Describing, transcribing, encoding, and editing modern correspondence material: a textbase approach

Edward Vanhoutte and Ron Van den Branden
Centre for Scholarly Editing and Document Studies, Royal Academy of Dutch Language and Literature, Belgium

Abstract

While letters and correspondence materials serve as (in)valuable sources of information for historians, philologists, (socio-)linguists, biographers, and textual critics, modern editorial theory merely assigns them a secondary role. Contrary to this traditional documentary view, the authors of this article argue for a treatment of epistolary materials as primary sources in their own right. They propose a generalized text-base approach of encoded and annotated correspondence materials that can accomodate the generation of versatile user-driven electronic editions. This approach needs to address current lacunae in markup theory and practice, resulting in a lack for either provisions for the encoding of letter-specific phenomena in texts, or encoding features for such generative editions. A closer look at broader editorial theories reveals a deeper lack of understanding of the nature and hence definition of correspondence materials. The authors propose a Jakobsonian communicative definition of letters that to a great deal can be mapped onto the textual model of the Text Encoding Initiative (TEI). The second part of this article discusses the motivation for and practical realization of *Digital Archive of Letters in Flanders* (DALF), a formal framework for encoding correspondence materials which is defined as a TEI customization. Its most important features for capturing detailed metadata as well as letter-specific source phenomena are analysed and discussed against the text-ontological background sketched out before.

Correspondence:
Edward Vanhoutte,
CTB-KANTL, Koningstraat 18, B-9000 Gent, Belgium.
E-mail:
edward.vanhoutte@kantl.be

This essay was written in 2002. It has not been significantly updated but reflects that particular moment in the debate on electronic text.

> *sharing the resources collected or cataloged by a documentary editing project can be one of the greatest intellectual challenges a scholar faces.*[1]

1 Introduction

It is almost impossible to list why the preservation, edition, and study of correspondence material is of great importance to our culture. Not only do archives, repositories, and collections of manuscripts and letters constitute the backbone of our cultural memory, they also provide the historian, biographer, literary and textual critic—amongst other scholars—with invaluable information concerning the people, topics, events, or works under study. The (socio-)linguist, the language geographer, and the linguist interested in diachronic or synchronic research can generate very rich data sets from well-constructed and consistently encoded corpora of transcribed correspondence material.

Literary and Linguistic Computing, Vol. 24, No. 1, 2009.
doi:10.1093/llc/fqn035

Such corpora can be subject stylometry and attribution studies, statistical research of all sorts, and lexicology.[2] With respect to scholarly editing, a conscientious analysis of the extant correspondence material between an author and their friends, publishers, illustrators, critics etc. is indispensable for a well-argumented reconstruction of the genesis of their works, and can shed new light on questions of authenticity, chronology, and assessment of the sources as well as the history of publication and reception. Letters can contain information about the dating of manuscripts, possible lost sources, (forced) revisions, and can for instance provide answers to documentary, aesthetic, authorial, sociological, and bibliographic questions. As Goethe has noticed, letters indeed are amongst the most important monuments which the individual can leave behind.[3]

In order to provide a multi-purpose and flexible access to correspondence material, the Centre for Scholarly Editing and Document Studies (CTB), a research centre of the Royal Academy of Dutch Language and Literature in Belgium, is building a gradually growing *Digital Archive of Letters in Flanders* focusing on epistolary material by authors and composers of the 19th and 20th centuries, which was acronymed *DALF*.[4] In setting up DALF as a textbase which could generate different products for both academia and a wider audience, a number of problems have been identified which will be dealt with in this article: text ontology, the role of the editor and the encoder, the choice between a textbase of transcribed correspondence material from which editions can be generated and a collection of electronic editions, the definition of a letter, the construction of a formal framework for the description and encoding of modern epistolary material, and the function of such a textbase in integrated networks assuring access to the material.

2 A Textbase of Letter Transcriptions

Letters are typically quoted or partially published in biographies and historical studies, or where appropriate (and the funding adequate) these monuments are traditionally made accessible in scholarly editions of epistolary material. But the theories of scholarly and documentary editing themselves and the medium of the printed book usually prevent a multifunctional use of their results. A scholarly edition always presents a reduced and constructed view by the editor on the available complex and simple documentary sources. This is done in compliance with specific methodological principles, serving a specific audience, and targeting at a specific goal. All of this implies many decisions on the part of the editor. This is why an edition which would serve every thinkable use and audience has never been produced. Therefore, the interested linguist for instance cannot use a collection of critically edited texts for research purposes, for they, strictly speaking, misrepresent the historical document. The call for including facsimiles of the originals to overcome this lack of documentalism, brings no solution to this problem. Facsimiles can however represent the original truthfully, but they are as static as the print medium, and leave the student with the immense task of transcribing the documentary source themselves, when they are not interested in the reading text presented by the edition. This is also true when computing the edition, for 'the goal of an electronic edition is to provide a version of the text which is encoded so as to permit *electronic* inspection, computer-assisted analysis, and retrieval, to which the raw image is inherently resistant'.[5] And, one can add, to which the print edition is inherently resistant as well.

When providing a version of the text which is encoded, the markup usually both looks backwards as representing something pre-existing (the transcription of the documentary source) and forwards to processing (the presentation of that transcription).[6] In the act of text-encoding, the structural, semantic, and renditional features are separated from the text by the use of markup which traditionally comes in two kinds: descriptive and procedural markup.[7] This severe dichotomy becomes problematic in the practice of electronic scholarly editing where editors and encoders mean and do different things according to which option is chosen for the production of an edition. Generally speaking there are three ways to produce a scholarly edition with the use of text-encoding:

(1) *Digitizing* an existing print edition.

Table 1 Allen Renear's classification of markup applied to the editorial debate in this essay

	Domain		
Mood	**Imperative**	**Indicative**	**Performative**
Renditional	<underline> *authorial*	<underline> *transcriptional*	<app> *authorial*
Logical	???	<name> *transcriptional*	<title> *authorial*

(2) *Creating* an electronic edition, e.g. by recording some or all of the known variations among different witnesses to the text in a critical apparatus of variants.
(3) *Generating* electronic editions from encoded transcriptions of the documentary source material.

