Преглед НЦД 5 (2004), 5–18

M. J. Driscoll
(Arnamagnæan Institute, Copenhagen)

THE TEXT ENCODING INITIATIVE

Abstract. The Text Encoding Initiative's *Guidelines for electronic text encoding and interchange* provide an extremely rich set of mechanisms for the transcription of primary textual sources. The present article presents a brief overview of some of the more important areas dealt with in the *Guidelines*, with examples taken from the author's own work with Old Norse and Early-Modern Icelandic texts.

1. The TEI

The Text Encoding Initiative is an international standards project established in 1987 to develop, maintain and promulgate hardware- and software-independent methods for encoding humanities data in electronic form using SGML (Standard Generalized Markup Language, ISO standard 8879). It began as a research effort cooperatively organised by three scholarly societies (the Association for Computers and the Humanities, the Association for Computational Linguistics and the Association for Literary and Linguistic Computing), and funded by substantial research grants from, among others, the US National Endowment for the Humanities, the European Union, the Canadian Social Science Research Council and the Mellon Foundation. In December 2000 the TEI Consortium, an independent and self-sustained non-profit organisation, was set up to maintain and develop the TEI standard. The Consortium, which has executive offices in Bergen, Norway, and hosts at the University of Bergen, Brown University, Oxford University and the University of Virginia, is managed by a Board of Directors and its technical work overseen by an elected Council. There are currently 81 members of the TEI Consortium, among them universities, research libraries, academic and other publishers, both non-profit and commercial, as well as scholarly societies and research projects concerned with the design, production and delivery of structured electronic text. One does not, of course, need to be a member of the Consortium to use the TEI, and indeed there are thousands of users worldwide (for further information see the TEI website: http://www.tei-c.org/).

Immediately after its inception the TEI began the task of developing a draft set of *Guidelines for electronic text encoding and interchange*, with working groups comprising scholars in many fields from all over the world drafting recommendations on various aspects of the problem. These were integrated into a first public draft, *TEI P1* (*P* for ‘Proposal’), published in June 1990. A second draft, *P2*, followed in 1992, and in 1994 the first official version of the guidelines, *P3*, appeared. The current version, *P4*, which appeared in 2001, other than correcting various errors and oversights, differed little from *P3*, apart from being expressed in XML, rather than SGML. This was intentional, so as to ensure that any document conforming to the original *P3* SGML DTD would also conform to the new XML version of it. In *P5*, which is scheduled to appear in the beginning of 2005, this will not be the case. Several existing chapters in the *Guidelines* will be dropped and several new ones added and the whole will undergo — is undergoing — a thorough revision. The major structural change will be the move from DTDs to XML schema language (Relax NG).

One of the strengths of the TEI encoding scheme is that it is modular, comprising a number of different tag sets which can be used in a variety of combinations, according to the needs of the encoder and nature of the material being encoded. There are core and header tag sets, the former defining elements which may be said to be universal, i.e. not specific to particular types of texts, and the latter elements which allow for the provision of documentary and bibliographic information about the electronic text itself (metadata). In addition to these there are tag sets to be used with texts of a specific type, such as prose, verse, drama, transcriptions of spoken material, dictionaries and terminological data, as well as a number of additional tag sets which are designed for use with particular types of processing or research, for example for the transcription of primary sources and for textual criticism. In this way the encoder can tailor the system to his or her individual needs, selecting from the very large number of elements available those which are most relevant to the material to be encoded.

2. The overall structure of a TEI document

All TEI conformant documents must contain two elements, a header, tagged `<teiHeader>`, in which, as was mentioned, metadata, information about the electronic document, is provided, and the text itself, tagged `<text>`, within which there is a `<body>` element, which contains the bulk of the text, as well as optional `<front>` and `<back>` elements for, respectively, any prefatory matter, such as title pages, prefaces, dedications etc., and back matter, such as appendices, indices and so on. What elements comprise the `<body>` will be determined to a large extent by the nature of the material.

