Computers and the Humanities **29:** 137–148, 1995.

Speaking with One Voice: Encoding Standards and the Prospects for an Integrated Approach to Computing in History

Daniel Greenstein *
Department of History, Glasgow University, Scotland, UK
e-mail: digger@arts.gla.ac.uk

and

Lou Burnard *
Oxford University Computing Service, 13 Banbury Road, Oxford OX2 6NN, England
e-mail: lou@vax.ox.ac.uk

Key words: SGML, TEI, text encoding, history, historical documents, Sasines

Abstract

This paper focusses on the types of questions that are raised in the encoding of historical documents. Using the example of a 17th century Scottish Sasine, the authors show how TEI-based encoding can produce a text which will be of major value to a variety of future historical researchers. Firstly, they show how to produce a machine-readable transcription which would be comprehensible to a word-processor as a text stream filled with print and formatting instructions; to a text analysis package as compilation of named text segments of some known structure; and to a statistical package as a set of observations each of which comprises a number of defined and named variables. Secondly, they make provision for a machine-readable transcription where the encoder's research agenda and assumptions are reversible or alterable by secondary analysts who will have access to a maximum amount of information contained in the original source.

1. The Textual Trinity

The computer-aided management, analysis and reproduction of historical sources has developed along three parallel and, at present, frequently incompatible lines. If we ask the computer-aware historian what it is exactly that he or she uses the computer for, it is very likely that the response we receive will focus on one of three areas: printing, publishing, and word-processing; linguistic analysis and engineering; or data storage, retrieval, and analysis. We begin therefore by examining the processing techniques which characterize each of these three application areas, before proceeding to suggest that the existence of this so-called textual trinity may impose unnecessary constraints upon the future of computer-aided research. To focus discussion, we examine each processing technique with reference to one particular source – a seventeenth-century Scottish Sasine, or record of landholding and transfer. A facsimile of part of such a document is given in Figure 1.[1]

Sasines of this kind from the 17th century exist in abundance in various Scottish record offices; like other documents of the period they also show much variation. Thus, for many historians, a simple printed compilation might be a worthy aim, and one replete with learned footnotes an even worthier one. Either way, the computer will be the obvious tool to handle the

* Daniel I. Greenstein is a Senior Lecturer in Modern History at Glasgow University where he teaches and publishes in American urban and political history and culture, modern British education, and computer methods. Computing works include *A Historian's Guide to Computing* (Oxford, 1994).

Lou Burnard is Director of the Oxford Text Archive at Oxford University Computing Services, with interests in electronic text and database technology. He is European Editor of the Text Encoding Initiative's Guidelines.

