MANHATTAN TRANSFER: TEI-BASED TEXT ENCODING

Description of the encoding process following TEI guidelines

Alessandra Failla, Laura Travaglini

DOWNLOAD THE FULL ENCODED TEXT HERE

Abstract *This article covers the process of encoding the novel Manhattan Transfer by author John Dos Passos following the guidelines of the Text Encoding Initiative. Some relevant categories were identified as the focus for the encoding for their potential as base for further analyses. Beyond the encoding of the structure of the novel, for which adequate TEI elements were selected, the encoding focused on the categories People, Places, Organizations, Language, Speech and Thought, Times and Measures.*

Introduction

This section covers the methodology of the encoding of the novel *Manhattan Transfer* by American novelist John Dos Passos, published in 1925. The novel deals with the stories of several characters and is set New York between the late 1890s and the 1920s, in the time often referred to as the *Jazz Era*. It is structured into three sections and each section is divided into five to eight chapters.
This report describes the process and choices related to the encoding of the novel, based on the guidelines of the *Text Encoding Initiative*. It focuses on different categories introduced in the following sections to subdivide the elements and the attributes used for the encoding. These categories are People, Places, Organizations, Language, Speech and Thought, Time and Measures, and Varia. Some examples are available below, however the full encoded text is available as xml file here or clicking on the download link below the title.

Resources and Methodology

The novel *Manhattan Transfer* is a public domain novel since it was published in 1925, thus more than seventy years ago. The full text is available on several online platforms; the text used by the authors of this work was downloaded from *Wikisource* (full text).
The novel was divided into two parts to carry out the encoding and each part was encoded by a member of the group. It was performed manually, both in a linear way or using regular expressions to find specific passages in the text and check if all occurrences were encoded correctly.
The primary resource used for the encoding of *Manhattan Transfer* are the guidelines of the *Text Encoding Initiative* (TEI guidelines), a consortium that develops and maintains the standards for the encoding of digital text. The explanations and examples contained in the guidelines were used to select the most appropriate elements and attributes.
The encoding related to the use of language by the author was developed based on the *Oxford Advanced Learner's Dictionary*, used as reference to define language variations, uncommon or informal uses, as explained in the Language section below.

Encoding Categories

The following sections illustrate the *TEI* elements and attributes used for the encoding of *Manhattan Transfer*, and the reasons behind each choice. They are divided into categories based on their use and meaning to allow the user to navigate more easily.

1. Structure

The initial **<teiHeader>** element, completely different from the *front matter*, supplies descriptive and declarative metadata associated with a digital resource or set of resources - in our case, the Manhattan Transfer version available in the public domain. Among the **<teiHeader>** tag's five principal components, we picked three: **<fileDesc>**, **<encodingDesc>**, and **<profileDesc>**.
<fileDesc> entails a file description that comprises a full bibliographical description of the computer file itself. It has generally three mandatory parts:

1. **<titleStmt>**:it contains child elements that provide basic metadata about the document, including the title of the resource (**<title>**), author (**<author>**), and editor names, as well as the names and roles of other people who contributed to the creation of the electronic document (**<respStmt>**).
2. **<publicationStmt>**: it contains basic child elements regarding publication information of the electronic text: **<publisher>** and **<distributor>**, **<date>**, **<availability>** and the specified **<licence>**.
3. **<sourceDesc>**: it contains child elements that describe the original source from which the electronic text was created. The following element was nested in it: **<bibl>** and **<listPerson>**:

   → **<bibl type="printSource">** and then its children: **<author>**, **<title>**, **<date>**, **<publisher>** as stated by the source.

A further element, **<extent>**, was added within **<fileDesc>**. This element provides infomation on the size of the novel. The type of measure and the related value are specified with the element **<measure>** and the attribute **@unit:**

```
<measure unit="words">
```

<encodingDesc>: documents the context in which the text was encoded.

<profileDesc>: provides a detailed description of non-bibliographic aspects of a text. We decided to specify:

- **<langUsage>** with its nested element **<language>**
- **<textClass>** through which we provided additional information about the topics of the text by providing a couple of **<keywords>** (and its nested tag **<term>**)
- **<textDesc>**, that contains the following elements

```
<channel mode="w"/>
<constitution type="single"/>
<derivation type="original"/>
<domain type="art"/>
<factuality type="fiction"/>
```

```
<interaction type="none"/>
<preparedness type="prepared"/>
<purpose type="entertain" degree="high"/>
```

Finally, the following element also appears in the **<teiHeader>**:

- **<revisionDesc>** and its child **<change>**, in order to provide information about the last revision of our work

The element **<standoff>**, generally used for contextual information and stand-off annotations was introduced after the teiHeader to include a list of the characters of the novel associated to their ID, using the element:

- **<listPerson>**, which contains a list of descriptions, each of which provides information about an identifiable person or a group of people, in our case the list of all Manhattan Transfer's characters. The TEI elements used to identify characters' names in the text are described thoroughly in the People section below.