A lengthy prose work such as a novel or saga will usually be divided into chapters, each generally with its own title; these may be further divided into sections, possibly also with titles, and each consisting of a number of paragraphs. These can be marked up as shown here, using nested `<div>` elements for the various divisions, with a **type** attribute to indicate the nature, and an **n** attribute to show the number, of each; paragraphs, finally, are tagged using `<p>`:

```
<body>
  <head><!-- title of work--></head>
  <div type="chapter" n="1">
    <head><!-- title of chapter--></head>
    <div type="section" n="1">
      <p></p>
      <p></p>
      <!-- etc. -->
    </div>
    <div type="section" n="2">
      <p></p>
      <p></p>
      <!-- etc. -->
    </div>
    <!-- etc. -->
  </div>
  <!-- etc. -->
</body>
```

A lengthy work in verse will similarly normally consist of several cantos or fits, each containing a number of stanzas, each made up of a number of lines. These can be marked up as follows, using `<lg>` for ‘line-group’, i.e. a group of lines functioning as a formal unit, again with a **type** attribute to identify the type of unit, e.g. ‘stanza’ or ‘couplet’, and `<l>` for ‘lines’.

```
<body>
  <head><!-- title of work--></head>
  <div type="canto" n="1">
    <head><!-- title of canto--></head>
    <lg type="stanza" n="1">
      <l></l>
      <l></l>
      <!-- etc. -->
    </lg>
    <lg type="stanza" n="2">
      <l></l>
      <l></l>
      <!-- etc. -->
    </lg>
    <!-- etc. -->
  </div>
  <!-- etc. -->
</body>
```

Regardless of the type of text, there will be two separate structural hierarchies: that of the work, i.e. its division into chapters and sections or fits and stanzas, as we have seen, and that of the physical object carrying the text of the work, i.e. the arrangement or layout of the text on the page. In medieval manuscripts and many early printed books, there is rarely any connection between these two: a new chapter will not necessarily begin on a new page — except by accident — and poetry was generally written out like prose. Both of these hierchies need ideally to be encoded, even in the most basic of transcriptions. In order to represent the former, the TEI recommends the markup shown above, where `<div>` elements are used for the largest structural divisions and then either further `<div>` elements for smaller divisions in prose texts and `<p>` for paragraphs within these divisions or `<lg>` and `<l>` for verse texts. For the structure of the physical document, it is recommended that empty ‘milestone’ elements be used, `<pb/>`, `<cb/>` and `<lb/>`, for page-, column- and line-boundaries respectively, which, like all TEI elements, can also be numbered and provided with an **id** for identification and linking. (Note that `<lb/>` is an empty element used to indicate physical divisions in a printed book or manuscript, while `<l>` is used for lines of verse.) The way these elements are processed and/or displayed is determined by a style sheet. Line-breaks can be made to display as such, i.e. as an actual break or ‘hard return’, or marked for example with a vertical bar, or they can be ignored altogether. What is more, different style sheets can be applied to the same text to provide different views.

3. Levels of transcription

At the very outset the editor or transcriber of a text must usually decide how much of the information in the original document is to be included (or otherwise noted) in his or her transcription. W. W. Gregg’s distinction between, on the one hand, ‘substantives’, the ‘actual words’ of the text, and, on the other, ‘accidentals’, the ‘surface features’ of the text such as spelling, punctuation, word division etc., is well established (originally set out in ‘The Rationale of Copy-Text’, *Studies in Bibliography* III (1950-51), pp. 19-36). The editor’s choice is by no means simple, however, as there is a great range of such features which may or may not be included in any transcription, particularly when one is working with manuscript materials or early printed books. At one end of the spectrum, there are transcriptions which may be called strictly diplomatic, in which every feature of the original document which may reasonably be reproduced in print is retained. These features include not only spelling and punctuation, but also capitalisation, word division and variant letter forms. The layout of the page is also retained, in terms of line-division, large initials etc. Abbreviations will not be expanded, and, in strictly diplomatic transcriptions, apparent slips of the pen will remain uncorrected. Such editions are often so close to the originals as to be all but unreadable for those unfamiliar with early palaeographical or typographical conventions, or in any case no easier to read than the originals. At the opposite end we have fully modernised transcriptions, where the substantives are retained but everything else is brought up to date. In between these two extremes a number of levels may be distinguished — ‘semi-diplomatic’, ‘semi-normalised’ etc. — depending on how these original features are dealt with. Practice varies greatly, however, and in some editorial traditions it may be common to normalise or regularise some features while retaining others.

3.1 Variant letter forms. Variant letter forms — high and round s, for example — are often distinguished in transcriptions of manuscripts and early printed materials. Ligatures in the text — those which are the result of scribal economy rather than representing separate phonemes — may also be retained or their presence otherwise noted in a strictly diplomatic transcription. Texts which are semi-diplomatic or semi-normalised will in general distinguish only between variant letter forms which are felt to have a basis in phonology distinctions, which the two forms of /s/, for example, normally do not. For statistical purposes, however, it may be desirable to register such palaeographical or typographical distinctions, even if one does not intend to display them.

