Document markup - Why? How?

Ore, Espen S.

Document Markup – Why? How?

Espen S. Ore [*]

Abstract: *»Dokumentenauszeichnung – Warum? Wie?«*. In this paper I argue that markup and writing belong to related systems for storing information and/or speech and that there is no clear border between the two. In addition I argue that marking up text done as more or less separate from ordinary writing has been used in Western scholarly work at least since the times of the library in the Museum in Alexandria and up until today. Markup means that some part of a document is identified and some statement is made about the linguistic and/or textual status and interpretative frame of that part or it is extracted for some scholarly purpose. The ways and means by which this is done may vary. It will depend both on the aim: why exactly do we wish to identify this part of the text? And on the technology available: papyrus scrolls and reed pens make for different markup than what is done with computer stored texts. In this paper selected uses for digital text and markup are discussed with examples mainly taken from the electronic edition of Henrik Ibsen's Writings.

Keywords: markup, text encoding, TEI, inline encoding, standoff encoding.

Introduction

In this paper I will look at how and why markup is used and has been used in various periods and across document types although the examples used will be text documents. I start from the point of view that markup has been with us more or less since the invention of writing and try to demonstrate that the use of markup and the explicit way markup is done depends on technology, the scholarly (if any) context, tradition, document type and various other variables. Writing seems to have been invented both as a way of recording abstract ideas such as numbers and as a way of recording speech. An example of the second use is found in, for instance, Runic texts: the texts are often written without word divisions, and when the final sound of one word is similar to the one at the beginning of the next word, one rune is often used for both[1] (see fig. 1). Both uses are probably found in the Linear B tablets – one of the more famous ones is Ta641 found in Blegen's excavations at Pylos where a vessel is described as "handle-less" (strictly "ear-less") in syllabic script while there is also a symbol showing a handle-less vessel.[2] In this tablet the logographs combined

[*] Address all communications to: Espen S. Ore, Department of Linguistics and Scandinavian Studies, University of Oslo, PO Box 1102 Blindern, 0317 Oslo, Norway; e-mail: espen.ore@gmail.com.

[1] See Barnes 2007, p 122.

[2] See Ventris & Chadwick 1973, p 336-7 and plate IIIb.

with the numbers make it possible for an illiterate person to read the accounts. The logographs represent or are summaries of the text written in the Linear B syllabary (see fig. 2). As with early Greek script, Runic script was, as mentioned, also often written continuously without word divisions (*scriptio continua*). The introduction of word separators did not fundamentally change the text as a transcription of a stream of sounds but it definitely gave information that made reading easier. Seen from this point of view, I find it difficult not to consider word separators as belonging to the class of markup codes, although they may belong to other classes as well.[3] With this in mind, the following will mainly discuss explicit markup separate from the running text, but it may at times look into the overlapping areas between these two classes.

Figure 1: The "Theodoric verses" on the Rök Rune Stone

Top left in the shown lines has "*raiþ þiaurikR"*, in normalized Eastern Norse "*Réð þjóðrikr*", written as "*raiþiaurikR*" with a Single "þ" used for both words.[4]

Figure 2: The "Tripod tablet"

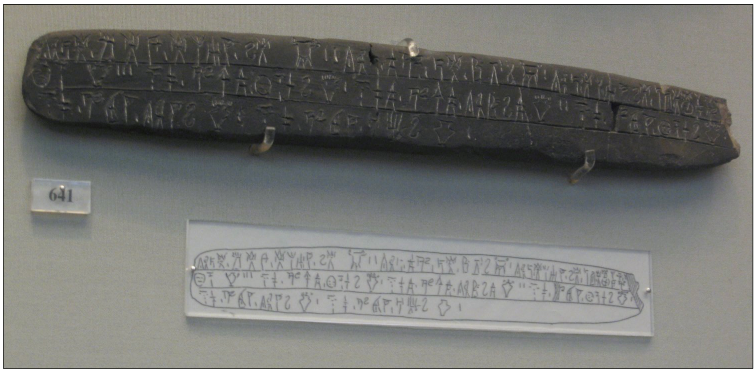


Photo by John Sie Yuen Lee, used with permission.[5]

[3] For a discussion and suggested definition of markup different from the one used here, see Schmidt 2010, p 338.

[4] The drawing is taken from *Östergötlands runinskrifter (SRI Band 2, 1911-1918),* digital version at <http://www.raa.se/cms/extern/kulturarv/arkeologi_och_fornlamningar/runstenar/digitala_sveriges_runinskrifter.html>.

Markup from Homer ...

Hellenistic scholarly editions basically came in two parts: the text itself and a commentary (*hypomnemata*). In the text there would be symbols identifying parts of the text that were discussed in the commentary. One early use of this markup system was the use of symbols (usually the *obelus*) to identify parts of a text that were not accepted as belonging to the original version of a text. The idea that there might be errors in the transmitted text goes back until it is lost in time. Plutarch in the first (or second) century C.E. may or may not include an anachronism when he writes that one of Alcibiades' teachers (in the fifth century B. C.E.) was capable of doing text critical work on Homer:

> ἑτέρου δὲ φήσαντος ἔχειν Ὅμηρον ὑφ' αὐτοῦ διωρθωμένον, 'εἶτ',' ἔφη, 'γράμματαδιδάσκεις, Ὅμηρον ἐπανορθοῦν ἱκανὸς ὤν; οὐχὶ τοὺς νέους παιδεύεις;'[6]
>
> (To another [teacher] who claimed that he himself was correcting Homer, [Alcibiades] said: how come that you who are able to revise Homer, are teaching children to read and not educating youths?)

