Including XML Markup in the Automated Collation of Literary Text

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Astrid Kulsdom

*R&D - Huygens Institute for the History of the Netherlands*
*KNAW Humanities Cluster*
Context: XML in the Humanities

Text modelling outside OHCO model: TEI (Text Encoding Initiative)

Textual phenomena (variation, overlap, etc.)

Focus: computational philology and textual variation

"Collation": the comparison of two or more text versions
Problem Statement

How can we make optimal use of the potential of XML for computational philology?

1. Modelling textual variation in TEI
2. Hypergraphs and hypergraph merging
3. Visualise/export results
Open Variants

now that they were the people were between
had intercourse with them -- & th
Open Variants

<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <text>
    <p>
      <s>were open to them, had intercourse <seg type="alternative" xml:id="alt1">with</seg> <seg type="alternative" xml:id="alt2">between</seg> them</s>
    </p>
  </text>
</TEI>
Open Variants

TEI

<text>

<p>

<s>

<seg>

had

<seg>

intercourse

<seg>

with

<seg>

between

<seg>

them
Simultaneity

pulse: & then gently mak muffling & folding the house again.

Only suddenly & for no perceptible reason, a board creaks; &
Simultaneity

<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <text>
    <p>
      <s>a board creaks <add>on the <choice><sic>landg</sic><corr>landing</corr></choice></add>;</s>
    </p>
  </text>
</TEI>
Simultaneity
the way of the beetles in the kitchen, sliding from corn dresser to carpet. Moreover, how And Further, what is one night?
A short space, especially in the north; especially where the
They spread their garments, they rose up and the waves rose and through the house there was a sullen wave of doom which surled and crashed. The wind seemed ruining and washing away in water.

III

But what after all, is one night? A short space, especially when the darkness dims so soon, and so soon a bird

earth seemed ruining and washing away in water.

III

But what after all, is one night? A short space, especially when the darkness dims so soon, and so soon a bird

earth seemed ruining and washing away in water.
Witness 1

<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <div type="chapter" n="1">
    <p>
      <s>dresser to carpet.</s>
      <add place="margin"><head>2</head></add>
      <s>Further, what is one night?</s>
    </p>
  </div>
</TEI>

(simplified) TEI-XML transcription of manuscript p.157

Witness 2

<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <div type="chapter" n="2">
    <p>
      <s>washing away in water.</s>
    </p>
  </div>
  <div type="chapter" n="3">
    <head>III</head>
    <p>
      <s>But what after all, is one night?</s>
    </p>
  </div>
</TEI>

(simplified) TEI-XML transcription of typescript p.5
Modelling Texts in the Humanities

Beyond OHCO:

- Intradocumentary variation ➔ multiple paths through one text, encoded with TEI tags
- Structure ➔ comparing different documents with different structures results in conflicting hierarchies
Conditions

Different types of XML data:

- fully ordered
- unordered

Text-centric XML of literary texts (in short: "TEI text") = partially ordered data
Functional Requirements

Processing and analysing TEI text in a native way

Automated collation: finding the minimum set of changes needed to turn one document into the other.

Outset:

- documents are partially ordered
- treat structure and text equally
- compare more than 2 documents
Technical Requirements

Data model:

- schema-aware
- store multiple hierarchies

Processing:

- find the smallest amount of differences between two TEI texts
- distinguish between words, punctuation, markup
- respect non-linearity of TEI text
HyperCollate

HyperCollate = automated collation tool that uses a hypergraph model for variation

Hypergraph: nodes, hyperedges (one-to-one; one-to-many; many-to-many)

HyperCollate processes complex features of text, including:

- overlapping hierarchies
- intradocumentary variation
Hypergraph Model for Textual Variation

Diagram showing nodes connected by arrows, with words like "a", "board", "creaks", "on", "the", "landing", and ";" indicating different layers or aspects of text variation.
had intercourse between them with
Steps of HyperCollate

1. Convert XML tree of each witness into a variant hypergraph
2. Align two variant hypergraphs
3. Merge two hypergraphs in one collation hypergraph
4. Repeat in case of >2 witnesses
5. Visualise/export collation hypergraph
1. Transform TEI text witnesses into variant hypergraphs
2. Align the variant hypergraphs
3. Merge variant hypergraphs into one collation hypergraph
Conclusion

TEI text, being partially ordered data, is especially challenging for processing.

Requirements for analysis of textual variation:

- process multiple paths (i.e. be schema-aware)
- store multiple hierarchies

A hypergraph data model allows us to make optimal use of XML's potential for humanities research.
Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Astrid Kulsdom

R&D - Huygens Institute for the History of the Netherlands

KNAW Humanities Cluster

@ellibleeker
@ronald_dekker
@bram_buitendijk