The scholarly editor and the text encoder as actors in the production process behave differently in each of these three scenarios:

(1) When *digitizing* an existing print edition the encoder encodes the edition the editor has created beforehand.
(2) When *creating* an electronic edition, the encoder and the editor are one and the same person, or work as a team. The edition does not exist outside the encoding.
(3) When *generating* an electronic edition from encoded transcriptions, the encoder is an editor in so far that transcribing is editing, and the editor steers the automated generating process resulting in spin-off products. Here, when user-driven generation is provided in the end product, the user of a textbase could as well be considered an editor.

Or put simpler:

(1) The encoder encodes (transcribes) an edition.
(2) The encoder/editor creates (authors) an edition.
(3) The encoder encodes (transcribes) documentary sources, the editor steers (authors) the generation of an edition.

When relating this back to the descriptive/procedural distinction, we see some problems occurring when trying to label the markup as such and combining this with the notions of *transcriptional* and *authorial* markup. In digitizing an existing print edition, for instance, the encoder uses transcriptional markup that describes content objects such as structural units and (semantic) functions of the text, next to formatting. As Allen Renear has pointed out, there has been a universal hesitation about calling this markup descriptive: 'The reason for this hesitation is obvious: the sort of thing that is being described, a formatting effect, was always seen as the proper business of procedural markup (to invoke); and not, typically, the business of descriptive markup (to describe).' Therefore Renear suggested to refine the descriptive/procedural distinction by two functional components: *mood* and *domain*.[8] This refinement proves to be useful for our purpose in that it both articulates the differences in encoding practice amongst the three ways to produce scholarly editions, and it clarifies the role of the encoder (see Table 1).

If *digitizing* an existing print edition results in a one to one relationship between the published original and the digital representation of that original, the transcriptional markup used to describe formatting (imposed on the text by the editor) cannot be considered procedural or descriptive exclusively, but is in fact markup which is in the indicative mood (like descriptive markup and unlike procedural markup) but with a renditional domain (like procedural markup and unlike descriptive markup). The pure descriptive markup then is in the indicative mood with a logical domain. In *creating* electronic editions through the use of markup, however, the encoder/editor establishes a many to one relationship between the documentary sources and the one electronic file which documents, e.g., textual variation in a critical apparatus. By doing so, the encoder/editor both transcribes the pre-existing originals *and* constructs (authors) a tool such as the *apparatus variorum* or *criticus* which is oriented towards processing. We can say that the encoder uses authorial markup in establishing such constructs (e.g. the use of the <app> element) which

is in the performative mood with a renditional domain.[9] The author/editor is transcribing the words in the different documentary sources, and *commands* them to be a lemma or a reading inside an apparatus of a newly constructed instance.[10]

In contrast with these two ways of producing a scholarly edition, the third option is in its encoding strategies not interested in the presentation of a text, and thus strictly speaking does not look forward to processing. Instead, the transcription of a documentary source establishes a one to one relationship with the unpublished original with the use of transcriptional markup in the indicative mood with either a logical domain (when encoding functions of the text and structural units) or a renditional domain (when encoding formatting such as writing material and [double] underlined words). Whereas this encoding strategy and practice seem identical to those used when digitizing an existing edition, the difference lies in the different kind of source material it wants to encode. Where the latter strategy is aimed at establishing a one to one relationship with a published and hence formatted *edition* which can never fully represent the documentary source materials because of the authorial mediation on the part of the editor, as we have argued, the former one aims at establishing a one to one relationship with the documentary source material itself. The presentation of a text, then, is the focus of the generating process which takes these transcriptions as a basis to deliver several alternate views on that text.[11]

In assessing these three possibilities, their theories of text-encoding, their consequential roles and functions for the text-encoder and the editor, and their processibility, the third option was considered most fit for the purpose of DALF which Vanhoutte described as follows: 'a methodology and an open system architecture [...] for the digitization, markup, and presentation in on-line, off-line and hard-copy spin-off products of correspondence material'.[12]

3 Computing the Edition: Markup

It may be clear from the previous that markup technology lies at the basis of such a methodology and system. Previous work on the *Electronic Streuvels*

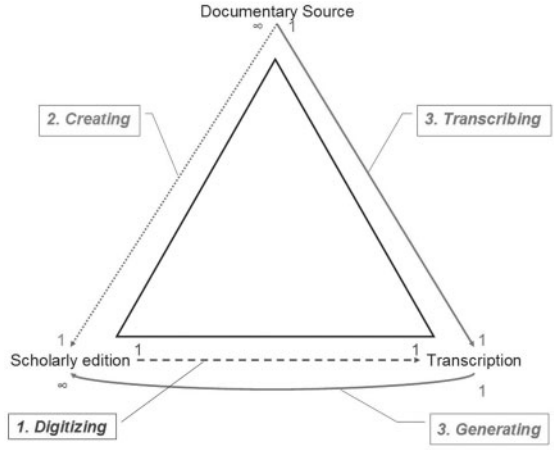


Fig. 1 The three production models in relation to their source materials

Project (ESP) which resulted in the publication of the electronic-critical edition of *De teleurgang van den Waterhoek* did recognize the fact that TEI could be used for the encoding of most of the letter contents, but added to this the need for a subset of letter-specific elements.[13] In the electronic diplomatic edition of 71 letters between the famous Flemish novelist Stijn Streuvels (1871–1969) and his publishers and friends, Vanhoutte justified the creation of a project-specific StreuLet Document Type Definition (DTD) as follows: 'Because the overall project uses the TEILite subset, and upgrading to the full DTD was at that time for practical reasons not debatable, I heavily relied on the TEI guidelines to develop a project-specific DTD which allows for both descriptive and procedural markup and which includes a set of letter-specific elements, rather than doing the theoretically unwise thing of modifying the TEILite DTD.'[14] But the 'full' TEI *could* be extended of course.