Whether we believe in a Peisistrean edition[7] or not, editorial work on Homer goes back at least to the library at the Museum in Alexandria. Plutarch also gives examples of this work, for instance:

> ὁ μὲν οὖν Ἀρίσταρχος ἐξεῖλε ταῦτα τὰ ἔπη φοβηθείς:[8]
>
> (Aristarchus removed these lines because of fear.)

The four lines quoted by Plutarch are not in the surviving medieval manuscripts. On the other hand, some modern editors have in fact included them as lines 458-61 in Il. IX.[9]

But how did the scholars in Alexandria remove the verses that they did not accept as belonging to Homer's text? An unacceptable verse (line) in the Iliad or the Odyssey would probably have been marked with an obelus by Zenodotus or Aristarchus,[10] two of the first known commentators/editors of the Homeric epics. The obelus is usually understood as a mark indicating that the marked verse is spurious and probably should be omitted. Here we can see an example of marking a part of a text (a verse in the case of Homer) and of giving the

---

5 Pylos tablet Ta641, National Museum, Athens, Greece.
6 Plutarch, *Alcibiades,* 7.1, Loeb Edition web published by the Persus project.
7 See for instance Davison 1955.
8 Plut. Adolescens 26f. Teubner Edition web published by the Perseus project.
9 See the Loeb-edition, Plutarch Moralia 1927. The lines are only included in the apparatus in Monro and Allen's 1902-edition (as seen in the 1969 edition).
10 Zenodotus fl. 280 BCE, Aristarchus ca 220-140 BCE.

marked text an attribute value: “spurious text”. Apart from the obelus, the Alexandrian scholars used other symbols such as the diple to mark lines that were commented in one way or another. As mentioned earlier, Hellenistic scholarly editions basically had two parts: the text itself with critical marks in the margin and the commentary (hypomnemata) where the marked passages were commented on. If we jump to modern technology we could say that the text had both in-line markup (the critical symbols) and stand-off encoding: the comments in the hypomnemata. The comments in the hypomnemata were in medieval times incorporated in the text manuscripts, the so-called *scholia*. In modern times scholia are edited and published separate from the main texts – in a way we are back where we began.

The Hellenistic scholars and grammarians did not only give us the critical symbols used to mark up the text and the comments which for a while were added to the text pages. They also added markup to the text itself: the polytonic Greek accent system is inherited from them. Although we would now usually consider these accents ordinary parts of the written texts, they were introduced to convey information that would not be obviously available to the reader of scriptio continua. A remnant of this encoding is still found in Modern Greek monotonic script.

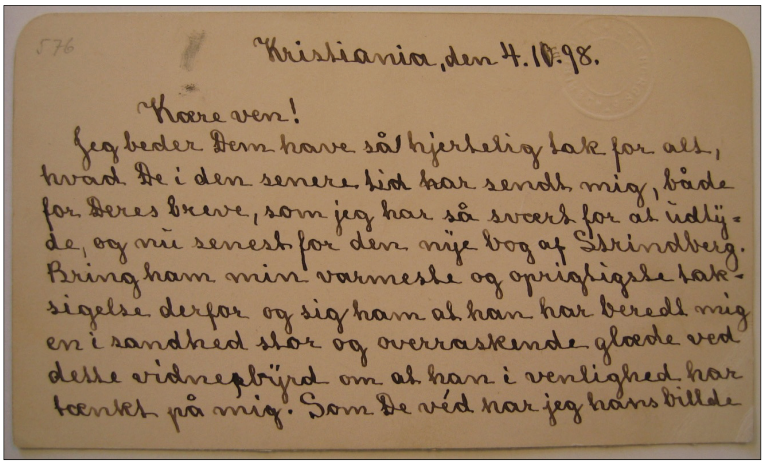
With the invention of movable type and the publishing of printed editions, a new level of markup was developed: the critical apparatus. The apparatus as we know it is typically placed at the bottom of the page or after the edited text: there are references to a line number and a variant reading along with references to witnesses containing it is listed. Thus, this can be called stand-off encoding since the markup or encoding is pointing to a location in the text and is not marked with symbols in the published text itself. But the modern printed editions have added their own in-line markup. The Hellenistic obelus in its incarnation as the dagger symbol is used to mark parts of a text that seem to be corrupted but where the editor has selected not to give a meaningful reading.[11] And especially in diplomatic editions such as editions of papyrus texts we find the so-called *Leiden Conventions* in use. This convention was agreed upon in 1931[12] and uses symbols such as [ ] for missing text and <abc> for missing text added by the editor. This is clearly an in-line encoding or markup system.

... to Ibsen

In the following, the work done in the project Henrik Ibsen’s Writings (HIW) will be used for examples and as a basis for discussion.

[11] See for instance Euripides 1966 1087-8.
[12] <http://en.wikipedia.org/wiki/Leiden_Conventions>, (Accessed April 5, 2012).

Figure 3: First Page of a Letter from Henrik Ibsen to Gustaf af Geijerstam Nov. (!) 4, 1898[13]


576

Kristiania, den 4.11.98.

Kære ven!