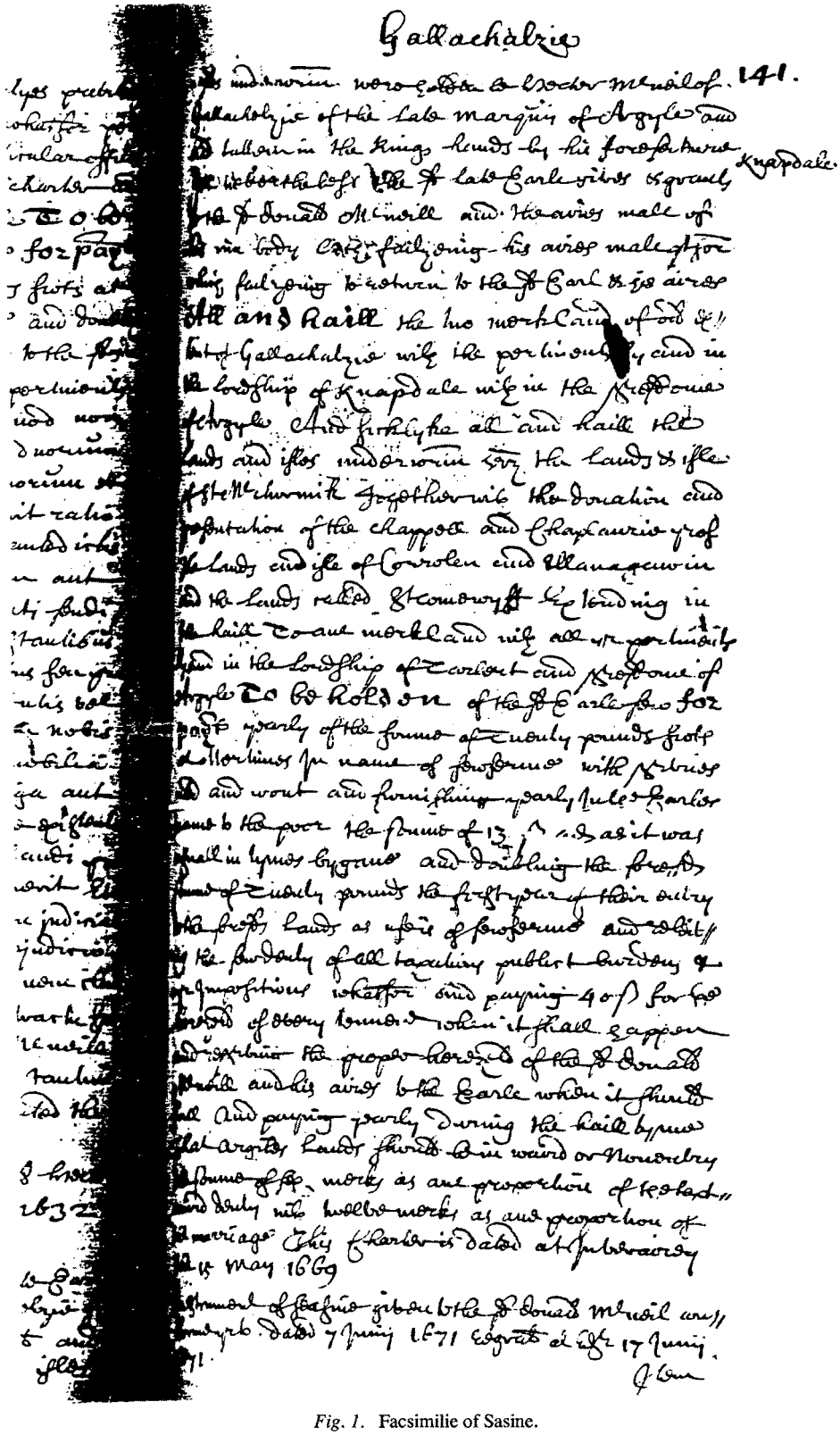

Fig. 1. Facsimilie of Sasine.

G320528 :Gallachalzie/140

Item instrument of Sasine given tha s[ai]d Hector Mcneill confirmed and dated 28 May 1632 [ms. illegible]; at Edinburgh upon the 15 June 1632

Item ane charter granted by *Archibald late earl of Argyle* to **Donald McNeill of Gallachalzie** wh[ich] makes mention that the lands [ms. illegible]; and Isles [p.141] underwritten were [ms. illegible]; by *Hector McNeill of Gallachalzie* of the late Marquis of Argyle and held fallen in the kings hand by his Foresfeiture which nevertheless the s[ai]d late Earl yields and grants to the s[ai]d Donald MacNeill and the aires male of his ane body which failing his aires male q[ua]tsoe[ve]r which failing to return to tha s[ai]d Earl and his aires .

All and hail the two merk land of old extent of Gallachalzie with the pertinents by and in the lordship of Knapdale within the sherrifdome of Argyll. And sicklike all and haill the lands and isles underwritten viz the lands & isles...

Fig. 2. Modernized reprint of Sasine.

drudgery of correcting errors, laying the text out neatly on the page and so forth. However, it is important to note that the production of a page like that shown in Figure 2 (the Sasine as it might be printed in an edited collection or critical edition) crucially depends on any number of interpretive acts, and is by no means a merely mechanical or even mechanizable process.

Even the most diplomatic of editions of a manuscript source such as this one necessarily enshrines in the authority of print one set of editorial assumptions about emphasis, illegibility, and even the interpretation of the letters of the text itself; a set with which other editors may wish to disagree or re-evaluate in the light of their own, perhaps more detailed or more specific knowledge.[2]

What is illegible to one editor may be perfectly plain to another, perhaps more conversant with Scottish legal practice; what one person regards as a standard abbreviation may require explicit expansion for another. Even experts may agree only that the decision as to whether a "u" or a "v" is intended at this point is not decidable.

For the historical linguist, the Sasine illuminates how vernacular Gaelic impinged upon English; for the historical geographer it provides valuable evidence for the evolution of place names. In such applications as these, the computer can be used to advantage but in a way rather different from that described above. The focus now is not on the orthography or presentation of the text but on its linguistic content.[3] We now present, as Figure 3, an extract from the same Sasine marked up in preparation for verbal analysis, using a standard concordance program such as OCP or TACT.

```
(( Examples of TEI encoding applied to primary
     historical sources: 17th c. Scottish Charters
   Transcribed by Lorna Hughes
   Encoded by Dan Greenstein and Lou Burnard ))
<P 140>
<S G320528><N Gallachalzie>
Item instrument of Sasine given tha said %Hector_ Mcneill confirmed and dated 28 May 1632 ((illegible)) at #Edinburgh upon the 15 June 1632 Item ane charter granted by %Archibald_late_earl_of_#Argyle to %Donald_McNeill of #Gallachalzie which makes mention that the lands ((illegible)) and /Isles
<P 141>
underwritten were ((illegible)) by %Hector_McNeill of #Gallachalzie of the late %Marquis_of_#Argyle and held fallen in the kings hand by his Foresfeiture which nevertheless the said late Earl yields and grants to the said %Donald_MacNeill and the aires male of his ane body which failing his aires male quatsoever failing to return to the said Earl and his aires All and hail the two merk land of old extent of #Gallachalzie with the pertinents by and in the lordship of #Knapdale within the sherrifdome of #Argyll And sicklike all and haill the
...
```

Fig. 3. Sasine marked up for verbal analysis.

Again, but perhaps rather more obviously than with the critical edition, the mark-up of the text is largely determined by the agenda of the analyst. Features of the text deemed of no relevance to his or her particular kind of research may be omitted from the transcription or included only partially.

In our example, parts of the text relating to proper names and currency measures have been identified with special markup codes. While this may gladden the heart of onomasticians and economic historians, clearly anyone interested in the lexical or linguistic

structure of rights and obligations as specified in a pseudo-legal document will be out of luck. Of course, no-one could be expected to identify every textual feature which might conceivably be of interest; however with arbitrary codes such as those used here, it may be difficult or impossible to tell which features an analyst has chosen to encode, let alone those which he or she has passed over in silence.

There is also no guarantee that those features which are encoded will be encoded to everyone's satisfaction. It is entirely conceivable that two researchers interested in proper names may handle the expression "I went to Rochester to get some help" in rather different ways, one preferring to interpret Rochester as a place name, and the other as a personal name. Interpretations can conflict as well where analysts agree that a particular feature exists in a text but cannot agree where it begins and ends. Where, for example, does the personal name "Archibald, late earl of Argyle" begin and end in our Sasine? Determining the boundaries of the calendar date contained in the following expression is even more problematic:

> A fortnight after the Michaelmas faire in the twelfth year of the reign of our Gracious King James.