The first initiative that addressed the issue of markup for correspondence material explicitly, and which extended the TEI for this purpose, was the *Model Edition Partnership* (MEP)[15] which 'is a consortium of twelve editorial projects which publish printed editions of historical documents for the use of students, scholars and the public at large' and wants 'to develop a foundation for the next generation of historical editions', i.e. 'electronic

editions disseminated via the internet or on CD-ROM (or its equivalent)'.[16] The authors of *A Prospectus for Electronic Historical Editions* note that '[i]f historical editions are to be part of tomorrow's digital libraries, extending the TEI markup is the logical path'.[17] For this purpose, the MEP created a set of 'additional markup tags required for historical editions' such as tags for grouping the <sender>, <addressee>, and <dateline> inside a <head> at the top of the edition of the letter in the <body> of the document, the <ps> tag which contains the postscript of a letter, tags for direct and indirect references to names, people, organizations, places, or ships, etc. However, an apparent preference for a procedural approach towards their encoding strategy makes this markup scheme inappropriate for DALF, as can be demonstrated by three examples:

- The possibility to opt out of the formal TEI header in favour of the less formal <mepHeader> with ten optional elements results in hardcoding in the body of the letter metadata such as sender, receiver, date, etc. which is inferred from the contents of the letter, the envelope or some extra epistolary material and which in our opinion should exclusively go in the header of a document.[18] By documenting this information in a virtual heading of the letter, which is put there in analogy with the heading of an edited letter in a conventional letter edition, the editor/encoder switches from transcribing documentary sources to creating an edition of these sources. The alternative is transcribing the documentary sources truthfully and considering inferred information such as sender, receiver, date and place as header material from which the heading of an edited letter can be generated.[19]
- The MEP distinguishes among three levels of transcription and provides four DTDs for gradual markup. The first and basic level is 'mostly motivated by typographic phenomena (paragraphs and other blocks of type, and font shifts within blocks)' and because of that MEP recognizes that 'it is not very useful for searching a collection of documents (the sender and addressee of a document, for example, are not identified as such), but it is sufficient for producing adequate paper or screen display of documents. Since both editors and users of editions spend a lot of time reading documents, level-one markup is essential.' The second level adds hyperlinks for cross-references and notes: 'This level of markup is thus essential for online display and use of the edition, whether by editorial staff or by readers'. In level 3 markup, then, 'additional markup is added for typographically indistinct, but intellectually important, phenomena like names, dates, etc. From the level 3 DTD the data can be translated automatically into the archival form.' We doubt that merely rendering the document on the screen or on paper should be the basic interest of an editorial enterprise and think that constituting an archival form of the letter comes first. Adding hyperlinks can then be automated, and online and paper displays generated. The danger of such a gradual markup system is that projects might chose, for financial reasons or other, not to go as far as level 3 encoding, and hence will never produce satisfactory archival files.
- The MEP divides historical editions in the electronic environment into three models: image editions, live text editions, and combined editions, and suggests transitional editions as a fourth model assuring continued access to existing scholarship.[20] Although the prospectus mentions that '[i]n general, a single electronic format should be chosen for the master archival copy of the edition from which published versions can be derived by automated means (i.e. by software),' and thus suggests kinship with the third option in producing electronic editions, i.e. to *generate* electronic editions (cf. *infra*), we have seen that only the third level markup can cater for that option. One can assume therefore that different kinds of historical editions are mostly *digitized* or *created* (the two other options in creating electronic editions), and thus focus on display more than markup.

Although the MEP has proven useful for the creation and digitization of historical editions, its theoretical approach and application is not fit for the use and purpose of DALF. MEP and DALF have

in common their striving for a framework which enables the production of scholarly (historical) editions and documenting that framework as an extension to and modification of the TEI. They differ, however, where MEP is oriented towards the creation and production of such editions, whereas DALF is focused on the creation of a textbase from which such and other editions can be generated. The markup used is hence more of a *transcriptional* nature and is based on the work proposed by the TEI, MASTER (*Manuscript Access through Standards for Electronic Records*), and StreuLet.[21]

Consequently, four considerations influenced the overall design of the DTD:

(1) First, the DTD should be designed for the transcription of primary source material, from which letter editions can be generated.
(2) Second, the DTD should allow to store detailed metadata about the transcribed document. This is the point where MASTER proved useful.
(3) Third, the DTD should be able to cater for the markup of letter-specific features, such as the envelope information, the postscript etc.
(4) Fourth, the DTD should allow for a general application to letter transcriptions and editions and should not restrict itself to the specific corpus of letters currently involved in the DALF project.

But in order to 'provide a version of the text which is encoded' as Julia Flanders put it, and develop a formal framework for that encoding we must have a fairly good idea of the nature of the texts we want to encode—i.e. letters—and the features of the text which need encoding.

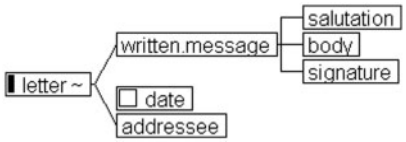
4 A Letter is a Letter is a Letter . . .

In scholarly editions of literary and musical works, relevant letters and other documents are very often not included as edited texts, but their contents is paraphrased and/or summarized in the commentary, the annotations, or the biographical or genetic article which accompanies the edition. Together with contracts, reviews, reports, diaries, etc. they belong to the so-called category of the secondary source material to which the theory and practice of historical, documentary, or non-critical editing is applied.[22] The primary source material, then, consists of the carriers of the extant versions of the work, and is edited in interaction with the theories and practices of the main schools of critical editing and literary, genetic, or textual criticism.[23] The status and treatment of correspondence material in scholarly editing, therefore, depends not on its appearance or form, but on its contents and function in a wider context. When a letter contains (unpublished) works or versions of a work, it is upgraded to the primary source league, and the versions or variants it contains could either feature as a reading text or could be mentioned in the apparatus criticus or variorum. Otherwise, if published, the (edited) correspondence material appears as a subsidiary 'product' to the scholarly edition. Stated otherwise: letters can be valuable sources for the annotation or commentary of critical edition of works, but they themselves need to be annotated and edited as well.

Not only is the treatment of correspondence material problematic in scholarly editing, there is also no common agreement on the very definition of a letter. A letter is not a type of text as is prose, poetry, or drama, but the concept of a letter hints at a form which can contain all types of texts, images, or objects. The first step in building a textbase of correspondence material, therefore, is to define precisely what the project considers a letter. Only then, find and retrieval tools, user interfaces etc. can be built for the exploitation of the textbase, and the material can be edited.