Jeg beder Dem have så hjertelig tak for alt, hvad De i den senere tid har sendt mig, både for Deres breve, som jeg har så svært for at udtyde, og nu senest for den nye bog af Strindberg. Bring ham min varmeste og oprigtigste taksigelse derfor og sig ham at han har beredt mig en i sandhed stor og overraskende glæde ved dette vidnesbyrd om at han i venlighed har tænkt på mig. Som De véd har jeg hans billede

In fig. 3 the first page of a letter from Henrik Ibsen to the Swedish author and literary critic Gustaf af Geijerstam seems to contain a very straight forward text where there may not seem to be much need for markup. Still we may see some links attached in the web edition of Henrik Ibsen's Writings:[14]

Figure 4: A Screen Dump of the Web Publication of the Text from Fig. 3

Kristiania, den 4.11.98.

Kære ven!

Jeg beder Dem have så hjertelig tak for alt, hvad De i den senere tid har sendt mig, både for Deres breve, som jeg har så svært for at udtyde, og nu senest for den nye bog af Strindberg. Bring ham min varmeste og oprigtigste taksigelse derfor og sig ham at han har beredt mig en i sandhed stor og overraskende glæde ved dette vidnesbyrd om at han i venlighed har tænkt på mig. Som De véd har jeg hans billede

This display is generated from an XML-encoded text taken from the published printed collection of letters. In the XML-file[15] the part from the letter displayed in figs. 3 and 4 looks like this:

[13] H 56: 12 from <www.ibsen.uio.no>, (Accessed April 10, 2012). Owner: The University Library, University of Gothenburg.

[14] <http://www.ibsen.uio.no/BREV_1890-1905ht%7CB18981104GaG.xhtml>, (Accessed April 9, 2012).

[15] The examples of TEI-XML encoded texts produced by the HIW show some particularities: a) some elements have been changed from standard TEI and have been renamed and moved to a project specific namespace, and b) since this project started out with SGML and TEI P3

```
<div type="letter"><pb n="[1]"/>
    <dateline>
        Kristiania, den <date>4.<HIS:hisRef type="tcNote"
        xml:id="noteT15_876" target="B1890-1905ht_noter.xml"
        corresp="noteT15_876">11</HIS:hisRef>.98</date>.
    </dateline>
    <salute>Kære ven!</salute>
    <p>
        Jeg beder Dem have så hjertelig tak for alt, hvad De i den senere tid
        har sendt mig, både for <anchor type="lemma"
        xml:id="koB15_2941"/> Deres breve, som jeg har så svært for at
        <anchor type= "lemma" xml:id"=koB15_2942"/> udtyde, og nu se-
        nest for <anchor type="lemma" xml:id="koB15_2943"/> den nye
        bog af <HIS:hisRef type="person" tar-
        get="Navneregister_HISe.xml#peASt">Strindberg</HIS:hisRef>.
        <anchor type="lemma" xml:id="koB15_2944"/> Bring ham min
        varmeste og oprigtigste taksigelse derfor og sig ham at han har be-
        redt mig en i sandhed stor og overraskende glæde ved dette
        <HIS:hisRef type="tcNote" xml:id="noteT15_877" target="B1890-
        1905ht_noter.xml" cor-
        resp="noteT15_877">vidnesbyrd</HIS:hisRef>
        om at han i venlighed har tænkt på mig. Som De véd har jeg <anchor
        type="lemma" xml:id="koB15_2945"/>hans <HIS:hisRef
        type="tcNote" xml:id="noteT15_878" target="B1890-
        1905ht_noter.xml" corresp="noteT15_878">billede</HIS:hisRef>
        <pb n="[2]"/>
       ...
     </p>
</div>
```

The markup shown in the example above is used for various purposes and to identify different features. The , <dateline>, <salute>, and <p> elements are used for the document's text structure. The <pb> element (and arguably the element>) is used for the physical document structure. The <hisRef> element is used for editorial text notes and information regarding individuals while the <anchor> element is used for reference links to the general factual comments (realia). In the web version these links are activated when the reader clicks on them. The text structure elements are used mainly by the style-sheets for display purposes while the physical structure element <pb> is used both by the style-sheet and for linking to page facsimiles. By clicking on the month number 11 in the dateline, the following note appears:

11] *umiddelbart endret fra* 10
(11] immediately changed from 10)

---

and has moved through TEI P4 to TEI P5, certain possible encoding strategies allowed by P5 are not used here – one can especially note the lack of <choice> elements. More on this later in this paper.

and by clicking on the word “billede” in the last line in the example this note appears:

billede] *HIS,* billde

In this way the traditional apparatus information is included – but with a difference from the printed book: the text(-segment) in question is encoded with markup in the running text. And it has a unique identifier inside the text, the XML:id attribute. As can be seen in the encoding, the markup also references an apparatus file or a note file in the *target* attribute, in the “billede”-example *B1890-1905ht_noter.xml* as well as a location inside the target file given in the *corresp* attribute: *noteT15_878*. If we look into the B1890-1905ht_noter.xml file, we find:

```
<note resp=“editor” xml:id=“noteT15_878”><HIS:hisRef
type=”tcNote” target=“B1890-1905ht.xml”
 corresp=“noteT15_878”>billede</HIS:hisRef>] <hi
rend=“italic”>HIS,</hi> billde</note>
```

This encoding is similar to the one found in the full text version of the letter. And indeed, this makes it possible to link from the note or apparatus entry into the text, in other words a two-way link.