Variant letter forms, and indeed any ‘exotic’ (read non-English) characters, can be represented using entity references, which may be given as a numeric entity reference, using either decimal or hexadecimal notation, in the Universal Character Set developed by the Unicode Consortium, or using a standardised name, which is then defined (in the DOCTYPE declaration subset) with reference to the Unicode standard. If one wished to distinguish between different allographs of a single letter or other palaeographical features for purposes of statistical analysis, one could define one's own entity references for this purpose, &a1;, &a2; and so on.

A TEI workgroup is currently reviewing the Writing System Declaration and related constructs and recommendations within the TEI Guidelines (Chapters 4 and 25) and will produce recommendations for an entirely new system for dealing with languages and writing systems to be incorporated into P5. One significant change will be that it will be possible to declare languages and writings systems independently, which has not previously been the case, there having been an erroneous assumption that a single language will only ever be written in a single script; one need look no further than Serbian to see an example of one that regularly makes use of two.

3.2. Abbreviations. The use of abbreviations is a characteristic feature of most medieval western manuscript traditions. In all but the strictest diplomatic transcriptions it is common practice to expand abbreviations as an aid to the reader. When a word or phrase is abbreviated a number of letters is suppressed and the expansion of the abbreviation thus involves supplying these letters. The letters so supplied are frequently marked in some way, printed in italic or given in brackets, for example, but even in transcriptions which are otherwise fairly diplomatic abbreviations may be expanded silently.

Abbreviations and their expansions can be marked up using either the `<abbr>` or the `<expan>` element. A transcription striving to be as ‘neutral’ as possible might give the unexpanded abbreviation, tagged `<abbr>`, as in the following example, where the Icelandic word *hann* (‘he’) is indicated by the letter h with a bar or stroke through the ascender (here indicated by the `&bar;` entity):

```
h<abbr>&bar;</abbr>
```

One could also use the expanded form, tagged `<expan>`:

```
h<expan>ann</expan>
```

In theory, it doesn’t matter which is used, since `<abbr>` and `<expan>` are so-called ‘Janus’ tags, meaning that each can also serve as an attribute of the other, as shown here:

```
h<expan abbr="&bar;">ann</expan>
```

Alternatively, one can provide both, and allow the style sheet to determine which is to be displayed:

```
h<abbr>&bar;</abbr><expan>ann</expan>
```

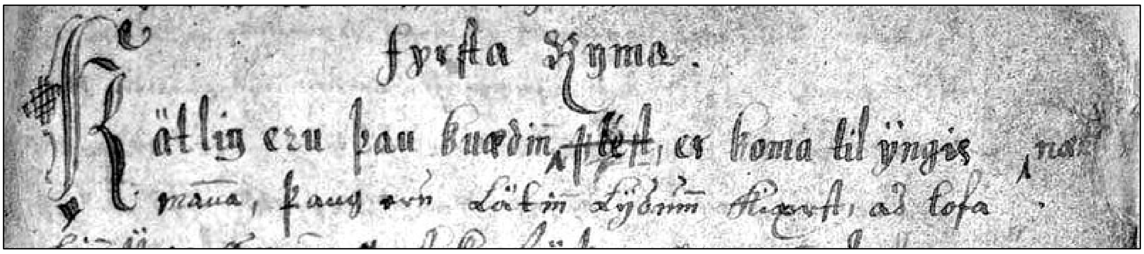
This last method is in keeping with recommendations to be made in P5, where the use of attributes the content of which is CDATA is to be deprecated, as will be discussed below.

It is customary to divide abbreviations into four types, seen from the point of view of the means through which the abbreviation is achieved:

1. Suspension, where only the first letter or letters of a word are written, generally followed (and frequently also preceded) by a point or with a superscript stroke.
2. Contraction, where the first and last letters are written, normally also with a superscript stroke, or, less commonly, a point or points.
3. Special signs or brevigraphs, such as the Tironian *nota*, used for ‘et’ in Latin and in the vernacular languages in the same sense.
4. Superscript letters, a superscript vowel normally representing that vowel preceded by *r* or *v*, while a superscript consonant represents that consonant preceded by *a*.