Preparing a document for linguistic content analysis shows how the research agenda not only determines which features are selected for encoding, but how the selected features are interpreted.

For the early modern historian, the Sasine is primarily of importance as neither a textual object nor a linguistic one, but rather as a means of constructing a real world system, the historical situation from which it emerged as an artefact. In conjunction with others, this Sasine might be used, for example, to determine the date at which standard taxation rates were adopted across Scotland, replacing a taxation system based on personal relations or patronage. Figure 4 shows such a database table, in which those parts of the Sasine relating (conjecturally) to land values and taxation have been re-expressed in terms of some standardized value.[4]

When completed, this table will allow us to determine whether any fixed relationship holds across whole areas in Scotland between the estimated value of land held by given individuals and the estimated value of taxes levied upon them. Here we see the same selectivity and the use of possibly contentious interpretions already implicit in the linguistic applications discussed above carried to their logical conclusion (or to a ridiculous extreme). Any features of the Sasine not considered relevant to the summary statement of land values and taxes paid are simply not present. Also missing from this record abstracted from our Sasine are the mathematical formulae used to standardize such radically different expressions of value as commodities (the two-merk land of old extent of Gallachalzie), obligations (the donation of the chappel and chaplainies), and currencies (13 shillings and 4 pence) before they could be included in the table.[5]

Of course, no-one would seriously regard such a bare abstraction as a faithful record of the original source, but from sheer expedience (consider the daunting bulk of many types of historical source materials) there is a natural tendency to confuse highly abstracted and possibly contentious interpretations of this nature with the source itself. Irrespective of what kind of processing is required by the historian, narrowly focussed research interests and assumptions inevitably impinge upon the creation of machine-readable textual transcripts and consequently the possibility of their later re-use.

This one example of how the same textual source is prepared very differently for computer processing, depending upon the particular processing aims in view, demonstrates how the textual trinity affects not only future prospects for computer-aided humanities research, but also the limitations of our current practices. Three problems in particular are worth emphasizing. Firstly, in pursuing a particular processing aim, the historian stores information in a way which optimizes only one kind of processing. The machine-readable critical edition is all but useless to both the historical linguist and the early modern economic historian. The inefficiencies are self-evident, especially when we consider the financial constraints under which most computer-aided research work is conducted and the high cost of creating machine-readable editions of large and important corpora.

Secondly, the meta-data that the historian brings to the interpretation of a textual source and which informs its transcription, is easily lost. In constructing the database table shown in Figure 4, the historian undoubtedly drew upon detailed expert knowledge of how best to interpret and equate early modern measures (whether currencies or commodities), and perhaps on machine-readable indices and translation tables of his or her own design. Likewise the publisher's expertise in early-modern paleography is lost to the secondary analyst and to other historians who may encounter and need to interpret documents which use orthographic and scribal conventions similar to those found in the

pid	date	trans_no	land_val	tax_due
M1	320612	T1	2678	34
M1	390523	T2	4576	22
..				

Fig. 4. Sasine 'marked up' for database analysis.

Sasine. The point is that in preparing their texts for computer-aided processing, historians rely upon background knowledge which might be useful to other analysts whether or not they are interested in the initial document. This background information is lost owing to the exigencies of text-preparation.

Thirdly, while historians may not yet be the world's most prolific producers of machine-readable information, they will surely be amongst the world's most profligate consumers of the machine-readable texts that are created by others. Machine-readable administrative data is daily churned out in miles worth of magnetic tape by government agencies, hospitals, financial institutions, and industries. Consider also the machine-readable versions of reference and scholarly works daily being produced by publishing houses, even if the vast majority of these are intermediate steps in production of a printed text which are destroyed as soon as the text appears in print. Frequently, these data comprise ASCII text strings which may be more or less immediately amenable to publishing but to little in the way of analysis. A machine-readable edition of the diary of Philip Hone or Sidney George Fischer might be used to analyze networks of nineteenth-century U.S. urban elites, but only if the edition has been prepared in a way that can support both printed publication and database-style analysis. Alas, the mechanism for process-independent encoding of machine-readable data does not permit us to bridge this processing divide. And it is rather alarming to reflect that the machine-readable administrative data which are rapidly becoming the primary sources for administrative, political and diplomatic historians, are virtually inaccessible to the kinds of linguistic, content and database-style or statistical analysis which take up the lion's share of the space in our toolkit of computer-aided analytical techniques. More alarming still is the fact that these data which are the stuff of predictive measures as employed from econometrics to epidemiology, are not at all conducive to the kinds of statistical manipulations from which such measures are made.[5]

2. The Promise of Standardization

The foregoing discussion anticipates at least one solution – the development and widespread adoption of a standard markup language for encoding machine-readable data – a markup language which would make the same body of information comprehensible to different processing applications in a way which conformed to the processing applications' particular view of how data should and must be modelled. In addition, the markup language would enable researcher to document where and how their own research agenda determined how a text was interpreted for encoding purposes. With respect to the Sasine, then, our object is two-fold. Firstly, we require a machine-readable transcription which would be comprehensible to a word-processor as a text stream filled with print and formatting instructions; to a text analysis package as compilation of named text segments of some known structure; and to a statistical package as a set of observations each of which comprises a number of defined and named variables. Secondly, we require a machine-readable transcription where the encoder's research agenda and assumptions are reversible or alterable by secondary analysts who will have access to a maximum amount of information contained in the original source. In our view, the guidelines offered by the Text Encoding Initiative meet both of these criteria. To exemplify the point, the second part of this paper develops a small number of TEI solutions at some length, again with respect to the three very different processing aims which might determine how the Scottish Sasine is handled by a computer.