In the letter quoted here Ibsen writes that he has a painting of August Strindberg hanging over his desk. In the text we find a link in front of “hans billede” (his image) that links to a factual note:

```
hans billde … i mit arbejdsrum
Strindberg-portrettet av Christian Krohg, jf. brev til Susanna Ibsen 12. mars
1895 med kommentar
(hans billde ... i mit arbeidrum
The Strindberg portrait by Christian Krohg, ...)
```

The name of Christian Krohg, the painter, in the comment is also linked to the data set concerning individuals, and based on the encoding in the comment-file the style-sheet used has generated this link:

```
<http://www.google.com/url?q=http%3A%2F%2Fwww.ibsen.uio.no%2FREGINF
O_peCK.xhtml>
```

This link is bound to a particular implementation of information sets and so belongs on a style-sheet level. It is generated, however, from a more general encoding. The comment quoted above is fetched from this <item> element in a comment-file:

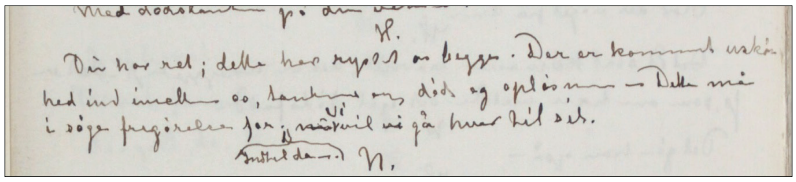
```
<item>
    <ptr type="lemma" target="B15ht.xml" corresp= "koB15_2945" xml:id="koB15_2945"/>
    <HIS:hisTerm>hans billde &hellip; i mit arbejdsrum</HIS:hisTerm>
    <HIS:hisGloss>Strindberg-portrettet av <HIS:hisRef type="person" target="Navneregister_HISe.xml#peCK">Christian Krohg</HIS:hisRef>, jf. brev til Susanna Ibsen 12. mars 1895 med kommentar</HIS:hisGloss>
</item>
```

And we can see here as well that the link between the text and the comment goes both ways.

The Dramatic Ibsen

Ibsen's letters are simple sources in some ways since there usually is only one original – no copies, no draft versions. When we move from Ibsen the letter writer to the more publicly known Ibsen the dramatist this changes: there are different manuscript versions, there are different printed editions and so on. So far HIW has selected to edit and publish the texts individually and the manuscripts in a more or less diplomatic form. The main edited critical edition is based on the first published edition, that is, the first edition is the base text for the editorial work in HIW. Going back to the diplomatic edition of manuscripts, a part of a work manuscript for A Doll's House is given in fig. 5.

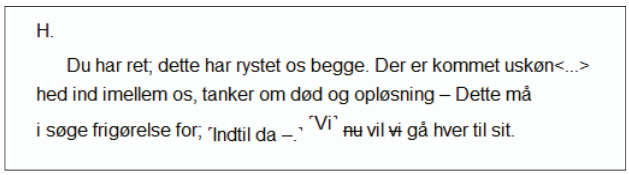
Figure 5: From a Manuscript Version of *A Doll's House*[16]



In the web publication this becomes:

[16] NBO Ms.4° 1113c, from <www.ibsen.uio.no>, (Accessed April 11, 2012). Owner: The National Library of Norway.

Figure 6: The Web Publication of the Text Shown in Fig. 5[17]


H.

Du har ret; dette har rystet os begge. Der er kommet uskøn<...>
hed ind imellem os, tanker om død og opløsning – Dette må
i søge frigørelse for; 'Indtil da –.' 'Vi' ~~nu~~ vil ~~vi~~ gå hver til sit.

The screen display showed in fig. 6 is built with the use of XSLT and CSS style-sheets from the following XML-encoded fragment:

```
<HIS:hisSp who="HELMER">
<HIS:spOpener><speaker>H.</speaker></HIS:spOpener>
<lb/>
<p>
Du har ret; dette har rystet os begge. Der er kommet uskøn<gap reason="binding"/><lb/>
hed ind imellem os, tanker om død og opløsning – Dette må<lb/>
<sic>i</sic> søge frigørelse for; <HIS:hisAdd place="infralinear">Indtil da –.</HIS:hisAdd>
<app type="alteration">
<lem>
<HIS:hisAdd place= "offline">Vi</HIS:hisAdd>
</lem>
<HIS:hisRdg>
<HIS:hisDel rend="overstrike">nu</HIS:hisDel>
</HIS:hisRdg>
</app>
vil <HIS:hisDel rend="overstrike">vi</HIS:hisDel> gå hver til sit.
</p>
</HIS:hisSp>
```

The manual for text editing in HIW says:

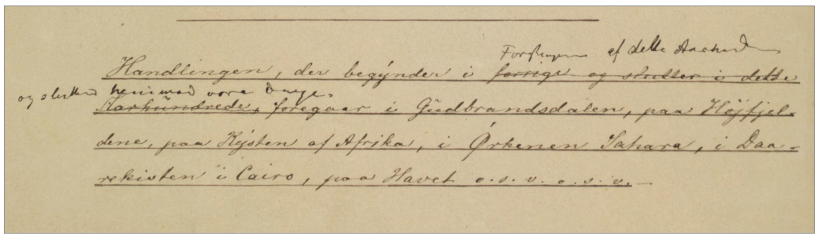
> Endringene er gjengitt så diplomatarisk som mulig, slik at tilføyelser er plassert der de er foretatt, for eksempel over linjen, og markert med innføyningstegn. Strykninger er markert med gjennomstrekning, ...
>
> (The changes are reproduced as diplomatic as possible, placing additional material where additions were made, for instance over the text line and marking the additions with special symbols. Text deleted is marked with overstrike, ...)[18]

[17] See <http://www.ibsen.uio.no/DRVIT_Du%7CDu41113c.xhtml?facs=Ja>, (Accessed April 11, 2012).