One fundamental question, to which there appears to be no simple answer, is what, exactly, is ‘the abbreviation’? Is it the mark, sign or letter that indicates that something has been suppressed, or is it the entire word? And similarly, is ‘the expansion’ only the letters which have been suppressed and have therefore been supplied by the transcriber, or is it, again, the whole word? Given that there is a clear distinction between abbreviations with a lexical reference, i.e. those which indicate that something has been omitted, without suggesting what that something may be (suspensions, contractions and some of the brevigraphs), and those with a graphemic reference, and those which always refer to a particular combination of graphemes, regardless of the lexical item in which they occur (superscript letters and signs and the remainder of the brevigraphs), the most reasonable answer would appear to be that it is both. It strikes one as counter-intuitive to treat abbreviations with a lexical reference on anything other than the whole-word level, and equally so those with a graphemic reference in any other way, especially in cases where there are more than one abbreviation within a single word; treating such abbreviations on a whole-word basis would blur the connection between the abbreviation sign and its expansion, making certain types of statistical analysis difficult or impossible. One solution would be to distinguish between the different types of abbreviations using the **type** attribute.

3.3. Additions, deletions and substitutions. Alterations made to the text, whether by the scribe or in some later hand, can be encoded using <add> and <del>; further information may be given as attribute values. The following is an example of a substitution, taken from the first line of an Icelandic metrical romance, preserved in a manuscript from the very end of the 17th century (1695). The scribe has crossed out the word ‘flest’ and substituted the word ‘nærst’, which provides a better rhyme (with ‘kærst’) without significantly altering the sense. The markup shows how all this information can be registered:



```
K&auml;tlig e&rrot;u &thorn;au ku&aelig;din<expan>n</expan>
<del type="subst" rend="overstrike">&fins;le&slong;t</del>
<add place="margin"
hand="scribe"></em>n&aelig;&rrot;&slong;t</add> er koma til
&ijlig;ngis
```

Which can be made to display as follows:

Kätlig e□u þau kuædin*n* ~~□leſt~~ \næ□ſt/ er koma til ijngis

3.4. Corrections and emendations. Obvious scribal errors and omissions — the sorts of things, it is argued, which the scribe himself would have corrected had they been brought to his attention — will normally be corrected by the transcriber. An editor may also want to emend the text on the basis of readings from other witnesses, common sense or artistic inspiration, ‘correcting’ things in the text which the scribe would presumably not have regarded as in need of correction. These corrections and emendations can be marked in a variety of ways in printed editions; one common way is to place letters or words assumed to have been inadvertently omitted by the scribe in angle brackets, while obvious misspellings are corrected and marked with an asterisk, the original form being given in a note. More extensive emendations are normally treated in the notes.

Alterations made to the text by the scribe or in a later hand, as we saw, can be encoded using <add>, for additions, and <del>, for deletions, while editorial emendations can be encoded with the <sic> and <corr> tags. The former pair, it has sometime been argued, could also be used for editorial emendations, on the grounds that there is, in essence, no great gulf between changes made to the text by a scribe or later reader and those made by the transcriber or editor of a scholarly edition, but it must also be said that there is no more fundamental distinction in textual scholarship than between what is physically present in the manuscript and what is not. The act of recording what actually is in the manuscript must therefore remain entirely separate from postulating what ought perhaps to have been there but for one reason or another isn’t. The simplest way to maintain this distinction is to employ separate sets of elements for the two.

The elements <corr> and <sic> function as mirror images of each other, in the same way as <abbr> and <expan>, and the choice as to which to use is made on the same grounds. In a strict diplomatic transcription, one may wish only to indicate an incorrect or suspicious reading in the manuscript without attempting to correct it, or, in a normalised text, to emend an obvious error without indicating what the original reading was. The two can also be combined, with the one then acting as an attribute of the other.

```
<corr sic="giorit">giorir</corr>
```

Or, as with <abbr> and <expan>, one may choose to use both side by side:

```
<sic>giorit</sic><corr>giorir</corr>
```

The style sheet can then determine which is to be displayed.