3. Applying the TEI Solutions

Before even considering how to ensure a machine-readable text's comprehensibility to different processing applications, agreement is necessary about how to describe the text itself. This is done in the TEI

header, one of the very few required portions of a TEI-conformant text. A header for our sample Sasine is reproduced below:

```
<TeiHeader>
<fileDesc>
   <titleStmt>
      <title>Examples of TEI encoding applied to primary
         historical sources: 17th c. Scottish Charters
         </title>
      <respStmt><resp>transcribed by</resp><name>
         Lorna Hughes</name></respStmt>
      <respStmt><resp>encoded by</resp>Dan Green-
         stein and Lou Burnard</name></respStmt>
   </titleStmt>
   <publicationStmt>
      <release>Demonstrated at AHC Conference, Bologna
         1992</release>
   </publicationStmt>
   <sourceDesc>
      <bibl>Register of title deeds for Argyle, National
         Library of Scotland Mss. Advocates 31:2:3. </bibl>
   </sourceDesc>
</fileDesc>
   <encodingDesc>
   <editorialDecl>
      <normalization method=tags>
         <p>The following common scribal abbreviations
            are expanded:
            <list type=gloss>
               <label>sd<item>said</item></label>
               <label>wch<item>which</item></label>
            </list>
         </p>
         <p>Lineation of the original preserved.</p>
      </normalization>
   </editorialDecl>
</encodingDesc>
<revisionDesc>
   <change><name>LB</name><date>April 1993
      </date><what>Revised for CHum Publication
      </what></change>
</revisionDesc>
</TeiHeader>
```

The header comprises two principal sections. A file description describes the machine-readable text itself, stating, for example, its title, the name of the person or persons responsible for its creation, information about the place, date and circumstances of its publication, and so forth. In the file description one will also find bibliographic information about the source or copy text (where relevant). A second component of the header – the encoding description – makes explicit the editorial practices followed in creating it, defining, for example, any normalization system, policy with respect to abbreviations etc. The header also contains information tracing the machine-readable text's revision history, indicating, for example, who did what to the transcript and when. The encoding description is particularly important to the secondary analyst seeking to identify, assess, and perhaps undo some of the interpretive gestures made by previous researchers who have helped to encode the text.

Within the body of a TEI-conformant text, we may expect to find an initially bewildering variety of different tags. For simplicity of discussion, we begin by distinguishing the core elements from the rest. The term *core* is used by the TEI to identify a large but not unmanageable set of textual features which are found in almost every kind of text (as opposed to specialized features characteristic of certain kinds of text only). Many, but by no means all, core tags identify what are known in the language of diplomatics as a text's *external properties*, that is, features which require little or no previous knowledge of the text or the circumstances of its creation.[7] These are features which even a Martian could recognize, though not perhaps understand. Some examples are shown in the following example, which shows the start of our Sasine marked up using the TEI core tagset:

```
<body>
<pb n='140'>
...
<div type=sasine id=G320528 n='Gallachalzie/140'><head>
<lb> Item instrument of Sasine given tha &sd; Hector
<lb> Mcneill confirmed and dated 28 May 1632
<lb> &illegible; at Edinburgh upon the 15 June 1632 </head>
<lb> <p>Item ane charter granted by Archibald late earl
<lb> of Argyle to Donald McNeill of Gallachalzie
   <expan abbrev='wh'>which</expan>
<lb> makes mention that the lands &illegible; and
<lb> Isles
<pb n='141'>
<lb> underwritten were &illegible; by Hector McNeill of
<lb> Gallachalzie of the late Marquis of Argyle and
<lb> held fallen in the kings hand by his Foresfeiture
   <note place=rmargin resp=scribal>Knapdale</note>
<lb> which nevertheless the &sd; late Earl yields and grants
<lb> to the &sd; Donald MacNeill and the aires male of
<lb> his ane body which failing his aires male
   <abbrev type=susp expan='quatsoever'>qtsoer</abbrev>
<lb> which failing to return to tha &sd; Earl and his aires
<lb> <hi rend=inked>All and hail</hi> the two merk land
      of old extent
```

<lb> of Gallachalzie with the pertinents by and in
<lb> the lordship of Knapdale within the sherrifdome
<lb> of Argyll And sicklike all and haill the
<lb> lands and isles underwritten viz the lands & isles
<lb> of Stemchmirk together with the donation and
<lb> &illegible; of the chappel and chaplainies of
<lb> the lands and isle of Cowden and Allanagenein
<lb> and the lands called Steonewytt extending in ...

Our putative Martian would have no difficulty identifying, for example, the places in the text where a page or line break occurs: these have been represented in our example by the tags <**pb**> and <**lb**> respectively. It would also probably notice that the text is not a homogenous stream but consists of small chunks or paragraphs. We have marked the start of each such chunk with the TEI tag <**p**>. Our Martian might even surmise that the Sasine is part of a larger body of material. The tag proposed by the TEI to indicate any form of textual subdivision is <**div**>, and we have adopted it here on the grounds that we wish to regard the whole collection of Sasines as a single text, of which each individual document forms a part.[8] The <**div**> tag marks the beginning of each new Sasine in the text, and a </**div**> tag marks its end.

We do not here intend to describe the syntax of the TEI scheme in any detail, since many other detailed descriptions are already available.[9] We will, however, call attention to some aspects of it relevant to our general argument. Note for example the way in which the syntax allows us to distinguish clearly the textual transcription itself from additional information which is not explicitly present in the text by presenting such housekeeping information in the form of attributes, contained within particular start-tags, for example the identification and characterization of the <**div**> element. With *attributes,* the TEI scheme allows users to employ typologies which are relevant to their own disciplines, specialisms, or research projects. Here, we see how *attributes* are used to identify relevant textual subdivisions, thus mitigating against the imposition of particular names for them such as "section" or "clause". Hence we can identify this particular div as being a "sasine". We can supply as attributes both a unique machine-friendly identifier ("G320528") and a possibly non-unique person-friendly name or number ("Marquis of Argyle 2") for the sasine, enabling it to be cross-referenced by other elements. The *id* attribute has many applications, as we will demonstrate below, and is for that reason global – any tag may carry it. Not demonstrated here, but also globally available, are attributes to document the language of a particular element and its rendition.