In his article 'Some Notes on Letter Editing' the German editor and theorist Siegfried Scheibe provides us with a useful definition which we take as our starting point. According to Scheibe, [b]y a ''letter'' we understand written messages, information or orders which are meant to inform other persons (the addressee) and which were not written for publication. As a rule, they have a standard form beginning with a salutation, ending with a signature and frequently containing a date.'[24]

This definition can be formally represented by the following algorithm:



Scheibe suggests this definition because of his dissatisfaction with the impracticality of the postal definition of a letter, but he fails to precise what that postal definition is. We assume that the definition the dictionary gives, approaches what Scheibe had in mind. For our purpose, we use the denotation given by *van Dale Groot Woordenboek der Nederlandse Taal*, the standard dictionary of Dutch. A letter is: 'Writing in the form of an address, to one or more absent persons, aimed to let him (them) know something, closed and sent supplied with an address.'[25]

This can be formalized as:

At first sight, both definitions indeed differ from each other, but when we assume that sending a letter implies a stamp, a postmark, and thus a date,[26] and when we take into account the knowledge from reality that the formal characteristics such as 'salutation' and 'signature' are but optional features, both definitions seem to be identical. That is, apart from the distinction Scheibe makes amongst several types of letters: 'written messages, information or orders', which cannot be formally represented but function as qualifiers.[27]

The Dutch editor and author of *Naar de Letter* a handbook on scholarly editing, Marita Mathijsen, elaborates on Scheibe's definition by pointing at the vital communicative relationship between the author of the letter and the addressee, and by adding to it that the letter, which she defines as 'text', is meant to be sent or handed over to the person to whom it is addressed.[28] With Scheibe, she concludes that therefore the public letter, the letter to the editor, and the epistolary novel can all be excluded from treatment in an edition of correspondence material, just as contracts, memoranda, claims, and proofs of payment. Scheibe adds to this that business-like documents which are transmitted in the form of a letter likewise do not belong in a letter edition.[29] We do not agree with Scheibe on this point, for the correspondences from and to authors, artists, scientists, statesmen, etc. not only contain invaluable information for the study of the genesis, production, meaning and reception of their work, for the reconstruction and a better understanding of the contemporary society and mentality in which they lived, worked, loved, and interacted with each other and with an audience, but correspondence material gradually becomes more important to economic, psychological, political, and social sciences, and to the history and philosophy of science as well. These disciplines consider the so-called *Aktenschriftstück* of equal importance as private letters, and very often they contain complementary information. The line along which private correspondence and business letters differ from each other is very hard to draw. In an essay on the edition of Charles Darwin's correspondence, Frederick Burkhardt contends that '[s]ome memoranda are signed and can with good reason be considered letters that simply lack a salutation and valedictory.'[30] One can even point out, be it as a rhetorical statement, that if the business of an author or composer is writing and composing respectively, each letter dealing with the act and the problems of writing and composing should thus be considered business letters and hence should not be included in an edition.

The formal (objective) structure which defines a text as a letter, thus, cannot be used as an argument in favour or against inclusion in an edition of letters. The quality of privateness which has to be (subjectively) assessed by the editor,[31] on the contrary, is an argument.[32] Therefore, even the most documentary or diplomatic treatment of correspondence material is always subjective in its selection.

In an attempt to draw principles for the selection of documents for an edition of correspondence material, the basis question 'What is a letter?' often is inverted and leads the debate of what to include in an edition to an enumeration of what *not* to include in an edition, and consequently to a negative definition of what a letter

is *not*. Irmtraut Schmid, for example defines a letter as follows: 'A letter is a communicative piece of writing which has not become a business document.'[33]

So far, the same negative principle was involved in the rejection of these definitions of a letter. In the following part, we will try to establish a positive definition by building on their commonalities.

5 A Simple Description of Letters

In the definitions we have seen so far, the communicative function of a letter features as a common denominator. According to Irmtraut Schmid, the urge to transfer a message and thus to establish a communicative situation is one of the common motives for the origin of all extant documentary material which she calls the 'testimony of the past'.[34] A relevant model for the analysis of this communicative situation has been offered by the Russian philologist Roman Jakobson who described six constitutive factors[35] in his communicative model which he linked to communicative functions. His model schematizes as follows[36]:

Table 2 Jakobson's communicative model

Semiotic factor	Illustration	Communicative function
Addresser	Sender	Emotive
Message	Message in code	Poetic
Addressee	Receiver	Conative
Context	Extra-linguistic reality	Referential
Code	Natural language	Meta-lingual
Contact	Letter	Phatic

When we consider the letter as physical channel through which the communicative situation is established, this model provides us with the elements for a simple formal description of a letter:

- the sender (<sender>);
- the message (<body>);
- the recipient (<recipient>);
- references to the extra-linguistic reality (e.g. <title>, <name>, <place>, <date>, etc.);
- the language of the letter (<language>); and
- the physical characteristics of a letter, e.g. the collation (<coll>), the envelope and its features (<envelope>), etc.

Almost all of these features can be encoded using the DTD subsets as proposed by the *TEI Guidelines for Electronic Text Encoding and Interchange*.[37] But letters are much more complicated material than suggested by the communicative model, as will become clear from the rest of this article. Therefore, we chose to extend these guidelines to enable a detailed transcription of correspondence material for input in our textbase.

In what follows we will briefly describe the customization process, then elaborate on the problems of storing detailed metadata about the transcribed document and the markup of letter-specific features, with a focus on an important problem area for the encoding of handwritten primary source material, namely the complex correlations of logical and physical structures.

6 Customization of the TEI

The TEI encoding scheme provides an excellent starting point for many of the features one would like to encode in letters. However, the TEI tagset has some generally acknowledged lacunae regarding the encoding of primary source material.[38] Moreover, apart from some features relating to primary manuscript material in general, an open collection of digitally encoded letters requires the means to encode some very specific textual and meta-textual features that are not covered by the TEI scheme. An attempt at overcoming this 'representational vacuum' could be made by looking for standard TEI tags that most closely correspond to the structural and semantic requirements of the (meta-)textual elements to encode. The advantage of having a plain vanilla TEI DTD for the encoding of correspondence material wouldn't, however, outweigh the disadvantage of this DTD only *looking* like a TEI DTD, for some of its elements would have different meanings or distributions than described in the *TEI Guidelines*. Moreover, such a solution does not guarantee a faithful representation of the text

ontology. For example, straightforward as it may seem, encoding a postscript as would force an uncomfortable view on the textual status of the letter. Since a element, as the *TEI Guidelines* document, is meant to indicate a subdivision of a text, this would not fit a postscript very well. There is no reason to consider a postscript more of a subdivision than a paragraph, with its own <p> tag, or a salutation formula, with its <salute> tag. Even if such an ontological-theoretical objection would be accepted in a model for letter encoding, it still leaves the markup-theoretical fact that in this case, a TEI element would be used for something other than a real subdivision.