[18] See <http://www.ibsen.uio.no/tekstkritiskeRetningslinjer.xhtml> 6.3.1, (Accessed April 11, 2012).

The project's aim is to reproduce what is seen in the original as far as possible, not to analyse or interpret what is seen. One might argue that the selection of added material in the <lem> element and deleted material in the <rdg> element is in fact the result of interpretation, but this is an area where there are no absolutes, rather a question of pragmatic choices. The way the principle listed above works in more complex cases can be demonstrated with the following text fragment from one of Ibsen's best known early (or fairly early) plays, *Peer Gynt*. In a work manuscript from 1867 we find the following general setting in the introduction after the role list:

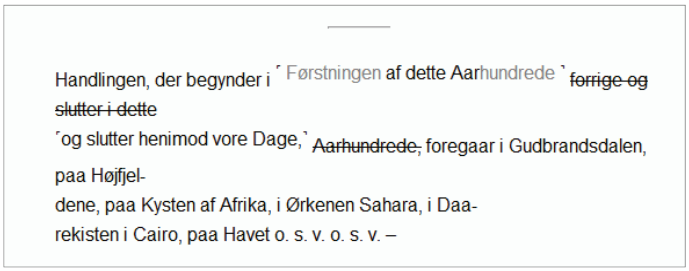
Figure 7: From a Manuscript of Peer Gynt, 1867[19]



(The action which takes place in the beginning of this century and ends near our own time and is located to Gudbrandsdalen (the Gudbrand valley), the high mountains, the coast of Africa, the desert Sahara, the lunatic asylum in Cairo, on the open sea, etc. etc.)

This is displayed in the web edition thus:

Figure 8: The Web Publication of the Text Shown in Fig. 7.[20]



[19] KBK NKS 2869, 4°, 2, from <www.ibsen.uio.no>, (Accessed April 10, 2012). Owner: The Royal Library, Copenhagen.

[20] See <http://www.ibsen.uio.no/DRVIT_PG%7CPG42869.xhtml?facs=Ja>, (Accessed April 10, 2012).

Apart from the broken lines – an artefact that comes from the column width in the web publication – we can see that the published version at least gives an indication of what the manuscript looks like. Some choices have been made, however, in what to encode and how to do it and these choices define the information content in the encoded text. This is the encoded version behind what is displayed in fig. 7:

```
<figure type="bar"/>
<lb/>
<set>
    <p>
        Handlingen, der begynder i
        <app type="alteration">
        <lem>
        <HIS:hisAdd place="offline">
        <unclear reason="writing">Førstningen</unclear> af dette Aar<unclear
        reason="writing">hundrede</unclear></HIS:hisAdd>
        </lem>
        <HIS:hisRdg>
        <HIS:hisDel rend="overstrike">forrige og slutter i dett</HIS:hisDel>e
        </HIS:hisRdg>
        </app>
        <lb/>
        <app type="alteration">
        <lem>
        <HIS:hisAdd place="offline">og slutter henimod vore
        Dage,</HIS:hisAdd>
        </lem>
        <HIS:hisRdg>
        <HIS:hisDel rend="overstrike">Aarhundrede,</HIS:hisDel>
        </HIS:hisRdg>
        </app>
        foregaar i Gudbrandsdalen, paa Højfjel-
        <lb/>
        dene, paa Kysten af Afrika, i Ørkenen Sahara, i Daa-
        <lb/>
        rekisten i Cairo, paa Havet o. s. v. o. s. v. –
    </p>
</set>
```

To a certain degree this means that the encoded version lacks some information compared to what we can see or believe we can see in the original. The most reasonable way to interpret the deletion and the insertion is that this is done as one act, that the text "forrige og slutter i dette Aarhundrede," has been deleted and the text "Førstningen af dette Aarhundrede og slutter henimod vore Dage," has been added. So why has the text been encoded like this:

```
<app><lem>...</lem><rdg>...</rdg></app><pb/>
<app><lem>...</lem><rdg>...</rdg></app>
```

and not in this way:

```
<app>
        <lem>...<pb/>...</lem>
        <rdg>...<pb/>...</rdg>
</app>?
```

One answer, as mentioned above, is that the project aims to reproduce what is seen in the original as far as possible, not to analyse or interpret what is seen. Accordingly the markup here is used for at least two different purposes. As in the example from the letter presented earlier, it is used both for the text structure, such as the use of the <p>-element, and for the document structure, where we find that the <lb>-elements are used. In what can be called the text encoding community, people now tend to show a weary and bored look if they hear the phrase "overlapping hierarchies" – much has been written and talked about this problem[21] or phenomenon. But as long as we apply an hierarchical encoding system – or an encoding system that presupposes an hierarchically organized document – this will continue to show up as a problem and will have to be handled in some way or other. For a play like Peer Gynt, the text itself shows at least two separate hierarchies in addition to the text/page hierarchies: it is a verse drama and the metrical verse lines overlap with the spoken parts. For the text/page hierarchies, the HIW made the decision to display the document structure rather than the interpreted text structure. Encoding the text or publishing a diplomatic edition of a manuscript will not make the edited published version a "copy" of the original document in every possible way so that it may always be accepted as a substitute for the original. What it can do is to give the user or reader a fairly good idea of what the original looks like. For a more exact likeness the user should be asked to look at the facsimile if the original document is not easily available. Having said this, it must be admitted that the solution chosen creates some problems for automatic handling of the text. There is nothing in the example above that can tell a computer program that the two <lem>-elements in fact belong to the same substitution. On the other hand, a computer program or a text filter that always selects the <lem> elements rather than the <rdg> elements inside the <app> elements will have the editors' selected text.