After the **<teiHeader>** and the **<standoff>**, the front matter was introduced in the encoding. The **<front>** tag contains any prefatory matter (such as headers, abstracts, title page, prefaces, dedications, etc.) found at the start of our document, before the main body.

First, we have the **<titlePage>** with its additional info **<docTitle>** (and its child **<titlePart>**), **<docAuthor>,** and **<docDate>** containing information about the title, the author of the text and the date of publication.

Then two ** **, one for an intermediary repetition of the title and another **<div type="authorSWorks>** to list all the author's other works. This list features a **<head>** and the actual **<list>** with its various **<item>**.

The element **<figure>** was exploited when to mark the presence of a couple of illustrations present in the digitized version. Nested inside of it is the **<graphic>** which provides the *url* for the original figure.

Once again, the title is repeated with the **<titlePage>** tag and its child elements: **<docTitle>** (nested the actual **<titlePart>**),

- **<byline>**
- **<docAuthor>**
- **<docDate>**
- **<figure>**
- **<docImprint>** contains the imprint statement as given (usually) at the foot of a title page, in our case: the **<publisher>** (with the nested **<name>** tag) and **<pubPlace>** (with the nested **<placeName>**)

This repetition is due to the reproduction of the first pages of the digitised version, available here.

Another generic ** ** specified with the attribute **@type** and the value contents was created in order to list the structure of the book. The ** ** is the main

container, each section introduced by a **<head>** tag and followed by a list with the **<list>** tag and its child elements **<item>** for each chapter
Each section of the book was grouped by the **<group>** tag - it "contains the body of a composite text, grouping together a sequence of distinct texts (or groups of such texts) which are regarded as a unit for some purpose, for example the collected works of an author, a sequence of prose essays, etc." (TEI) - then followed by the **<text>** tag - a container element for the actual content - and **<body>** - a container for the whole body of a single unitary text, excluding any front or back matter.
Each section is nested inside a main ** ** container featuring two attributes, one for the number of the section and the other for the type of ** ** (ex. First Section: ** **). Same goes for the chapters (ex. ** **).
Each chapter is prefaced by a paragraph of what can only be described as a prose poem, which signals the theme of the chapter. For this "impressionistic paragraph" the tag **<epigraph>** was chosen: it identifies a quotation at the start of a division.

2. People

Characters' names were encoded using multiple elements, listed below. The most relevant among them were also given an ID in the form **#NameSurname** (or **#Name** if the surname is unknown). IDs are listed as attributes in the **<person>** element using the attribute **@xml:id** to assign a unique identifier to each character. These can be found in the characters' list contained in the element **<listPerson>** included in the **<standOff>** element. The element **<standOff>** can be used to include information that is not meant to appear on screen. In this case this element was used to include information about individuals mentioned in *Manhattan Transfer*, including their IDs. These are used to identify the same characters in the text and are included in the element **<persName>** using the attribute **@ref**, in the form **ref="#NameSurname"**. Find a list of all elements used for the encoding of individuals in the following paragraphs:

- **<persName>** contains a proper noun or noun phrase referred to a person. It can contain one or more of the elements listed below. If related to a relevant character of the novel, the element **<persName>** includes the attribute **@xml:id** containing the identifier of the character:

```
<persName ref="#EllenThatcher">
      <forename>Ellen</forename>
</persName>
```

- **<forename>** contains characters forenames or most used name. In some cases, the most used name was encoded as forename, as for Jimmy Herf; Jimmy's first name is James, but since the character is referred to as Jimmy in the greatest part novel to distinguish him from his cousin James Merivale, the form Jimmy was adopted as his forename. The occurrences of his birthname *James* were encoded using the attribute **@type** with value *birth* (example below). The element **<forename>** is also used to encode second names, identified through attribute **@sort**, with value *2*; if the second name is abbreviated, the attribute **@full** with value *abb* was introduced:

```
<persName ref="#JimmyHerf">
      <addName type="birth">James</addName>
</persName>
```

```
<persName ref="#PhineasBlackhead">
      <forename>Phineas</forename>
      <forename sort="2" full="abb">P.</forename>
      <surname>Blackhead</surname>
</persName>
```