Realizing the limitations of the text ontology it offers and anticipating the danger of intolerable stretching of TEI semantics, the TEI DTD has been provided with structural extension mechanisms that are clearly documented in Chapter 29 of the *TEI Guidelines*. The arguments pointed out above motivated the adoption of this approach in the construction of the DALF DTD. Of the 281 elements it contains, 221 are taken over from the TEI tagset and 60 are uniquely defined for DALF. The extensions and modifications—encoded in DALFExtns.ent and DALFExtns.dtd files—were input in the Pizza Chef program on the interactive TEI website,[39] which neatly produced the DALF.dtd file. The following parameters were used:

- a 'mixed base' tagset was selected, consisting of the prose and drama bases,
- elements from the additional tagsets linking, figures, analysis, transcr, textcrit, and names.dates were selected
- the entity sets ISOlat1, ISOlat2, ISOnum, and ISOpub were selected.

These selection parameters are inspired by the diverging types of information conveyed in letters. Being highly authorial ego-documents, they come in many form(at)s, thus requiring many representational means to capture them. The choice for a mixed base tagset reflects the multi-faceted status of the letter regarding conventional text 'genres'. It is highly plausible to encounter fragments of poems, dramas, novels, synopses and the like in the correspondence of writers.[40] Also the elements of the additional tagsets are selected to anticipate the manifold textual features in letters on the one hand and the desires of encoders to make more elaborate annotations on the other hand. The ISO entity sets ensure a computer- and human-readable annotation of non-standard textual characters, allowing the encoding of, for example, the character ü with the ISO reference *ü* rather than the more obscure Unicode reference *ü*.

As the emphasis of this article lies in specific issues concerning the encoding of correspondence material, the following sections will illuminate the way in which some of those motivated the declaration of the 60 unique DALF elements, rather than discuss the TEI elements included in the DALF DTD (for a discussion of which the reader is directed to the excellent TEI reference documentation).

7 Some Relevant DALF Elements

The most typical structure for a letter encoded with the DALF scheme may look like this:

```
<TEI.2>
    <teiHeader>...</teiHeader>
    <text>
        <envelope>...</envelope>
        <body>
            <opener>
              <address>...</address>
              <dateline>...</dateline>
              <salute>...</salute>
              . . .
            </opener>
            <p>...</p>
            <closer>
              <salute>...</salute>
              <signed>...</signed>
              <ps>...</ps>
              . . .
            </closer>
        </body>
        <back>
            <note>...</note>
```

```
                <join>...</join>
                ...
            </back>
        </text>
</TEI.2>
```

The conception of DALF as a growing textbase linking archive records to the actual contents of the transcribed letters requires the possibility to encode a rich amount of letter-specific meta-information in the header from which catalogue entries, indexes, etc. can be generated. This is of vital importance to the organization and management of the textbase, functioning either as a stand-alone digital archive or in an integrated network. The text then is the place for the actual transcription of the documentary source.

7.1 The header

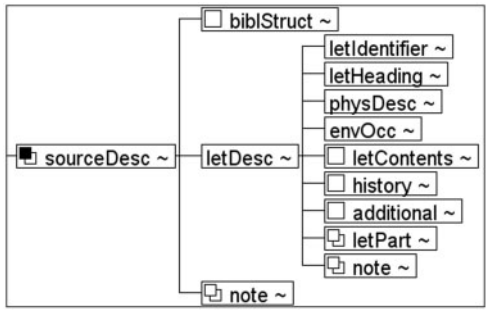
As already mentioned, a useful model for the extension of the TEI header was found in the DTD and guidelines developed for the MASTER project, which was 'intended primarily for the detailed cataloguing of medieval and early modern manuscript materials in the Western European tradition'. The envisioned outline of DALF differs, however, in two major aspects from the goals of the MASTER project:

- DALF is designed as a textbase of *transcriptions* and *electronic editions* of primary sources (whereas, the MASTER project aims at a database of manuscript description records).
- DALF focuses on an archive of transcriptions of *modern correspondence* (whereas, the MASTER project is aimed at documenting medieval manuscripts).

Instead of extending the MASTER DTD (a TEI extension itself), we used the MASTER scheme as a source of inspiration for a distinct TEI customization. This allowed us to add or delete elements, and to streamline their naming, attributes and content models where appropriate, adhering to the following principles:

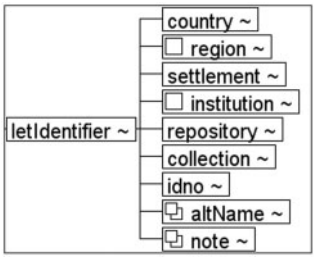
- In order to keep the analogy and semantics transparent, all letter-specific elements in the header have been named starting with 'let-'.
- In order to ensure consistent encoding of DALF documents and to facilitate their integration into a searchable electronic database, we opted for a fairly strict design of the header. This resulted in several mandatory elements or choices between alternatives.
- To ensure flexibility, optional <note> elements are allowed after mandatory contents.

The main feature of the DALF DTD is the extension of the <sourceDesc> element of the standard TEI header with <letDesc>, grouping meta-information about each encoded document. This <letDesc> element is required for each DALF document, and contains mandatory elements for the encoding of information about the identification, catalogue summary, physical description, and presence of an envelope. Optionally, information about the contents, history, additional aspects, or specific parts of the letter may be provided.



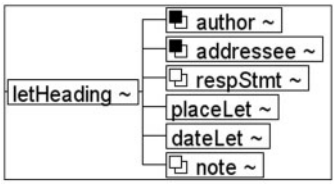
We will not deal with each element of <letDesc> in detail, but instead we will just run through a presentation of the element content models which should be more or less self-declarative.

- *<letIdentifier>:* identifies the letter both on the macro-level (<country>, <region>, <settlement>, <institution>, and <repository>) and on the micro-level (<collection>, <idno>, and <altName>). Most of the letters that will constitute the DALF database will be unpublished primary manuscripts that are stored in private or public collections, located at particular places. Therefore, there is need for a more detailed level of cataloguing description than for published source materials, providing researchers with a uniform and accurate system of archival reference for each letter.