3.5. Supplied text. Where a word has been supplied by the editor, the <supplied> tag can be used. It is customary in textual scholarship to distinguish between text now illegible or lost through damage but assumed originally to have been in the manuscript, and text assumed to have been inadvertently omitted by the scribe. This distinction is indicated in the markup through the use of the reason attribute, and can be made to display in different ways (e.g. in angle brackets in the case of the former and square brackets in the case of latter):

```
gieck sijdan <supplied reason="omitted">j burt</supplied>
```

gieck sijdan <j burt>

```
lid<supplied reason="illegible">z</supplied>
```

lid[z]

With <supplied> the attribute **resp** can be used to indicate the scholar or previous editor responsible for the conjectural emendation. Where the reading of another witness supports the reconstruction it is also possible to use the **source** attribute:

```
ath &thorn;eir <supplied reason="omitted" source="AM 152
fol., 76ra" resp="NN">mundu</supplied> sundr ganga
```

The <supplied> element should only be used when the missing text can be reconstructed with some degree of certainty. When such is not the case <gap> should be used instead, with both a **reason** and an **extent** attribute. The extent should be given as the number of characters presumed missing, which can then be made to display as a series of small noughts, as is customary in a printed edition.

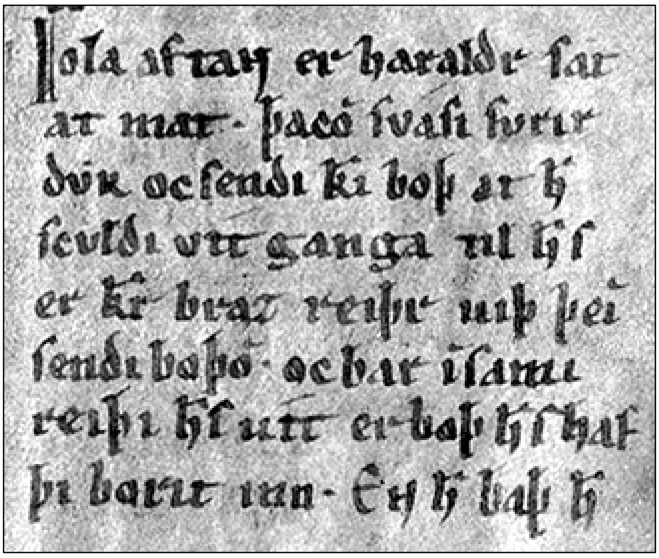
3.6. Normalisation. Finally, there is the question of normalisation/regularisation. One can use the `<reg>` element to give regularised forms of variant or archaic spellings, or the `<orig>` element to indicate that a spelling is archaic or non-standard. As with several other of the elements already mentioned, the two mirror each other, and both allow for both the regularised and unregularised forms to be given, the one as the content of the element, the other as an attribute value, or, as has been said, as two separate elements, i.e. either:

```
<orig reg="l&iacute;ka&eth;i">lijkadi</orig>
```