- **<surname>** contains characters' family (inherited) name; in case of changes of surnames after marriage the attribute **@type** with value *married* was used:

```
<persName ref="#EllenThatcher">
      <addName type="nickname">Elaine</addName>
      <surname type="married">Oglethorpe</surname>
</persName>
```

- **<roleName>** contains components referred to a particular role or position of the character in society; examples are among others honorifics, royal titles, office, occupation, military. Role names can be abbreviated in the forms *Mr.*, *Mrs.*, *Dr.* etc. In this case, the attribute **@full** with value *abb* was added:

```
<persName ref="#GeorgeBaldwin">
      <roleName type="honorific" full="abb">Mr.</roleName>
      <surname>Baldwin</surname>
</persName>
```

- **<addName>** contains an additional name component, such as a nickname. In the encoding of *Manhattan Transfer* it was used for nicknames or name variations, such as the ones related to one of the main characters, Ellen Thatcher, referred to as *Ellie*, *Helena* or *Elaine*:

```
<persName ref="#EllenThatcher">
      <addName type="nickname">Ellie</addName>
</persName>
```

Further elements were used to encode information about individuals, not strictly related to their names. These are:

- **<occupation>**, used to encode information about a person's occupation or job.
- **<age value="#">**, used to encode information about a person's age, indicating the value as attribute.
- **<nationality>**, used to encode information about a person's nationality.

3. Places

When it came to the places cited in the text, we made a distinction between:

1. **<geogName>** for all the places associated with some geographical features specifying their type with other nested tags and
2. **<placeName>** for other, generic places.

And regarding the second type we decided to nest:

- **<geogFeat>** contains a common noun identifying some geographical feature contained within a geographic name, such as valley, mount, etc.
- **<bloc>** for the name of a geo-political unit consisting of two or more nation states or countries (ex. **<bloc type="continent">**).
- **<country>** for the name of a geo-political unit, such as a nation, country, colony, or commonwealth, larger than or administratively superior to a region and smaller than a bloc.
- **<region>** for any administrative unit such as a state, province, or county, larger than a settlement, but smaller than a country (ex. **<region type="state">**).
- **<settlement>** for settlement such as a city, town, or village identified as a single geopolitical or administrative unit.
- **<district>** for any kind of subdivision of a settlement, such as a parish, ward, or other administrative or geographic units. Given the American administrative configuration, we added type="borough" for the five main New York boroughs and type="neighborhood" for the various neighborhoods.
- **<address>** for postal address (and if necessary a **<num>** nested)
- **<offset>** for the direction of the offset between the two place names, dates, or times involved in the expression.

4. Organizations

Manhattan Transfer can be considered a big, complex picture about "how to make it - by any means - in New York". Therefore, numerous organizations, brands, etc. appear in the text. In order to highlight them we picked the befitting tag, **<orgName>**, and when necessary specified with additional nested tags (such as **<country>** or **<placeName>**) or an inner **type="partnerNames"**.

5. Language

Language variations and unusual forms are typical of John Dos Passos's literary style. In the encoding, **different** elements were adopted to express different kinds of language variations.

<orig> was used for the encoding of misspelled words. John Dos Passos frequently adopts incorrect spellings to imitate spoken language, dialects or accents. This is the case of words such as *fur* (sometimes used instead of for in direct speech), *nothin* or *nutten* (instead of nothing), *'em* (instead of them). The same element is used to encode incorrect grammatical forms, such as dont, wont, cant and so on, where the apostrophe was omitted for the author's choice. **<orig>** was also used to encode misspellings that imitate the accent of a foreign language, such as in the case of the character Marcus Antonius Zucher, a German man who Ed **Thatcher** meets at the beginning of the first section and that uses terms such as *vill* (will), *fif* (five), *vife* (wife) etc.

To provide a differentiation among the several possible language variations, the element **<distinct>** was introduced. This element allowed to specify particular language forms, such as slang or informal words, paired to the attribute **@type** (example X). The selection of these values was based on the definitions of the Oxford Advanced Learner's Dictionary, used as reference to specify a particular use of a word. The following values were used, as suggested by the dictionary:

- **<distinct type="informal">**, for words such as feller, yer, ain't, cops, yessir etc.

- **<distinct type="slang">** for words defined as slang or North American English, such as ye, howdy, bum (homeless person).
- **<distinct type="taboo">** used for swearing and curse words, defined as taboo by the Oxford Advanced Learner's Dictionary. These include profanities as well as swearing in different languages.
- **<distinct type="non-standard">** used for non-standard spellings of words, such as *dunno* ("don't know").