However intelligent, our Martian is likely to lack wide experience of terrestrial paleography and diplomatic skills which are essential to much historical research. Consequently, it might not immediately notice that this Sasine includes an annotation, nor could it reasonably be expected to identify the abbreviations within the text. These are instances of core textual features whose identification requires a modicum of background knowledge about the text and the circumstances of its creation. To identify such features an interpretive act is necessary, but they are included with the TEI core because there is at least widespread agreement that they do exist in most kinds of text, although there may be controversy about whether or not a given instance has been genuinely so classed. The marked up text includes a <**note**> element, marking the existence, whereabouts, and authorship of a marginal comment which appears on the Sasine. It also demonstrates two complementary ways that the TEI provides for the handling of abbreviations – a notorious difficulty in manuscripts from this period. The first is to specify the expanded version of an abbreviation as the element's content, at the same time retaining the original form of the abbreviation as an attribute; see the treatment of "st" for example. The second does exactly the opposite; see the treatment of the abbreviation "qtsoer", for example.

Finally, note the two so-called *entity references* used here to indicate illegible portions of the text and the familiar abbreviation "said". We wish to note the advantages they offer the encoder wishing to defer a decision as to how a particular textual phenomenon should be encoded. Some abbreviations occur so frequently that they approximate to lexical or even graphemic status: (in Modern English "Mr" is a case in point). It may be that the scribal abbreviation for "said" is in this category; alternatively, it may be that we will wish to record each occurrence using the full-blown <**abbrev**> element discussed above. The entity reference "&sd" may be expanded by an SGML processor as "said" on one occasion, as "<**abbrev expan='said'**>sd.</**abbrev**>" on another, and as some machine-specific code representing the actual symbol in the text on a third.

It is often the case that the textual features whose existence our encoding is intended to flag will be simple consecutive sequences of words or phrases in the text. It is not, however, invariably so: and the TEI proposals therefore include a number of suggestions

not only for ways of arbitrarily segmenting texts into smaller units, but also ways of recombining these arbitrary segments into meaningful analytic units of various kinds, irrespective of their physical sequence within the text.

This has a particular relevance when dealing with historical sources of information, for which such aggregation is often not only desirable, but may even be essential to an accurate analysis. A typical example is that of a geographical place name which cannot be interpreted correctly in isolation of temporal information. Thus, the expression "Germany" will have very different cartographic representations when it is used in conjunction with 1870, 1872, 1943, 1964, and 1993, respectively. Consequently, aggregation is necessary; it alone enables the researcher to combine a geographical place named somewhere in a text with the temporal information necessary to interpret it but which is available (perhaps implicitly) in a distant text segment.

As a simpler example, we consider the phrase "two merk land of old extent of Gallachalzie" in our Sasine. For the purposes of discussion, we will assume that at least one meaningful reading of these words would group "two merks of old extent" as one phrase, and "land of Gallachalzie" as another, and that it is this reading which we wish to encode. The problem is (of course) a generic one, for which the TEI proposes the following generic solution. Each fragment of the phrase must be identified as an <**s**> (segment) element and given an identifier.

```
<s id=s1>two merk</s>
<s id=s2>land</s>
<s id=s3>of old extent</s>
<s id=s4>of Gallachalzie</s>
```

Two methods may then be used to associate segments "s1" and "s3", and "s2", and "s4". The first makes use of two special attributes *next* and *prev*:

```
<s id=s1 next=s3>two merk</s>
<s id=s2 next=s4>land</s>
<s id=s3 prev=s1>of old extent</s>
<s id=s4 prev=s2>of Gallachalzie</s>
```

In the second method, the encoder simply creates a virtual <**join**> element, which does not enclose any part of the text, but simply exists to point at the two <**s**> elements and state that they are to be regarded as an instance of the element indicated by its *type* attribute:

```
<join type=measure parts='s1 s3'>
<join type=place parts='s2 s4'>
```

It is at around this point that encoders begin to worry that interpretation is taking over from transcription. As stated above, we are unconvinced that there can ever be a fundamental distinction between the two, other than that established by consensus. A neutral transcription is, we suggest, merely one in which the set of interpretative distinctions made happens to coincide with the set of such distinctions most people would wish to make most of the time; an analytic or interpretative transcription is one in which the set of distinctions made is peculiar to some specific analytic goal or agenda. It is hard to establish a more exact distinction, even with the aid of a Martian observer.

The point at which the consensus as to which textual features are actually present in a text breaks down is hard to define. For example, for many uses of the machine-readable Sasine, there is no need to postulate an analytic feature such as "cash value". It is hardly likely to be relevant to the production of a printed edition. Yet, if we wish to use the Sasine as evidence in a study of taxation practices in 17th century Scotland, some such element must be postulated. Even assuming that we are agreed on the necessity of including such an element within the encoded text, its internal structure is likely to be almost entirely different for different uses. For the historical linguist it might be sufficient simply to surround an expression of cash value with a begin and end cash value tag, or to analyze the language used to express cash valuation. For the economic historian a substructure composed of elements such as quantity, currency measure, denomination etc. would be essential. And even here, a substructure appropriate for cash values appearing in early modern historical sources would be utterly useless to the modern historian interested in cash values quoted on the New York Stock exchange say, in 1931.