or:

```
<reg>l&iacute;ka&eth;i</reg><orig>lijkadi</orig>
```

4. Multi-level markup



Shown here are eight lines, the beginning of chapter 3, from f. 1v, column b, of the manuscript AM 235 II 4to, an Icelandic vellum of the early 13th century containing a history of the kings of Norway. The text contains numerous abbreviations, a horizontal bar or stroke being used both to represent a suppressed nasal or as a general mark of abbreviation, as well as a number of unusual letter forms, the small capitals n and r (representing geminates), dotless i, insular f, long s and uncial e. There is also a single error, where the scribe has written the non-sensical ‘er’ (‘is’ or the relative ‘which’) instead of ‘en’ (‘but’). All these features can be registered in the markup, as shown here:

```
<div type="chapter" n="3"><p><orig
reg="J&oacute;laaptan"><hi rend="3">I</hi>ola
afta&nscap;</orig> <orig reg="er">er</orig> <name
type="person"><orig reg="Haraldr">haraldr</orig></name>
<orig reg="sat">&stall;at</orig> <lb n="19"/> <orig
reg="at">at</orig> <orig reg="mat">mat</orig>. <orig
reg="&thorn;&aacute;">&thorn;a</orig><orig
reg="kom">co<abbr>&bar;</abbr><expan>m</expan></orig> <name
type="person"><orig
reg="Sv&aacute;si">&stall;v&aacute;&stall;&idl;</orig></nam
e> <orig reg="fyrir">fvr&idl;r</orig> <lb n="20"/> <orig
reg="dyrr">d&vdot;&rscap;</orig> <orig reg="ok">oc</orig>
<orig reg="sendi">&stall;end&idl;</orig> <orig
reg="konungi">k<abbr>&bar;</abbr><expan>onung</expan>&idl;<
/orig> <orig reg="bo&edh;">bo&thorn;</orig> <orig
```

```
reg="at">at</orig> <orig
reg="hann">h<abbr>&bar;</abbr><expan>ann</expan></orig> <lb
n="21"/> <orig reg="skyldi">&stall;cvld&idl;</orig> <orig
reg="&uacute;t">&vacute;tt</orig> <orig
reg="ganga">ganga</orig> <orig reg="til">t&idl;l</orig>
<orig
reg="hans">h<abbr>&bar;</abbr><expan>an</expan>&stall;</ori
g>. <lb n="22"/> <corr sic="er"><orig
reg="en">en</orig></corr> <orig
reg="konungr">k<abbr>&bar;</abbr><expan>onung</expan>r</ori
g> <orig reg="br&aacute;sk">braz</orig> <orig
reg="rei&edh;r">re&idl;&thorn;r</orig> <orig
reg="vi&edh;">u&idl;&thorn;</orig> <orig
reg="&thorn;eim">&thorn;e&idl;<abbr>&bar;</abbr><expan>m</e
xpan></orig> <lb n="23"/> <orig
reg="sendibo&edh;um">&stall;end&idl;
bo&thorn;o<abbr>&bar;</abbr><expan>m</expan></orig>. <orig
reg="ok">oc</orig> <orig reg="bar">bar</orig> <orig
reg="inn">&idl;<abbr>&bar;</abbr><expan>nn</expan></orig>
<orig reg="sami">&stall;am&idl;</orig> <lb n="24"/> <orig
reg="rei&edh;i">re&idl;&thorn;&idl;</orig> <orig
reg="hans">h<abbr>&bar;</abbr><expan>an</expan>&stall;</ori
g> <orig reg="&uacute;t">&uacute;tt</orig> <orig
reg="er">er</orig> <orig reg="bo&edh;">bo&thorn;</orig>
<orig
reg="hans">h<abbr>&bar;</abbr><expan>an</expan>&stall;</ori
g> <orig reg="haf&edh;i">ha&fins;<lb
n="25"/>&thorn;&idl;</orig> <orig
reg="borit">bor&idl;t</orig> <orig
reg="inn">&idl;nn</orig>. <orig
reg="En">&eunc;&nscap;</orig> <orig
reg="hann">h<abbr>&bar;</abbr><expan>ann</expan></orig>
<orig reg="ba&edh;">ba&thorn;</orig> <orig
reg="hann">h<abbr>&bar;</abbr><expan>ann</expan></orig><!--
rest of chapter --></p></div>
```

Using different style sheets, one can display the text in a variety of ways, from a strictly diplomatic text, retaining the line-breaks, variant letter forms, unexpanded abbreviations, and so on of the original:

lola aftaɴ er haraldr ſat
at mat. þaco̅ ſváſı fvrır
dv̇ʀ oc ſendı kı̅ boþ at h̅
ſcvldı vtt ganga tıl hſ̅.
er k̅r braz reıþr uıþ þeı̅
ſendı boþo̅. oc bar ıſ̅amı

reıþı hſ̄ útt er boþ hſ̄ ha□
þı borıt ınn. Ɛɴ h̄ baþ h̄

to a semi-diplomatic text, where the abbreviations have been expanded and the expansions displayed in italic and the obvious error has been corrected:

Iola aftaɴ er haraldr ſat
at mat. þaco*m* ſváſı fvrır
dv̇ʀ oc ſendı k*onung*ı boþ at h*ann*
ſcvldı vtt ganga tıl h*an*ſ.
*en k*onung*r braz reıþr uıþ þeı*m*
ſendı boþo*m*. oc bar ı*nn*ſamı
reıþı h*an*ſ útt er boþ h*an*ſ ha□
þı borıt ınn. Ɛɴ h*ann* baþ h*ann*

to a semi-normalised text, where abbreviations are expanded silently, the line breaks are not retained, most variant letter forms have been replaced and names, which were tagged using the <name> element, have been capitalised:

Iola aftaɴ er Haraldr sat at mat. þacom Svási fvrir dv̇ʀ oc sendi konungi boþ at hann scvldi vtt ganga til hans. *en konungr braz reiþr uiþ þeim sendi boþom. oc bar innsami reiþi hans útt er boþ hans hafþi borit inn. Ɛɴ hann baþ hann

and, finally, to a fully normalised text, where the spelling, punctuation, capitalisation, word division and so on have all been regularised, by taking the values of the **reg** attribute:

Jólaaptan, er Haraldr sat at mat, þá kom Svási fyrir dyrr ok sendi konungi boð at hann skyldi út ganga til hans, en konungr brásk reiðr við þeim sendiboðum, ok bar inn sami reiði hans út er boð hans hafði borit inn. En hann bað hann

But this is only the beginning. The TEI also provides mechanisms for associating any kind of semantic or syntactic analysis and interpretation which an encoder might wish to attach to all or part of a text, including familiar linguistic categorisations such as ‘clause’, ‘morpheme’, ‘part-of-speech’ etc., as well as characterisations of narrative structure, such as ‘theme’.