The elements **<orig>** and **<distinct>** were also nested in case of misspelling of forms included in the dictionary as one of the categories listed above. An example is the word *aint*, encoded as

```
<distinct type=”informal”>
      <orig>aint</orig>
</distinct>
```

or *awright*, encoded as

```
<distinct type="informal">
      <orig>awright</orig>
</distinct>
```

6. Speech and Thoughts

Dialogues are a prominent mean of communication in the novel. The element **<said>** was used for the encoding of direct and indirect speech, specifying the type of speech using the attribute **@direct** paired to either the values true or false. Whenever possible or relevant, the name of the speaking person was also introduced in this element using the attribute **@who**, paired with characters' IDs. **<said>** was also used to encode characters' thoughts, expressed in direct or indirect form, using the attribute **@aloud** paired with either the values true or false, and specifying the character's ID as in the case of speech. The inclusion of characters' IDs allows to keep track of characters mentions and interactions, that will be used in section 3 for the network analysis of the novel.
The element **<foreign>**, finally, was nested inside the **<said>** element whenever a character used a language different from English, as for example in the case of French, quite popular in the novel due to the presence of several French characters. The attribute @xml:lang was used to specify the language.

```
<said who="#JimmyHerf" aloud="true" direct="true">"I promise."</said>

<said who="#CongoJake" aloud="true" direct="true">
      <foreign xml:lang="fr">J'ai fait trois fois le tour du monde Dans mes
      voyages,</foreign>
</said>
```

7. Time and Measures

Measurements in time and money were managed through the following tags.
Money were **marked** with **<measure>**, a generic tag containing *"a word or phrase referring to some quantity of an object or commodity, usually comprising a number, a unit, and a commodity name"*, then better described through an attribute **<type="currency">**.

Precise dates were marked with the **<date when="">** tag and other moments in time instead with:

- **<time when="">** when better specified, for example like:

  ```
  <time when="08:30:00">half past eight</time>
  ```

- **<time dur-iso="">** when indicating a temporal duration, in which case we rely on the ISO 8601, an international standard covering time-related data:

  ```
  <time dur-iso="PT4D">four days</time>
  ```

8. Varia

In this last section, we are going to summarize the still not mentioned tags covering very different matters.
<floatingText> "contains a single text of any kind, whether unitary or composite, which interrupts the text containing it at any point and after which the surrounding text resumes" (TEI). We made use of this tag for example whenever a character would stop and read a piece of newspaper, a label, a sign, etc., which happens quite often.
Being so heterogeneous, *Manhattan Transfer* is full of characters from different extractions, may it be ethnic or religious or else. For this reason, we added the tag **<lang>** for the various mentions of the languages present.
<abbr> (abbreviation) tag was used for abbreviation of any sort. Some examples:

```
<abbr>Co.</abbr>
```

```
<abbr>D.S.C.</abbr>
```

Given the careers or work environment in which the characters are usually involved, Dos Passos's novel is full of songs, newspapers, musicals, theatrical works, etc.: for this category, we exploited the generic tag **<title>**, which - as the tei Guidelines states - "contains a title for any kind of work".
We made use of the salutation tag, **<salute>**, when needed, like for example at the very beginning of this letter:

```
<salute>
Dear
    <persName ref="#EllenThatcher">
        <addName type="nickname"> Elaine</addName>
    </persName>,
</salute>
```

Conclusion

The aim of this section was to introduce all TEI elements used for the encoding of Dos Passos's Novel *Manhattan Transfer* and the reasoning behind the choices. The selected TEI elements were selected based to represent the different types of phenomena present in the novel.
The encoding focused in particular on individuals and dialogues, in which characters' IDs are provided, and was performed thoroughly as a base for the third section of the project covering the network analysis of *Manhattan Transfer*. A further issue of the text encoding was related to linguistic phenomena: two elements were adopted to provide a specification of different uses of language and, in particular, a distinction between informal or slang words and the author's personal style.
The category related to places was as well provided with a detailed encoding was provided. Nonetheless, there is the possibility of extending the encoding by adding, for example, unique identifiers to element identifying the places mentioned in the text.
A further possible approach could be the analysis of linguistic phenomena based on the provided elements and on a possible extension of the encoding with further categories related to language phenomena, possibly integrating multiple dictionaries as resources. An analysis of such phenomena would allow a comparison with other literary works, opening up new perspectives of analysis with the use of computational means.

Bibliography and Sitography

- Text Encoding Initiative: Guidelines
- Oxford Advanced Learner's Dictionary
- Dos Passos, John: Manhattan Transfer (Wikisource)