In short, for analytic applications, the set of textual features which our markup must identify needs to be as open-ended as the number of different database table configurations which might be used to manage the data taken from a given historical source. To handle this situation, the TEI recommendations move "up a level". Rather than saying "the feature <**cashValue**> (or <**personalName**> or <**administrativeUnit**>) exists and has the following internal structure" it allows the encoder to provide a detailed formal *description* of each such textual feature, saying, effectively, "for my purposes, a feature exists which I will call "cash value" (or "personal name" etc), and I define it to have the following internal structure". Because this formal description uses SGML it has all the bene-

fits of interchangeability associated with other parts of the TEI scheme; what is standardized, however, is not how the analytical features are conceived but how they are described in SGML. This formal description is expressed as a bundle of *feature structures*,[10] of which the following might be an example:

```
<fs type=cashValue>
   <f name=quantity><nbr value=2900></f>
   <f name=units><sym value=ecu></f>
   <f name=reliability><minus></f>
</fs>
```

For the purposes of analysis, we have decided that we wish to decompose the single notion "cash value" into three particular *features*. Each feature has a *name* ("quantity", "units" and "reliability" in the example above), and a value, concerning which we shall have more to say in a moment. This combination or bundle of features contains all we wish to record about "cash value" for this purpose; of course, other analysts might well decide to bundle other features together under this heading or type. Somehow, therefore, we will still need to specify what features are legal within our definition, and what it means if some feature of that definition is missing; that is the function of the *feature system declaration*, which is further discussed below (see also Langendoen and Simons in this volume).

The example above also demonstrates the range of different values which features may have, including number (the amount), symbol (i.e. the name "ecu" – drawn, we assume, from a closed domain of currency names) and Boolean plus or minus (cash values, according to this analysis, are either reliable or the reverse). If this description is starting to sound not unlike the traditional description of a database table, with its predefined columns each of a specific type, that is not entirely surprising. One of the things that the feature structure notation is intended to do is precisely to permit the encoder to embed within the text the database-like structure most appropriate to his or her particular analytical aims.

However, because they use SGML, feature structures can be used to build structures rather more complex than simple columns and rows. In the following example, we demonstrate both some more examples of different feature values and the crucial fact that one feature structure can itself form the value of some feature within another feature structure:

```
<fs type=payment>
   <f name=regular><plus></f>
   <f name=duedate><str>Mertimes</str></f>
   <f name=period><sym value='annual'></f>
   <f name=place target=P1></f>
   <f name=amount>
      <fs type=cashValue>
         <f name=quantity><nbr value=20></f>
         <f name=currency><sym value='pounds Scots'>
            </f>
      </fs></f>
   <f name=stdVal>
      <fs type=cashValue>
         <f name=quantity><nbr value=2900></f>
         <f name=units><sym value=ecu></f>
         <f name=reliability><minus><//f>
      </fs>
   </f>
   ...
</fs>
```

Note here that we have two occurrences of the feature structure called "cashValue". The first appears as the value for a feature called "amount", and the second as the value for a feature called "stdVal" (i.e. standardized value). These two features are themselves part of the large bundle making up the feature structure called "payment", along with several others ("regular", "duedate", etc.). This recursion (the ability to regard a feature value as itself a feature structure, containing values which are themselves bundles of other features, which...) gives this scheme much of its expressive power, as we shall see. Note also the use of the cross-referencing ability of SGML in the feature called "place": here the value has been specified simply as a *target* – in other words, the value for this feature is a pointer to some other SGML element somewhere in the same document with the identifier "P1". Again, any similarity to the characteristics which database cognoscenti might expect or require is entirely intentional.

As mentioned above, the TEI recommendations also provide the ability to define the whole of a "Feature System Declaration", or FSD, in which the analyst specifies, inter alia,

- names, descriptions and legal values for all features;
- constraints on which features may co-exist within a given feature structure;
- rules concerning the interpretation of any underspecifed feature structure, for example default or missing values for its constituent features.

Space precludes detailed discussion of this somewhat technical topic; we note here only its striking similarity to a database schema. In the not-too dis-

tant future, standard database and statistical packages should be capable of using a TEI-conformant FSD to construct tables like the following from the feature structure given above:

Table PAYMENT

PaytNo	Regular	DueDate	Period	Amt	StdAmt
1001	Y	Mertimes	Annual	P1	A1	A2

Table CASHVALUE

Key	Qty	Currency	Reliable
A1	20	PoundScots	Y
A2	2900	Ecu	N

Reinventing the capabilities of standard statistical or relational database models, while it may keep computer science departments in business, is not in itself particularly impressive. But feature structures are far more powerful than databases. They are, for example, far better at representing two notorious areas of difficulty for the traditional "atomic" valued database approach: features whose values are composite or ambiguous. The value of a feature may be a *conjunction* or a *disjunction* of one or more feature structures. Suppose, for example, that we wish to represent the fact that a payment has been made which is a combination of two cash values, using different currencies:

```
<fs type=payment>
  <f name=amount>
    <valGrp type=conj>
      <fs type=cashValue>
        <f name=quantity><nbr value=13></f>
        <f name=currency><sym value='shillings'>
          </f>
      </fs>
      <fs type=cashValue>
        <f name=quantity><nbr value=4></f>
        <f name=currency><sym value='pence'>
          </f>
      </fs>
    </valGrp></f>
  <f name=payee><str>the poor</str></f>
</fs>
```

The two "cash values" in this example are not independent feature structures as they were in the earlier example. Here they form the constituent components of a **<valGrp>** (value group) which is being supplied as the value of the "amount" feature. The whole "payment" feature structure represents the analysis "thirteen shillings and four pence payable to the poor".