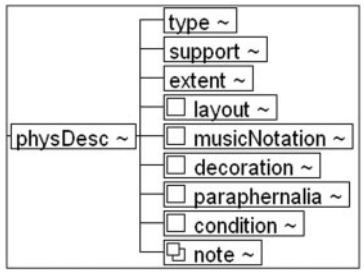

```
<letIdentifier>
        <country>Belgium</country>
        <settlement>Antwerp</settlement>
        <repository>AMVC</repository>
        <collection>S 935 / 62295</collection>
        <idno>171373/2882</idno>
</letIdentifier>
```

- *<letHeading>:* allows for a structured description of bibliographical information of a letter. One of the essential characteristics of letters is their close relationship with the particular communicative context in which they are created. Of course, this also holds for published books, written by a certain author and at a certain place and time. Yet, as bibliographic references to books show, those particular communicative circumstances of the writing act are deemed less important than the circumstances of publication. In contrast, when referring to letters as unambiguously as possible, one has to include as much of the communicative particularities possible. Those are so important that they may be considered an essential part of the bibliographical identification of a letter. This element is the place for abstractions about sender, addressee, place, and date of the letter. The elements comprised by this element have the attribute 'attested' with the possible values 'yes' when the abstraction is made on the basis of evidence inside the letter, 'added' when it is made on the basis of material accompanying the letter, or 'no' when it is derived from external evidence.




```
<letHeading>
        <author attested="yes">Stijn
                Streuvels</author>
        <addressee attested="yes">Maurice
                De Meyer</addressee>
        <placeLet attested="no">Ingooigem
                </placeLet>
        <dateLet attested="added">1945-01-13
                </dateLet>
</letHeading>
```

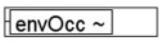
- *<physDesc>:* describes aspects of the physical appearance of letters. As letters can be very different with regard to their physical realization, the description of physical aspects contains a limited set of elements for features that are shared by all letters, and additionally provides the possibility to encode a rich array of specific phenomena when they occur. Required elements are a characterization of the format of the letter, a description of the material on which the letter is written, and an indication of the physical size of the document. Additionally, general aspects of layout can be pointed out, as well as a characterization of possible fragments of musical notation, a description of possible decorative elements or paraphernalia, and the condition of the letter as a physical object.



- *<envOcc/>:* documents the occurrence of an envelope. The envelope can contain valuable information for the contextualization of a letter, or even contain text that may be closely related to the contents of the letter. However, the question whether or not to regard this text as part of the letter is a theoretical one. Some encoders may wish to exclude envelope contents completely; others may consider it relevant enough to include

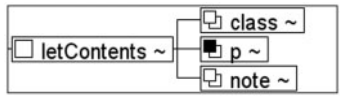
```
<physDesc>
        <type>letter</type>
        <support>
            <p>single page with pre-printed
                letterhead, writing on one
                side only</p>
        </support>
        <extent>
            <dimensions>
              <height units="mm">214</height>
              <width units="mm">276</width>
            </dimensions>
        </extent>
</physDesc>
```

it as part of the letter. The DALF encoding scheme allows (and strongly suggests) a middle road, by providing the special <envelope> body element (see further) that enables the encoding of envelope content while at the same time keeping it separate from the letter proper. In order to facilitate retrieval of documents in the DALF textbase, the element <envOcc/> that explicates the presence or absence of an envelope is adopted as a mandatory element of the letter description. It is an empty element that has one required attribute @occ, which can have either 'yes' or 'no' as its value.




```
<envOcc occ="no"/>
```

- *<letContents>:* describes the contents of a letter. The contents of the letter are of course a major source of interest for users of encoded DALF materials. In order to ensure a rich possibility of exploiting the DALF textbase, structured access to descriptive records of its cumulative contents is an interesting starting point. It can be used to implement a search mechanism that lets users do thematic queries throughout several (selections of) DALF letters, or provide the means to generate thematic editions like a regest or calendar edition. This is also the place to document a formal characterization of a letter according to a certain typology in a <class> element. We think of a functional-communicative 'genre'-indication of letters (cf. the Jakobsonian communicative functions).



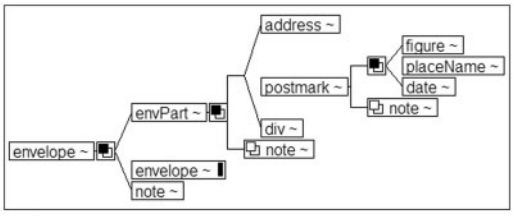

```
<letContents>
    <class>love letter</class>
    <p>Streuvels proposes to his girlfriend</p>
</letContents>
```

7.2 Letter-specific text elements

Considering now the actual encoding of letters itself, the TEI had to be extended to cater for some letter-specific features. Obviously, there are some structural elements that are unique to letters, like the envelope and postscripts. Others are more generally bound to primary manuscript material, and thus occur very frequently in letters, such as calculations, pre- and post-printed materials, and decorative elements.

- *<envelope>:* contains the information on the envelope. The typical letter is delivered within an envelope. Often, when letters are stored, their accompanying envelopes are stored with them. For an encoder, there are good reasons to provide transcriptions of letters with a transcription of their envelopes. One is that envelopes may contain valuable information for the identification of letters. When the letter is lacking some of the indicators of the communicative context that are important for an unambiguous identification of a letter (communicative participants, time, and place of writing), there is good chance they still can be deduced from the postal information on the envelope. Furthermore, the envelope may contain significant information, apart from the postal data. Some authors create on their envelopes pieces of art in their own right that may closely relate and contribute to the letter content. Some receivers may use the envelope to write quick notes about the letter content or other contextual circumstances. Envelopes may even contain drafts of successive letters. In response to such diverse types of textual content, the means are provided to document the occurrence of a

postmark, addresses on front and/or back, possibly additional plain text, or no text at all, and even the containment of another envelope. The format can be supplied as a value for the 'type' attribute in <envelope>. The occurrence of an empty envelope can be signalled within <note>. Text on an envelope can be encoded within an <envPart> element, with an attribute @side to specify the envelope part concerned. Envelope text can be an address of the sender or recipient, a postmark, or plain text.