5. Changes in P5

As was mentioned, there are some major changes to be expected in the next version of the TEI Guidelines, P5, some, but by no means all, a result of the move from DTDs to schemas.

5.1. The ODD format. One of the more interesting of these is the ODD format, originally developed for internal use as a means of generating the TEI Guidelines (ODD stands for 'One Document Does it all') but now revamped as an XML implementation which can be used for the documentation of XML elements and element classes and also used in the automatic generation of schemas or DTDs conforming to that documentation; Relax NG XML and DTD are generatable direct from the P5 source, while W3C Schema and Relax NG compact syntax are generated by post-processing using Roma, the new replacement for the TEI 'Pizza Chef', which generated SGML (and later XML) DTDs. It is primarily intended for use by those wishing to customise or modify the TEI Guidelines in a conformant manner, but may also be used for the documentation of any comparable encoding scheme.

One result of the move to schema language and the new ODD system is that it will be possible to produce localised versions of the TEI in different languages. This will involve the automatic translation of element and attribute names (and values) from one language to another in any document instance, in the DTD or schema and in the documentation itself.

5.2. Attribute values. It has already been mentioned that the intention is to deprecate — or indeed ban altogether — attribute values with a content of CDATA, either replacing them with tokens, i.e. a fixed number of possibilities, or by employing separate elements. One immediate advantage of this would be in cases where markup, which is not allowed in atribute values, is required in the alternative views.

We have already seen examples of so-called 'Janus' elements (Janus, the Roman god of doors, was often depicted as having two faces, and could look both backwards and forwards) such as `<abbr>` and `<expan>`, `<sic>` and `<corr>` and `<orig>` and `<reg>`, where the one can also function as an attribute of the other:

```
h<expan abbr="&bar;">ann</expan>
```

Instead of this, since the content of the **abbr** attribute is theoretically infinite, and may need to contain markup, one would be obliged to used separate elements:

```
h<abbr>&bar;</abbr><expan>ann</expan>
```

Doing so, however, breaks a long-standing — but unwritten — rule of text-encoding, which views markup as something one adds to the text, such that if the markup is removed one should be left with what one started with, something that makes sense. Removing the markup from the example shown here would leave one with two views of the same word side by side. Partially for this reason, but principally to make explicit the notion that these are alternatives, a new grouping element, tentatively called `<choice>`, has been proposed:

```
h<choice><abbr>&bar;</abbr><expan>ann</expan></choice>
```

One could also wrap the entire word within `<choice>` tags, which would allow for further possibilites. Where choices are available within choices, the `<choice>` element could nest:

```
<choice>
  <orig>
    <choice>
      <abbr>h&bar;</abbr>
      <expan>hann</expan>
    </choice>
  </orig>
  <reg>hann</reg>
</choice>
```

5.3. Manuscript description. P5 will contain a major new chapter on manuscript description, based chiefly on the work of the TEI Medieval Manuscript Description Work Group, headed by Consuelo Dutschke and Ambrogio Piazzoni, and MASTER (Manuscript Access through Standards for Electronic Records), an EU-funded project headed by Peter Robinson, but with significant input also from the Repertorium of Old Bulgarian Literature and Letters project, based in Sofia and Pittsburg (http://clover.slavic.pitt.edu/~repertorium/).

The chapter defines a set of special purpose elements which can be used to provide detailed descriptive information about any kind of ancient inscribed artifact. Although originally developed with the needs of manuscript scholars working in the European tradition, the scheme is general enough that it can also be extended to other kinds of material and other traditions.

The <msDescription> element is the framing element into which the manuscript description is put. This will normally appear within the <sourceDescription> element of the header of a TEI conformant document, where the document being encoded is a digital representation of some manuscript original, whether as a transcription, as a collection of digital images or as a combination of the two. However, in cases where the document being encoded is essentially a collection of manuscript descriptions, the <msDescription> element may be used in the same way as the bibliographic elements, i.e. within the listBibl element.