Our second example illustrates the case where the value of a feature is understood to be *one or other* of a number of feature structures:

```
<fs type=persName>
  <f name=forename><str>Archibald</str></f>
  <f name=rank>
    <valgrp type=excl>
      <sym value=earl>
      <sym value=marquis>
    </valgrp></f>
  <f name=placeName><str>Argyle</str></f>
</fs>
```

Here, we know that there is a personal name "Archibald" referred to variously in the text as having a rank of "Earl" and "Marquis" (presumably Archie remained the Marquis until his father died, at which point he became the Earl). The **<valGrp>** tag enables us to express an interpretive gesture of a type which characterizes much historical research, in this case that a given object (a person named Archie) can also be known by one of two other names but not both (he can be either the Earl or the Marquis). Try designing that into a database table!

Just as important, the feature structure links a highly structured analysis of a text with the text itself using one of two mechanisms. In this way, it is very *unlike* a database where highly structured summaries (call them interpretations if you will) are entirely divorced from the textual information which underlies them. The TEI recommends two chief mechanisms for effecting this linkage. The first embeds the feature structure within the running text which it interprets:

```
...payt yearly of the summe of
<s>twenty pounds Scots
  <fs type=cashValue>
    <f name=quantity><nbr value=20></f>
    <f name=currency><sym value='pounds Scots'>
      </f>
    <f name=stdVal>
      <fs type=cashValue>
        <f name=quantity><nbr value=2900></f>
        <f name=units><sym value=ecu></f>
        <f name=reliability><minus></f>
      </fs>
    </f>
  </fs>
</s>
at Mertimes...
```

The second uses a neater referencing mechanism whereby the text that is to be interpreted is marked up as orthographic sentences which are pointed at by feature structure notations stored elsewhere in the file or even in another file entirely:

```
<s>payt yearly of the summe of</s>
<s analysis=SCL20>twenty pound Scots</s>
<s>at Mertimes</s>

<fsLib>
   <fs id=SCL20 type=cashValue>
      ...
   </fs>
</fsLib>
...
```

Together with the FSD, this mechanism for linking analysis to the text will in future enable the researcher to treat the two as a seamless whole, to speak with one voice.

4. A Vision of the Future

The feature structure, and the linkage it promises between a machine-readable text and database or statistical style computer processing brings us full circle back to the textual trinity with which we began. Three parallel developmental trajectories are joined. The time, effort, and expense involved in the creation of machine-readable data need no longer be dedicated to one particular processing aim. The fruits of our computer-aided research need no longer be constrained by the exigencies of our initial processing aims. Imagine, then, if you will, a machine-readable transcription of Sasines prepared for the purpose of creating a critical edition. The transcription uses many of the tags described above (divs, paragraphs, line, and page breaks), some of them possibly resulting from detailed analysis of the surface features of the text (for example the abbrev, expand, and note tags). Because all such features are explicitly tagged, converting them to the appropriate set of codes for ones favourite formatting system (be it TeX, Microsoft Word, or whatever) is a simple task. Because all the encoding in the text is done in the same way, filtering out any tags which are meaningless to the printing process is equally straightforward.

At the same time, the text is available for content analysis. Another piece of interchange software enables OCP, for example, to recognize the corpus as comprising a series of independent Sasines (each treated as a distinct <**div**> element) and to filter out entity references to illegible text portions, editorial notes, etc. Admittedly, the analyst may want to add an additional layer of markup to satisfy his or her own interpretive needs, for example, by encoding place names. A TEI-conformant editor enables the analyst to scroll through the Sasines as they appear in WYSIWYG fashion on the screen (i.e. with SGML tags suppressed from view). Our content analyst uses a mouse to highlight a place name, clicks, and a semi-intelligent interface responds with the question: "what feature are you highlighting?" It provides a user-defined list of optional responses: "place name", "currency measure", etc. The encoder chooses place name, the appropriate tags are inserted in the underlying transcription, and the editing continues. Undoubtedly, the analyst will want to use some of the segmentation and aggregation methods described above to create virtual elements such as "two merks of old extent" out of text segments which are physically separate from one another. The base text is still usable for printed production because the interchange program used for producing printed texts filters out unwanted place name tags; but it is also usable now for more advanced forms of content analysis.

Our database enthusiast can also turn to the text, in this case to pursue an investigation of Scottish taxation. No database-like analyses are as yet sustained by the Sasine, but the situation is easily rectified with a semi-intelligent interface similar to that just described. Highlighting a text segment with a mouse and clicking, the encoder is prompted: "what feature structure are you highlighting?" and is provided with a range of choices defined for the software by the user-designed FSD. The choices include, "cash value", "payment", etc. Upon choosing "cash value" the user is prompted to highlight any of the features allowed in the feature structure "cash value". Again, the angle-brackets are entered behind the scenes.

Notice, too, that however the machine-readable text is processed – by a wordprocessor, text-analysis package, or database management system – it is not irretrievably transformed by a single set of narrow research interests and assumptions. Features selected by the primary encoder for linguistic-content or database-style analysis can be ignored and others marked up in their place. Disagreement about the features that are encoded is also possible because the TEI recommendations readily support multiple and even competing interpretations of a text. The textual critic can therefore re-evaluate abbreviations where

these were expanded "incorrectly"; the feature structure "cashValue" can be re-evaluated according to a different standard currency measure and/or re-worked to comprise a different array of features.