[21] See, for instance, the introduction to and discussion of this problem in Schmidt 2010, especially pp 341-44 or an early listing of areas where overlapping hierarchies exist in Durand et al. 1996.

So far we have looked at examples of in-line encoding in XML, using a modified TEI P5.[22] This solution has been chosen for the project as a matter of convenience. The project (HIW) started up in 1998/99 and the first texts were encoded in SGML (Structured Generalized Markup Language[23]) in TEI P3. In 2000 the encoding was updated to TEI P4 and XML, and in 2010 the conversion to TEI P5 started. This last conversion has basically been an adjustment of the (modified) P4 text to a (modified) P5 text that can be validated rather than a deeper change of encoding strategy such as using the <choice> element introduced in P5. One might ask, what is use of TEI (P3, P4 and P5) if it has to be modified for the project? And what is the value for the project of using a modified standard – if it can be called a standard at all in this case? Would it not be just as convenient to use something tailored especially for HIW's needs?

Although some people may not believe this, the editors and the organization behind the TEI Guidelines do not claim that the TEI P<whatever> covers every possible need. Instead, ever since the first public launching of TEI, TEI P3 in 1994[24] there have always been well defined methods for changing (adding elements, removing elements or even changing elements, changing attribute lists and more) as part of the TEI standard. How this can or should be done is described in chapters 23.2 Personalization and Customization and 23.3 Conformance.[25] If we check the modifications of the TEI made for the HIW against the check list for conformance in the Guidelines' chapter 23.3, we find that HIW documents are not TEI conform. One reason is that some of the TEI elements in the TEI namespace have had their attribute lists modified, typically by adding attributes with no specific namespace. But these changes are documented in a so-called ODD file.[26] This ODD-file can be automatically edited so that the added attributes are moved into the HIS (HIW) namespace, thus removing this problem. Still the Ibsen texts use TEI extensions so that it is impossible to generate a text document that is validated by TEI or TEI-all without loss of information just from the encoded file and the ODD-file. To simplify this: there would not have been any need for additional or modified elements if the existing elements available in TEI had been considered sufficient for the information the project wishes to encode. It is, however, possible to remove additional information and make slightly stripped-down versions of the Ibsen text files that would be TEI conform. And the question of conformance only appears in certain situations:

---

[22] See <http://www.tei-c.org/Guidelines/P5/>, (Accessed April 11, 2012).
[23] See Goldfarb 1998.
[24] See Sperberg-McQueen and Burnard 1994.
[25] See <http://www.tei-c.org/Guidelines/P5/> (Accessed April 11, 2012).
[26] ODD: "One Document Does it all" See <http://www.tei-c.org/Guidelines/Customization/odds.xml#Note1>, (Accessed April 12, 2012).

The notion of *TEI Conformance* is intended to assist in the description of the format and contents of a particular XML document instance or set of documents. It may be found useful in such situations as:

- interchange or integration of documents amongst different researchers or users;
- software specifications for TEI-aware processing tools;
- agreements for the deposit of texts in, and distribution of texts from, archives;
- specifying the form of documents to be produced by or for a given project.[27]

And even if the HIW is mainly concerned with having a codebook or a schema that allows the HIW to encode the information it wishes to encode, it has seemed well worth while to use as much as possible from the TEI Guidelines and the enormous work including abstract text and document analysis that lies behind the development of the TEI. Since the HIW works within the umbrella of TEI and TEI extensions, it also means that the project can use tools developed elsewhere, such as the Roma tools for developing schemas and ODD-files for TEI and TEI extensions, style-sheets and more. A way of encoding the Ibsen texts invented from scratch would probably have seemed less of a straitjacket – at first. But as the project has dug deeper into the texts, manuscripts and the various features that should be marked in some way or other, an unbelievable amount of work and time has been saved since one could always start out with the analysis present in the TEI Guidelines.

Another important feature with the TEI is that it uses XML as its encoding system. This means that even if a project like the HIW uses a modified version of TEI, it still uses legal, validated XML. A large amount of tools and enhancements are available for XML-encoded data. For the time being the HIW web-edition leans heavily on:

- eXist XML-textbase
- XQuery
- XSLT
- Cocoon and generators/transformations

If the project had used its own tailor-made system, tools for all these functions would also have to be not only developed but also maintained.

Standards come in many ways and in different areas. The TEI grew out of the text encoding community in the 1980s. I have argued that text encoding in its deepest form goes back to the invention of writing. But there are communities who have a history of marking and storing information that also goes back as far if not longer. When it comes to formally represented metadata the TEI itself is in some areas not very exact or complete. The <teiHeader> element where metadata about an encoded text (among other things) are stored is fairly

[27] Ibid, section 23.3.

open in some places where it allows free text (<p> elements) where other metadata standards may have more explicit requirements. In HIW, for instance, we find that groups of texts/documents are related to or represent works (for instance a play) in some way or other. In the library community the FRBR (Functional Requirements for Bibliographic Records) standard[28] aims (to simplify things a little) to organize the relationships between manifestations and works. The FRBR has its background in the library community. But then it was taken in by the museum community. ICOM's CIDOC (International Council Of Museums, International Committee for Documentation) had developed a Conceptual Reference Model (CRM)[29] which was used as a basis for what is called FRBR-oo[30] (FRBR object-oriented). And in 2008 an analysis of a possible relationship between the CIDOC CRM (and thus the FRBR-oo) and the elements in the TEI-header was presented at the DH2008 conference in Oulu[31]. The CIDOC CRM and FRBR-oo have been implemented in relational databases, for example. Having the information found in the TEI-header stored in FRBR-oo-model implemented in a relational database means that we may have some of the encoding stored outside the text itself.