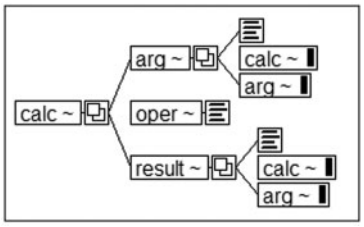

```
<envelope>
  <envPart side="front">
    <address type="addressee">
      <addrLine>De Heer Styn Streuvels
          </addrLine>
      <addrLine>"Lijsternest"</addrLine>
      <addrLine><hi rend="underlined">
          INGOYGHEM</hi></addrLine>
    </address>
    <postmark>
      <date value="1924-01-04">4.I.1924
          </date>
      <placeName><place>ANTWERPEN
          </place></placeName>
    </postmark>
  </envPart>
</envelope>
```

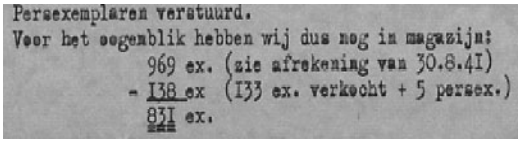
- *<ps>:* contains a postscript in a letter or letter part. Postscripts are a typical phenomenon for letters. Occurring after the closing formulae and salutation, they form a last addition to the contents of the letter. Moreover, the author often explicitly signals this additional status with the abbreviation 'P.S.'. Their formulaic use and meaning justify an own tag. Therefore, the <ps> element is adopted in the DALF DTD to appear only at the back within the <closer> element.

```
<closer>
  <salute>Met vriendelijken groet</salute>
  <signed>(Styn Streuvels)</signed>
  <ps><p id="xr2"><add id="add1">
    <abbr expan="postscriptum">P.S.</abbr> Ze
    jubileeren bij de firma Veen
    (60 jaar bestaan)<ptr target="n8"/>
    en er wordt me daarom gevraagd, door
    het comit&eacute;: hoeveel geld ik
    daarvoor als feestgave wensch te geven!
    Zonderlinge zeden? Als ik nu eens
    vroeg: hoeveel ze voor mij beschikken
    als 75-jarige jubilaris!</add></p>
  </ps>
</closer>
```

- *<calc>:* contains a calculation. Calculations frequently occur in modern correspondence. Formally, calculations are often set apart from running text, and it may be desirable to mark them with explicit encoding features. This provides researchers with greater control over the textual features they want to study. Calculations have an internal structure the semantics of which cannot be captured sufficiently with the standard TEI <num> element. We considered the option to incorporate MathML, an existing W3C standard providing a specialized tagset for mathematical formulae. In Chapter 22.2 of the TEI P4 guidelines, directions are given for specifying external tagsets as XML notations that can be used in the <formula> tag. However, testing that mechanism with the literal examples given in that chapter turned out unsuccessfully. Further investigation of postings on the TEI public mailing list[41] showed that other TEI users encountered the same problems, and learned that the incorporation mechanism itself does not provide the inclusion functionality we had in mind. These troubles and the unwieldy suggestions to get around the incorporation of external tagsets like MathML,[42] as well as the complexity of the MathML standard itself made this option less favourable than devising a specialized element that can encode at least some of the semantic structure of calculations in a

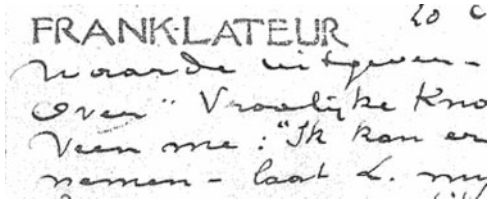
rudimentary way. Calculations may partly or entirely consist of plain prose (thus possibly needing some phrase-level TEI elements), and can contain embedded calculations. The basic structure, however, is made up of one or more arguments, an operator, and a result.




Persexemplaren verstuurd.
Voor het ogenblik hebben wij dus nog in magazijn:
969 ex. (zie afrekening van 30.8.41)
- 138 ex (133 ex. verkocht + 5 persex.)
831 ex.

```
<calc>
  <arg>969 <abbr expan="exemplaren">ex.
      </abbr> (zie afrekening van
      30.8.41)</arg>
  <oper>-</oper>
  <arg>138<abbr expan="exemplaren">ex
      </abbr> (
    <calc>
    <arg>133 <abbr expan="exemplaren"> ex.
        </abbr> verkocht</arg>
    <oper>+</oper>
    <arg>5 <abbr expan="persexemplaren">
        persex.</abbr></arg>)
    </calc>
  </arg>
  <result><hi rend="double underlined">
        831</hi>
    <abbr expan="exemplaren">ex.
      </abbr>
  </result>
</calc>
```

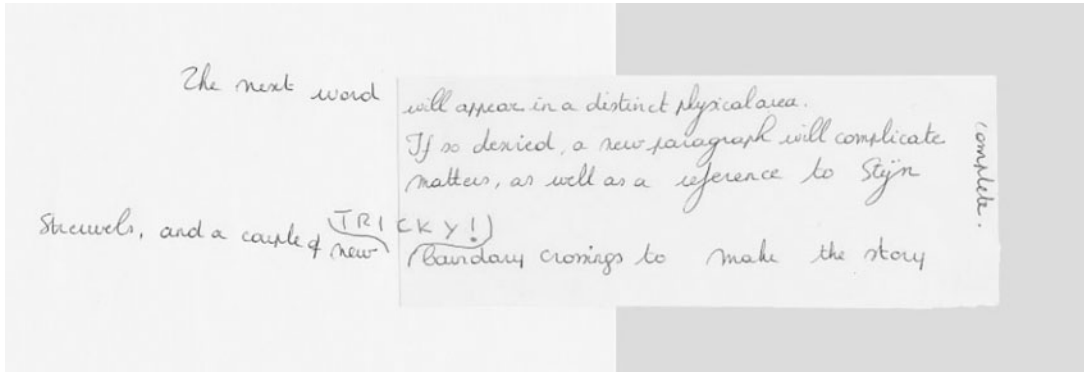
- *<print>:* signals printed material that was present on the carrier of the letter, or added afterwards. Letters may be written (or printed) on paper (or other support material) containing pre-printed text like letterheads, form data, newspaper articles, ads and so on. There are also similar formal text elements, like stamps, that may be added after the composition of the letter. Such text fragments can be seen as part of the letter, but may need to be distinguished from more 'authorial' parts of the letter, as they mostly have an impersonal character. It is imaginable that such material would be excluded from e.g. a linguistic study on the language of a writer, or selected in a study on stamps in letters. The TEI tagset does not contain any element that can accurately indicate pre-printed text material. Post-printed material, like stamps, could possibly be tagged with the TEI <add> element. However, as that element is reserved for 'letters, words, or phrases inserted in the text by an author, scribe, annotator, or corrector', it is questionable whether mostly impersonal stamps can be regarded as genuine additions in that sense. Therefore, in order to provide consistent treatment for all pre- and post-printed material, the <print> element is included in the DALF DTD.


