION TO EXISTING PRACTICE

Profiles an existing standard

#### SCIENTIFIC CONTEXT, SOURCES, AND SCHOLARLY DISCUSSION

This section describes the scientific background, underlying work, and sources of the specification, together with the research groups and scholarly settings in which the proposed solution or its formalisation was presented or discussed. It documents the context and nature of these scholarly exchanges.

Interpreter-training corpora bring together a source utterance, its oral interpretation, a subsequent written translation and the annotations used to analyse the exercise. Comparing these materials requires an explicit account of segment identity, timing and alignment. L. Liégeois et al. (2016) demonstrate the use of TEI as a pivot format for oral and multimodal corpora. C. Bendazzoli, M. Russo and B. Defrancq (2018), together with M. Russo (2021), place machine-readable corpora at the centre of empirical work on interpreting and its pedagogical applications.

Corpus metadata and annotation interoperability were addressed in CORLI's working meeting at the Maison de la Recherche, Sorbonne Nouvelle, on 28 November 2024, and in the CORLI-MoDyCo workshop at Université Paris Nanterre on 23 June 2025. The latter combined demonstrations with participants' projects, covering inter-annotator agreement, file interoperability and conversion within annotation workflows. These activities form part of the methodological context for exchanging aligned and annotated language resources.

TEI-SIT applies these methods to small training datasets through the identifiers, speech elements, timelines and stand-off links available in TEI P5. Separate rendition types preserve the distinction between oral interpretation and written translation; many-to-many links record alignment across different segmentations. Repair annotations and analytical notes retain the observations needed for comparison and classroom discussion. Fixing the profile to TEI P5 4.9.0 gives the exchange format a reproducible technical reference.

The research setting combines corpus linguistics and interpreting studies, with the LLF (UMR 7110) and its computational-linguistics platform providing a relevant environment for corpus preparation and exchange. The TEI community and the corpus-based interpreting

literature supply the shared methods through which alignment, annotation and pedagogical reuse can be examined across research and training groups. The profile formalises a common representation for that work, allowing the same source-interpretation-translation relationships to be inspected in different tools.

#### References:

- TEI Consortium (2025), TEI P5 4.9.0 release (24 January 2025), Source: <https://tei-c.org/release-4-9-0/>
- TEI P5, Linking, Segmentation and Alignment, Source: <https://www.tei-c.org/release/doc/tei-p5-doc/en/html/SA.html>
- Liégeois, Étienne, Parisse, Benzitoun & Chanard (2016), Using the TEI as pivot format for oral and multimodal language corpora, Source: <https://shs.hal.science/halshs-01357343>
- Bendazzoli, Russo & Defrancq (2018), Corpus-based Interpreting Studies: a booming research field, Source: <https://www.intraline.org/specials/article/2304>
- Russo (2021), Corpus-based interpreting studies, DOI 10.1075/hts.5.cor2, DOI: <https://doi.org/10.1075/hts.5.cor2>
- RFC 6129 (2011), The application/tei+xml Media Type, Source: <https://www.rfc-editor.org/rfc/rfc6129>
- CORLI. Working meeting on metadata, Maison de la Recherche, Sorbonne Nouvelle, 28 November 2024. Programme: <https://corli.huma-num.fr/fr/events/reunion-de-travail-metadonnees/>
- CORLI-MoDyCo. Advanced annotation workshop, Université Paris Nanterre, 23 June 2025. Programme: <https://corli.huma-num.fr/fr/events/inception-avance/>
- LLF (UMR 7110). Research themes: [https://www.llf.cnrs.fr/fr/themes\\_de\\_recherche](https://www.llf.cnrs.fr/fr/themes_de_recherche)
- Université Paris Cité. LLF computational-linguistics platform: <https://plateformes.u-paris.fr/plateau-de-linguistique-informatique-du-llf-laboratoire-de-linguistique-formelle-umr7110/>

#### EXTERNAL STANDARDISATION STATUS

None intended

#### LICENCE FOR THE SUBMITTED SPECIFICATION

Creative Commons Attribution 4.0

#### PUBLICATION DECLARATION

Yes

## Data model or interchange format

---

### ENTITIES, RECORDS, AND RELATIONSHIPS

SESSION describes one recorded or textual exercise and links consent, source language and date. PERSON is a pseudonymised speaker, interpreter, translator or analyst. MEDIA references an external audio/video object and its checksum. TIMELINE contains ordered time points. UTTERANCE is a speaker-attributed `u`; SEGMENT is an addressable `seg` inside a `u`. RENDITION is a segment whose role is `source`, `interpreted` or `writtenTranslation`. ALIGNMENT is a stand-off TEI `link` connecting one or more source segments to one or more rendition segments. REPAIR is a stand-off span over an interpreted rendition. NOTE targets one entity and has a controlled pedagogical or analytic function. Relationships are identifier references; Alignment is asserted only through explicit identifier references.

### SERIALIZATION AND MEDIA TYPE

XML 1.0 encoded in UTF-8, rooted in TEI P5 `TEI` and served as `application/tei+xml` as registered by RFC 6129. Canonical files use the TEI namespace `http://www.tei-c.org/ns/1.0`. A human-readable HTML view and CSV alignment export may be derived, but XML is normative. Audio and video remain external and are referenced by URI plus SHA-256 checksum.