The HIW has built an archive of encoded text files, one for each manuscript and for each edition included. The individual files have their own TEI-headers. Especially when it comes to witnesses for the same "work" or "text", the TEI-headers contain much duplicated information, and updating this information in individual files opens up for errors and inconsequences when the same data end up stored in slightly different ways. One simple solution is to use text entities where much of the header-information is inserted from a common file. Another way is to store the data in an external database. For a project such as the HIW, a typical content block – both for publication and for search and retrieval – is a Work in the FRBR meaning of the term. For the HIW then it seems natural to look at the FRBR, and since the project to a large part works with unique documents (for instance manuscripts) the museum-oriented approach in FRBR-oo (including the CIDOC CRM) has been selected and we are now (spring 2012) moving metadata into an FRBR-oo database.

Inline or Stand-Off Encoding?

In the letter from Ibsen to af Geijerstam shown earlier there was a short discussion of the links between comments/notes and the text of the letter:

[28] See <http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records> (Acessed April 11, 2012).
[29] See Crofts et al. 2011.
[30] See Doerr and LeBouf 2007.
[31] See Ore and Eide 2009.

```
<HIS:hisRef type="tcNote" xml:id="noteT15_878" target="B1890-1905ht_noter.xml" corresp="noteT15_878">billede</HIS:hisRef>
```

in the letter, and

```
<note resp="editor" xml:id="noteT15_878"><HIS:hisRef type="tcNote" target="B1890-1905ht.xml" corresp="noteT15_878">billede</HIS:hisRef>] <hi rend="italic">HIS,</hi> billde</note>
```

in the note file. This two-way link uses named anchors (in this case xml:id-attributes). Depending on the direction of the link we might also say that the note is a stand-off encoding of the text inside the <hisRef> element in the Ibsen source text. In principle, we might as well have had just a filename and a byte count, something like:

```
<HIS:hisRef type="tcNote" target="B1890-1905ht_noter.xml" corresp="byte count">billede</HIS:hisRef>
```

in the letter, and

```
<note resp="editor"><HIS:hisRef type="tcNote" target="B1890-1905ht.xml" corresp="byte count">billede</HIS:hisRef>] <hi rend="italic">HIS,</hi> billde</note>
```

in the note file. But if we introduce links as filenames and byte counts, why do we need this reference in the text at all? Another solution might be to store the link-information separately, either in a text file or in a database. This could give us the following three part information set:

```
billede
```

in the letter,

```
<note resp="editor">billede <hi rend="italic">HIS,</hi> billde</note>
```

in the note file, and, for example, a record in a database:

Note-Id	Textfile	T-from	T-to	Notefile	N-from	N-to
noteT15_878	B1890-1905ht.xml	<byte count>	<byte count>	B1890-1905ht_noter.xml	<byte count>	<byte count>

This data structure is identical to the one in the pure XML-files shown above in the sense that one can automatically convert from the one to the other. Stand-off encoding can also allow for encoding that is not hierarchical – just as any personally invented inline encoding scheme that allows overlap may be. This brings us back to the TEI. The intended use for the TEI is described as:

We envisage three primary functions for these Guidelines:

- guidance for individual or local practice in text creation and data capture;

- support of data interchange;
- support of application-independent local processing.[32]

As for the first bullet point, this was discussed earlier in this paper, and so to a certain degree the last one. When it comes to support of data interchange, one has to realize that some planning and work are needed if one uses an encoding system locally – inline or offline – which is essentially different from well-formed XML, such as one full of overlapping elements.

Stand-off encoding in itself has some large advantages over inline encoding. One of the most important ones is that a general stand-off encoding system that can use an URI and a byte count as a linking mechanism allows for the encoding of data stored in other places or on read-only media. Stand-off encoding using character location has also been suggested for early modern Chinese texts.[33] One practical drawback with systems using byte counts or similar addressing mechanisms is, of course, that the data and so the byte (vel sim.) address will change as well. This means that stand-off systems either have to rely on advanced systems for synchronising data and data counts or the data must be frozen or at least under control of the organization responsible for the stand-off codes. This again has led to solutions where, for instance, normal inline XML encoding is used while the documents are edited, and then the encoding is extracted and stored offline only when the document is published or frozen.[34] If for a moment we take for granted that we either have frozen data or that changes can be handled some way or other, stand-off encoding also opens up the possibility of having real Hypertext[35] and even a sort of Memex[36] and that we can share our webs of links. And there are some tools being developed right now, for instance the CATMA[37] system at the University of Hamburg. In CATMA the tags and markup information is stored in a TEI P5 XML-file with character offsets connecting the tag to a part of the source file. Version 3 of CATMA works on local files while a web-based version will be launched this spring (2012).

