

TEI for Interactive Concordances: The New Menota Search System

Øyvind Eide and Vemund Olstad, Unit for Digital Documentation, University of Oslo

Background

Menota is a network of leading Nordic archives, libraries and research departments working with medieval texts and manuscript facsimiles.¹ The aim of Menota is to preserve and publish medieval texts in digital form and to adapt and develop encoding standards necessary for this work. The Menota Archive has published 17 Medieval Nordic texts (approx. 923,000 words) online. The sources for the online versions are XML encoded according to the Menota handbook,² which is compatible with TEI P5.

The Menota search and display system was recently re-implemented at the Unit for Digital Documentation at the University of Oslo. In this paper, we will present the new implementation, focusing on the use of TEI as a retrieval format for corpus based KWIC concordances.

In the presentation, the system architecture and the data format used for the TEI concordances will be described in some detail. A number of open questions will be presented, hopefully starting a discussion on how to best use TEI for purposes such as the one presented in this paper.

Architecture

The Menota web system is served from a Cocoon platform, where the Menota/TEI encoded texts are converted on the fly to HTML using XSLT. The search form is part of the Cocoon system. Search expressions are sent from Cocoon to a HTTP/CGI frontend for a corpus system. The corpus is stored in Corpus Workbench and the search system is implemented in PERL using the Corpus Workbench API.³ The result of a corpus search is encoded as a TEI compliant XML document. An outline of the architecture can be found in figure 1.

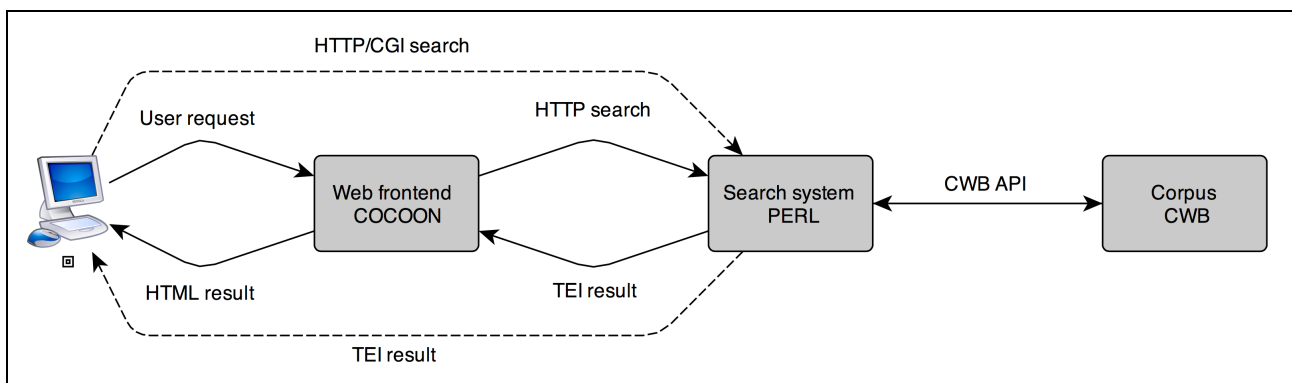


Figure 1. System architecture.

In the current implementation of the system, HTML versions of the search results are returned to the user through the COCOON frontend, as shown with the solid lines. In order to receive a TEI result, the user has to query the PERL/CGI system directly, as indicated with the dashed lines. We are planning on offering a more user friendly option for receiving TEI results in the future.

¹ <http://www.menota.org/> (checked 2013-08-27)

² http://www.menota.org/HB_index.xml (checked 2013-08-27)

³ <http://cwb.sourceforge.net> (checked 2013-08-27)

KWIC format

The TEI encoded KWIC concordance documents have short TEI headers giving basic metadata for the computer generated files, including some details about the search on which the TEI concordance was based. An example of a title statement from the TEI header can be found in figure 2.

```
<titleStmt>
  <title type="main">Search result from Menota corpus</title>
  <title type="sub">Searched for [word="talaðe" ] and had 30 hits.</title>
</titleStmt>
```

Figure 2. Title element from header of TEI concordance document.

```
<list>
  <item>
    <ref target="#HolmPerg-17-4to"/>
    <w n="w07888" lemma="braut" me:msaX="NC" me:msaI="" me:msaG="F"
me:msaN="S" me:msaC="A" me:msaS="I" me:msaR="" me:msaP="" me:msaT=""
me:msaM="" me:msaV="" me:msaF="" me:msaE="" me:msaY="" context="[TEI][text
xml:lang='onw']][body][div org='uniform' part='N' sample='complete'
type='chapter']][p]">brott</w>
    [several more w elements omitted]
    <w type="keyword" n="w07934" lemma="tala" me:msaX="VB" me:msaI="U"
me:msaG="" me:msaN="S" me:msaC="" me:msaS="" me:msaR="" me:msaP="3"
me:msaT="PT" me:msaM="IN" me:msaV="A" me:msaF="F" me:msaE=""
me:msaY="" context="[TEI][text xml:lang='onw']][body][div org='uniform' part='N'
sample='complete' type='chapter']][p]">talaðe</w>
    [several more w elements omitted]
  </item>
  [several more item elements omitted]
</list>
```

Figure 3: List element with one shortened item from the KWIC concordance.

The main part of a TEI encoded KWIC concordance document contains a number of concordance lines as items in a list, where each concordance line is presented in an **item** element. A shortened example can be found in figure 3. Each line on the KWIC concordance is represented as a list item containing a number of **w** elements: left context, the hit, and right context. To simplify the example, only one word from the left context and the hit word are included in the example.

Each word can be identified based on two pieces of information. First the **ref** element which is seen as the first child of the **item** element. The **target** of the **ref** points to the source Menota document. Second, the **n** attribute of each **w** element identifies the word within the source document. The **PCDATA** content of the **w** element contains the word as it is found in the running Menota text.

The rest of the attributes give further details for the word. The **lemma** attribute includes the lemma as coded in the source file. The **me:msa...** attributes offer detailed grammatical information. The **context** attribute gives the context vector, that is, the path from the XML root element. While this is not strictly speaking necessary, it is cached to simplify processing; it makes page formatting easier.

The hit word of each concordance line has the **type** attribute set to **keyword**.

Discussion

The use of TEI for the KWIC concordance data is convenient for our system. However, we would be equally well served by a non-standard internal format. An important purpose for the use of TEI is to open up for integration with external systems. If other corpus search interfaces were developed using TEI as the output format, combining results from several search would be convenient, opening up for tight integration between diverse corpus resources.

Another possibility for the future is to see results from corpus searches as first class documents. A TEI document which is the result of a corpus search could be used as the input for further processing, it may be published, and it can be preserved future use. By storing original ID values a link back to the corpus and, through that, a link back to the sources behind the corpus, can be preserved. It is important to be able to deliver the TEI concordance files to the users as unparsed data packages, in line with the way it is often done for simple formats such as comma separated lists. This saves us the time and processing power needed to parse large files on the online system.

We are currently considering how to best deliver the TEI corpus document. Should it be generated only when a user requests it at search time, or should it be offered as a linked option to all users at result time? We are also uncertain about how to offer this services to external systems. Demanding HTTP/CGI based search expression is not ideal. Should the search expression be formatted in TEI in line with the formatting of the resulting concordance? Or are there other formats more suitable for encoding search patterns?