At last, the richly and fully encoded text can be used within any of the three processing environments. Assume access to an enormous terminal and we can have a quick look into four windows on the future. In the first is the encoded text itself, replete in all its ugliness with a plethora of angle brackets and their ungainly contents. In the second is the formatted text as it would appear in a printed critical edition. In yet a third window is a text as it might appear to a content analysis package consisting only of a series of divisions in which are embedded the occasional place name or currency measure. And in a fourth window is a database table whose few records show a highly summarized version of the financial information contained in the Sasine. At last we have found a single voice to express all that we want to say about our information source, and our transition from text to table is complete. Fantastic? Perhaps. But at this point in time we can say confidently that the markup language is there to sustain this level of interchange. What is required now is the software to implement it.

Notes

[1] Taken from a register of title deeds for Argyle, National Library of Scotland Mss. Advocates 31:2:3. The authors are deeply grateful to Lorna Hughes who has been particularly generous with her dual expertise in Scottish history and historical computation.

[2] See I. H. Kropac, "Medieval Documents", in D. I. Greenstein, ed., *Modelling Historical Data: Towards a Standard for Encoding and Exchanging Machine-Readable Texts* (St Katharinen, 1991), 111–16.

[3] Historians have made too little use of linguistic content analyses and stylistic measures which are so fruitfully exploited in other humanities disciplines. Fortunately there is at last some movement in this direction after a slow start. For an early discussion of how content analysis might be applied in history see T. F. Carney, "Content Analysis: A Review Essay", *Historical Methods Newsletter*, 4 (1971), 52–61. For substantive work see R. L. Merritt, "The Emergence of American Nationalism: A Quantitative Approach", *American Quarterly*, (1965), 319–35. More recent work includes P. Dawdler, "Les Déclarations des Droits de l'Homme: Une Approche Quantitative", Centre National de la Recherche Scientifique, IVe Congrès, History and Computing, Talence, 14–16 Septembre 1989: Volume Des Actes (Bordeaux, 1990), 65–73; P. Tavernier, "L'héritage de 1789 et de 1848 dans la Déclaration universelle de 1948", *Les droits de l'homme et la conquête des libertés* (Grenoble, 1988); M. Olsen and L-G. Harvey, "Computers in Intellectual History: Lexical Statistics and the Analysis of Political Discourse", *Journal of Interdisciplinary History*, 18 (1988), 456–8; D. I. Greenstein, *A Historian's Guide to Computing* (Oxford, 1994), chapter 5.

[4] Database and statistical processing is much more developed amongst historians than linguistic content analysis. For a brief review see L. D. Burnard, "Relational Theory, SQL and Historical Practice", in P. R. Denley *et al.*, eds., *History and Computing II* (Manchester, 1989), 63–71; L. D. Burnard, "The Principles of Database Design", in Sebastian Rahtz, ed., *Information Technology in the Humanities* (Chichester, 1987), 54–68; D. I. Greenstein, "Multi-Sourced and Integrated Databases for the Prosopographer", in E. Mawdsley *et al.*, eds., *History and Computing III: Historians, Computers and Data. Applications in Research and Teaching* (Manchester, 1990), 60–6 and "A Source-Oriented Approach to History and Computing: The Relational Database", *Historical Social Research*, 14:3 (1989), 9–16; P. Hartland and C. Harvey, "Information Engineering and Historical Databases", in P. R. Denley *et al.*, eds., *History and Computing II*, 44–62; C. Harvey and J. Press, "Relational Data Analysis: Value, Concepts and Methods", *History and Computing*, 4 (1992), 98–109; S. Pasleau, *Les Bases de Données en Sciences Humaines* (Liège, 1988). For the use of relational databases in medieval history see C. Bourlet and J-L. Minel, "An Expert System Decision Support System for a Prosopographical Database", in L. J. McCrank, ed., *Databases in the Humanities and Social Sciences* (Auburn, 1987); D. I. Greenstein, *A Historian's Guide to Computing* (Oxford, 1994), chapter 3.

[5] Amongst the most obviously contentious interpretions made when encoding machine-readable data are those concerned with the classification of occupational data. See M. B. Katz, "Occupational Classification in History", *Journal of Interdisciplinary History*, 3 (1972); D. J. Treiman, "A Standard Occupational Prestige Scale for Use with Historical Data", *Journal of Interdisciplinary History*, 7 (1976), 283–304; D. I. Greenstein, "Standard, Meta-Standard: A Framework for Coding Occupational Data", *Historical Social Research*, 16 (1991), 3–22. Coding notoriously fuzzy data including dates, place names, and currency measures is also problematic. See M. Thaller, "The Need for Standards: Data Modelling and Exchange", 5, and "A Draft Proposal for the Coding of Machine Readable Sources", 39, both in D. I. Greenstein, ed., *Modelling Historical Data*, pp. 1–64.

[6] See J. M. Clubb, "Computer Technology and the Source Materials of Social History", *Social Science History*, 10 (1986), 97–114; D. I. Greenstein, "Historians as Producers or Consumers of Standard-Conformant, Full-Text Datasets? Some Sources of Modern History as a Test Case", in D. I. Greenstein ed., *Modelling Historical Data*, 179–94; R. W. Zweig, "Virtual Records and Real History", *History and Computing*, 4 (1992), 174–82.

[7] See I. H. Kropac, "Medieval Documents".

[8] In fact, the TEI proposals for the encoding of corpora are rather more subtle than our simple example demonstrates, distinguishing for example between subdivisions of unitary texts (which are marked as ** **s) and composite texts or **<group>**s; a collection of Sasines would probably be better encoded using the latter method.

[9] The current Guidelines being one obvious example.

[10] At the time of writing, no public version of the TEI recommendations for the encoding of feature structures has appeared more recent than that published in P1. Our examples are based on TEI internal drafts for the relevant chapters of P2, and may therefore be inaccurate with respect to the final published version. Further and more detailed discussion of this mechanism is provided by the article by Langendoen and Simons in this volume.