Conclusion

Markup is what is done so that a part of a document (a text, an image, a sound file etc.) can be identified, pointed at, and have some information or data con-

[32] See <http://www.tei-c.org/Guidelines/P5/ section iv.1>, (Accessed April 11, 2012).
[33] See Wittern 2009.
[34] Such a system, JITM (Just In Time Markup) has been suggested by Philip Berrie, see, for instance, Berrie 2000.
[35] The term Hypertext was probably coined by Ted Nelson, possible in 1965. See: <http://faculty.vassar.edu/mijoyce/Ted_sed.html>, (Accessed April 11, 2012).
[36] See Bush 1945.
[37] See <http://www.catma.de/>, (Accessed April 11, 2012).

nected with the selected data. This may be some sort of attribute value, it may be a link between two documents or between two places in the same document and so on. This markup can be done by adding codes in a text document such as XML-tags, it can be done by stand-off encoding or by a mixture of these two strategies. For local processing anything that works may be chosen, although there are good reasons to adhere to standard markup systems and tools as far as possible. For data interchange it becomes even more important to follow standards, preferably the standards most commonly used within a scholarly community. This may often lead to a setup with export and import routines between the local work formats and an exchange format.

References

[Anonymous]. *Östergötlands runinskrifter (SRI Band 2, 1911-1918)*, ed. Erik Brate. Stockholm, Kungl. Vitterhets Historie- och Antikvitetsakademien, 1911-18.

Barnes, Michael P. 2007. Rök-steinen – noen runologiske og språklige overveielser. *Maal og Minne* 2.

Berrie, Philip William. 2000. Just In Time Markup for Electronic Editions. Paper given at the Apple University Consortium Conference in Wollongong, Australia in April 2000, <http://hass.unsw.adfa.edu.au/ASEC/JITM/Wollongong200004PWB.pdf>, (Accessed April 11, 2012).

Bush, Vannevar. 1945. As We May Think. *The Atlantic*, July.

Crofts, N., M. Doerr, T. Gill, S. Stead, and M. Still, eds. 2011. Definition of the CIDOC Conceptual Reference Model. <http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf>, (Accessed April 11, 2012).

Davison, J. A. 1955. Peisistratus and Homer. In *Transactions and Proceedings of the American Philological Association*, Vol. 86: 1-22. <http://www.jstor.org/stable/283605>, (Accessed April 5, 2012).

Doerr, M., and P. LeBouf. 2007. Modelling Intellectual Processes: The FRBR – CRM Harmonization. In *Digital Libraries: Research and Development*, LNSC 4877. Berlin, Heidelberg: Springer.

Durand, David, Elli Mylonas, and Steven J. DeRose. 1996. What Should Markup Really Be? Applying theories of text to the design of markup systems. In *ALLC-ACH '96 Abstracts*, ed. Anne Lindebjerg, Espen S. Ore and Øystein Reigem, 67-70. Bergen: University of Bergen.

Euripides. 1966. *Evripidis Fabulae*, ed. *Gilbert Murray*. Oxford: Oxford University Press.

Goldfarb, Charles F. 1998. *The SGML Handbook*. Oxford: Oxford University Press. (originally published 1990).

Homer. 1969. *Homeri Opera, Tomus I, Iliadis Libros I-XII Continens,* eds. David B. Monro and Thomas W. Allen. Oxford: Oxford University Press.

Ibsen, Henrik. 2012. *Henrik Ibsens Skrifter/Henrik Ibsen's Writings*, <http://www.ibsen.uio.no>.

IFLA. <http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records>, (Accessed April 11, 2012).

Ore, C. E., and Øyvind Eide. 2009. TEI and cultural heritage ontologies: Exchange of information? In *Literary & Linguistic Computing*, ed. L. L. Opas-Hänninen, E. S. Ore and C. Warwick, Vol. 24, No. 9, Special issue, 161-72. Oxford: Oxford University Press.

Perseus Digital Library Project, ed. Gregory R. Crane. Tufts University. <http://www.perseus.tufts.edu>, (Accessed April 3, 2012).

Plutarch. 1888. *Moralia*, ed. Gregorius N. Bernardakis. Leipzig: Teubner, part 1.

Plutarch. 1927. *Moralia,* translated by Frank Cole Babbitt. Cambridge, MA: Harvard University Press. London: William Heinemann Ltd. 1.

Plutarch. 1916. *Plutarch's Lives*, translated by Bernadotte Perrin. Cambridge, MA.: Harvard University Press. London: William Heinemann Ltd. 4.

TEI P5 Guidelines. <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html>.

Schmidt, Desmond. 2010. The inadequacy of embedded markup for cultural heritage texts. *LLC – The Journal of Digital Scholarship in the Humanities* 25 (3): 237-56.

Sperberg-McQueen, C. M., and Lou Burnard. *Guidelines for electronic text encoding and interchange.* Chicago and Oxford: Text Encoding Initiative.

Ventris, Michael, and John Chadwick. 1973. *Documents in Mycenean Greek.* Cambridge: Cambridge University Press.

Wittern, Christian. 2009. Digital Editions of premodern Chinese texts: Methods and Problems – exemplified using the Daozang jiyao. In *Early Chán Manuscripts among the Dūnhuáng Findings – Resources in the Mark-up and Digitization of Historical Texts.* University of Oslo (Sep. 28 to Oct. 3, 2009), preprint PDF available. <http://kanji.zinbun.kyoto-u.ac.jp/~wittern/data/digital-editions-dzjy.pdf>, (Accessed, April 12, 2012).