### REQUIRED FIELDS AND CARDINALITIES

TEI/@xml:id: exactly 1 per document

teiHeader/fileDesc/titleStmt/title: exactly 1

teiHeader/fileDesc/publicationStmt: exactly 1

teiHeader/fileDesc/sourceDesc: exactly 1

profileDesc/langUsage/language: 2..n, each with @ident (BCP 47)

profileDesc/particDesc/person: 1..n, each with @xml:id and controlled @role

recordingStmt/recording/media: 0..n; if present, @url, @mimeType and SHA-256 note required

timeline/when: 2..n for timed speech, each with @xml:id and monotonic @absolute or @interval

text/body/div[@type='session']: 1..n, each with @xml:id and @xml:lang

u: 1..n per session, with @xml:id and @who; @start/@end required when media exists

seg: 1..n per u, each with @xml:id and @type in source|interpreted|writtenTranslation

standOff/linkGrp[@type='sit:alignment']: exactly 1

link: 1..n, each with @xml:id, @type in oral|written and @target with at least two segment references

spanGrp[@type='sit:repair']: 0..1; every span has @xml:id, @from, @to and @ana

note[@type='sit:pedagogical'|'sit:analytic']: 0..n, each with @target and @resp

**DATATYPES, CONTROLLED VALUES, NULLS, AND UNKNOWN VALUES**

`xml:id` and local references obey XML ID/URI-reference rules. Languages use BCP 47 strings such as `fr` and `en`. Roles are controlled values `speaker`, `interpreter`, `translator`, `analyst`; segment type is `source`, `interpreted`, or `writtenTranslation`; alignment type is `oral` or `written`; repair analysis is `repetition`, `substitution`, `deletion`, `insertion`, or `unclassified`. Time is non-negative decimal seconds from media start. Dates are ISO 8601. Checksums are lower-case hexadecimal SHA-256. Missing optional information is absent, never an empty string. Unknown language is `und`; unknown time omits start/end and must not be invented; unknown repair category is `unclassified`.

**MACHINE VALIDATION MECHANISM**

XML Schema

**VALIDATOR, SCHEMA, OR GRAMMAR DETAILS**

A TEI ODD customization generates Relax NG and XSD artefacts pinned to TEI P5 4.9.0. XSD validates namespace, element vocabulary and base datatypes. Schematron rules enforce role/type value sets, uniqueness of identifiers, resolvable references, at least one source and one rendition target per alignment, monotonic timelines, `oral` links targeting interpreted segments, `written` links targeting writtenTranslation segments, and prohibition of direct identifiers in PERSON.

**FORWARD AND BACKWARD COMPATIBILITY**

Every TEI-SIT 1.0 document is intended to remain a valid TEI P5 document. Generic TEI processors may ignore `sit:\*` controlled values and still preserve text and links. A minor release may add optional roles, note types or repair categories; consumers must preserve unknown optional values. A major release is required to rename elements, change link target order or make optional fields mandatory. Converters must retain the declared TEI-SIT version in `appInfo`. Derived CSV is lossy and may not be converted back without the XML. Pinned ODD/XSD files remain available for every registered version.

**COMPLETE MINIMAL EXAMPLE**

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="sit-example-01">
  <teiHeader>
    <fileDesc>
      <titleStmt><title>TEI-SIT example</title></titleStmt>
      <publicationStmt><p>CC BY 4.0</p></publicationStmt>
      <sourceDesc><p>Example of source, interpretation and written translation
alignment</p></sourceDesc>
    </fileDesc>
    <encodingDesc>
      <appInfo><application ident="TEI-SIT" version="1.0"/></appInfo>
    </encodingDesc>
```

```

    <profileDesc>
      <langUsage><language ident="fr">French</language><language
ident="en">English</language></langUsage>
      <particDesc><listPerson>
        <person xml:id="p-src" role="speaker"><persName>S01</persName></person>
        <person xml:id="p-int" role="interpreter"><persName>I01</persName></person>
        <person xml:id="p-tr" role="translator"><persName>T01</persName></person>
      </listPerson></particDesc>
    </profileDesc>
  </teiHeader>
  <text><body><div type="session" xml:id="session-1" xml:lang="fr">
    <u xml:id="u-src" who="#p-src"><seg xml:id="s1" type="source">La séance
commence.</seg></u>
    <u xml:id="u-int" who="#p-int" xml:lang="en"><seg xml:id="i1"
type="interpreted">The session starts.</seg></u>
    <u xml:id="u-tr" who="#p-tr" xml:lang="en"><seg xml:id="t1"
type="writtenTranslation">The session begins.</seg></u>
  </div></body></text>
  <standOff>
    <linkGrp type="sit:alignment">
      <link xml:id="a1" type="oral" target="#s1 #i1"/>
      <link xml:id="a2" type="written" target="#s1 #t1"/>
    </linkGrp>
    <note type="sit:pedagogical" target="#i1" resp="#p-tr">Lexical contrast only; not an
error score.</note>
  </standOff>
</TEI>

```

## Registry review record

---

### Submission received

08 September 2025, 12:00 UTC

Submission received.

### Registered

15 December 2025, 12:00 UTC

Registered as a narrow TEI P5 project profile. Documentary review confirms that the cited TEI and corpus-interpreting sources predate the submission. Technical review confirms a stable conformance target, resolvable many-to-many links, controlled unknown values, specified ODD/XSD validation rules, and a complete minimal example. The specification appropriately leaves transcription, translation-quality judgement and restricted-media access outside scope.