Within the <msDescription> comes a required <msIdentifier> element, which groups information identifying the manuscrip, its location, holding institution and shelfmark, followed by an optional <head>, which can be used to provide in a brief, unstructured way information on contents, date and place of origin, and the language or languages of the manuscript. These are then followed either by one or more paragraphs, for those engaged for example in retrospective conversion of existing catalogues into machine readable form, who may not want (or have the option of) more structured data, or one or more of the following specialised elements:

- <msContents>, which contains an itemised list of the intellectual content of the manuscript, with transcriptions of rubrics, incipits, explicits etc, as well as primary bibliographic references
- <physDesc>, which groups information concerning all physical aspects of the manuscript, its material, size, format, script, decoration, binding, marginalia etc.
- <history>, which provides information on the history of the manuscript, its origin, provenance and acquisition by its holding institution
- <additional>, groups other information about the manuscript, in particular, administrative information relating to its availability, custodial history, surrogates etc.
- <msPart>, which contains in essence a nested <msDescription>, in cases of composite manuscripts now regarded as constituting a single unit but made up of two or more parts which were originally physically distinct.

Within each of these elements a number of sub-elements is available; <msContents>, for example, will normally consist of one or more <msItem> elements, each in turn containing specific elements for <rubric>, <incipit>,

<explicit> and <colophon>, as well as the standard TEI elements <author>, <title> and <bibl> for bibliographic references. As with <msDescription> itself, however, the contents of these first-level and second-level elements need not be this structured, since there is also the option of using paragraphs.

5.4. Names. One of the innovations of the MASTER project was a way of centralising and cross-referencing information about persons by defining <person> and <listPerson> elements, based on the existing <partic> and <particDesc> elements available in the Corpora module for participants in a transcribed text, such as a conversation, which allowed for a number of so-called ‘demographic’ subelements, such as <birth>, <occupation> and <residence>, to which several more were added, e.g. <death>, along with attributes indicating such features as gender and role. This system will become not just part of the new Manuscript Description module, but will be made available throughout the TEI. According to the system, all names are tagged as such using the <name> element, with a type attribute to indicate whether they are the names of persons or places, and key attribute the value of which refers to more detailed information provided in <listPerson> and <listPlace> elements within the <profileDesc> element in the TEI header.

The advantage of treating personal names in this way should be readily apparent. Each person is uniquely identified with an ID, to which all individual instances of that person's name then refer, whatever form those instances take. This solves the problem not only of variant spellings but also in cases where, for example, a medieval author is known by a standard Latin, Greek or Old Church Slavonic name and any number of vernacular forms, many or indeed all of which may have claims to ‘authenticity’. In order to ensure uniformity, the method generally employed in the library world has been to accept the form found in some authority file, for example that of the American Library of Congress, as the ‘base’ or ‘neutral’ form. Feelings can run high on this matter, however, and people are frequently reluctant to accept as ‘neutral’ an overtly ‘foreign’ form of the name of some local saint or hero. Within the <person> tag any number of variant forms of a name can be given, with no prioritisation, and hence, less likelihood of offense.

6. Conclusion

The TEI has established itself as the *de facto* standard for the electronic encoding of literary and linguistic texts. The changes to be introduced in *P5* will only increase its robustness and flexibility and should thus ensure that it remains so for the forseeable future.

References

M.J. Driscoll, 'Encoding Old Norse/Icelandic primary sources using TEI-conformant SGML', *Literary and linguistic computing*, XV, 1 (2000), pp. 81-91.

TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML-compatible edition, ed. C. M. Sperberg-McQueen and Lou Burnard ([Oxford], 2001); also available on-line at: http://www.tei-c.org/P4X/.

Text encoding initiative: Background and context, ed. Nancy Ide and Jean Véronis (Dordrecht, 1995).

The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources, Version 1.1, ed. M. J. Driscoll, Odd Einar Haugen, Karl. G. Johansson and Rune Kyrkjebø (Bergen, 2004); available on-line at: www.hit.uib.no/menota/guidelines.

The Unicode Standard, Version 4.0, ed. Joan Aliprand et al. (Boston, MA, 2003); available on-line at: http://www.unicode.org/versions/Unicode4.0.0/.

mailto:mjd@hum.ku.dk