FRANK·LATEUR

```
<print
type="letterhead">FRANK&middot;LATEUR
</print>
```

8 Correlations of Logical and Physical Structures

As emphasized in the chapter 'A Gentle Introduction to XML' of the TEI Guidelines, every encoding effort involves the abstraction of an underlying interpretation: 'Encoding a text for computer processing is in principle, like transcribing a manuscript from scriptio continua, a process of making explicit what is conjectural or implicit, a process of directing the user as to how the content of the text should be (or has been)



interpreted'. The development of the DALF DTD has thus been an effort to establish an interpretative framework for the encoding of letters. However, in this process, we saw ourselves confronted with more fundamental boundaries of the format chosen to express this framework. The well-formedness constraint of proper nesting of elements in XML imposes a so-called 'Ordered Hierarchy of Content Objects (OHCO)' view on XML documents, and hence on XML-encoded versions of existing texts. This forces some abstraction on the encoding scheme that can remain fairly implicit as long as no competing hierarchies are involved. A problem arises, however, when one wants to make explicit assumptions of the text which are situated on different structural levels. Since our approach to the encoding of handwritten primary source material is deliberately documentary, we encountered some markup-technical as well as theoretical difficulties in aligning this with the mainly logical orientation of the TEI markup scheme on which the DALF scheme is built

8.1 Technical difficulties: overlapping hierarchies

In the course of the encoding process of the letters in this project, an important issue was raised concerning the use of the TEI scheme for the transcription of chronological, authorial, and physical aspects of (handwritten) primary sources.

Consider the above (fake) example which illustrates the documentary complexity we're sometimes dealing with: a number of paragraphs by author 1, including a shift in writing direction, with an addition by a second author 2 written across two physically distinct pieces of paper.

The overlap of logical structures (paragraphs), authorial and chronological features (the addition), and physical structures (the distinct pieces of paper, the writing orientation) are of interest to us.

The TEI P4 guidelines propose a number of ways to deal with similar overlapping structures. One is the use of *empty elements* 'to mark the beginnings and endings of regions of the text which have something in common'. As these elements do not contain the text region as their content, they can interfere with other hierarchical units without causing overlap. Although logically transparent, this approach encumbers the processing of these structures by XML software. Another possibility is to use start and end tags for all these different structural levels, but to avoid overlap by splitting up overlapping structures into several elements. This, however, complicates the encoding and does not allow straightforward processing.

This problem of encoding multiple overlapping hierarchies with XML is generally acknowledged, and a number of other solutions (independent of TEI) are proposed. Thompson and McKelvie suggest the construction of separate files containing (a premature form of) XLink that create virtual spans of annotation in a source text.[43] Durusau and O'Donnell propose a model of 'Bottom Up Virtual Hierarchies', which involves the physical separation of different markup hierarchies into different files (each file holding a different markup 'view' on the text), and a virtual join of these markup schemes in

a derived base file. In that base file, each word is encoded separately, and an attribute is included for each distinct hierarchical layer. Those attributes get as their value XPath expressions that record the hierarchical position of the word concerned in the according hierarchy.[44] Other suggested solutions abandon the XML paradigm and investigate the development of other formats allowing real encoding of overlapping elements. These include MECS and TexMECS.[45] More recently, a data structure was proposed, named GODDAG, that should allow the representation in and translation between MECS, TexMECS, and XML, but 'is still very much a work in progress'.[46] Durusau and O'Donnell note that the standoff methods of approaches such as TECS and TexMECS 'suffer from a variety of defects, ranging from non-implementation to a fairly high degree of notational complexity.' Also, their own BUVH approach does not seem to be a really practical working solution, as the project website states that the verbosity and processing difficulties of the BUVH model inspired them to investigate a new thread of research, named 'Just-In-Time-Trees' (JITTs).[47] In another paper, the authors explain that this proposal involves adapting the XML processing model itself so that it allows multiple root elements and thus overlapping hierarchies. Also this proposal is very much (the announcement of) work in progress.[48]

In the absence of a ready-to-use, generally accepted way of dealing with overlapping hierarchies in XML, a TEI-conformant solution was adopted in the DALF DTD, exploiting the above mentioned suggestion of using empty elements, and centred around a notion of textual 'layers'. Bearing in mind the processing difficulties inherent to empty elements, a mechanism was adopted to express a minimal 'layer' view on documents, in a way that avoids both the possibility of overlap with the encoding of logical structures, and also an overgeneralised use of empty elements. However interesting a very neutral 'layer' concept might be, allowing the encoder to distinguish and define several types of layers (chronological, physical, authorial, thematic,...), this would hold the danger of allowing a great deal of textual structures to be marked up with empty elements. In order to minimize this strategy as much as possible, the semantics of a 'layer' concept was reduced to a physical level. This means that a letter can be seen as a complex of physical layers, pieces of physical containers for the logical structures. When different layers are distinguished, their boundaries can be indicated with empty start and end tags to virtually enclose the physical area concerned, in a way that does not disturb the proper nesting of other elements. The logical aspects that were distinguished at the start of this discussion (chronology and authorship) can then be represented by standard TEI tags and attributes, like <add> and 'hand'. The physical phenomenon of writing orientation can perhaps be indicated with a <seg> element and a 'rend' attribute. For the demarcation of physically distinct layers the tags <layerStart/> and <layerEnd/> are introduced, with a 'layer' attribute referring to the corresponding definition in the header. Those elements can be given an own 'id', in order to facilitate the virtual linking of these points in the text by means of a <link/> or <join/> element. Thus, assuming a layer definition with id "l2" for the added paper in the <profileDesc>, a suggested encoding of the example given above looks like this:

```
<TEI.2>
  <teiHeader>
  ...
    <profileDesc>
      <handList>
        <hand id="hand2"/>
      </handList>
      <layerList>
        <layer id="l2" type="post-it"/>
      </layerList>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
    ...
    <p>The next word <layerStart layer="l2"
        id="ls1"/>will appear in a
        distinct physical area</p>
```

```
    <p> If so desired, a new paragraph will
        complicate matters, as well as a
        reference to <name>Stijn <layerEnd
        layer="l2" id="le1"/>
        Streuvels</name>, and a couple of new
        <add hand="hand2">
        tri<layerStart layer="l2"
        id="ls1b"/>cky!
        </add> boundary crossings to make
        the story <seg rend="90">com
        <layerEnd
        layer="l2"
        id="le1b"/>plete.</seg>
    </p>
    ...
    </body>
    <back>
        <join targets="ls1 le1 ls1b le1b"
            result="div" desc=
            "physical layer"/>
    </back>
  </text>
</TEI.2>
```

```
<teiHeader>
...
  <layerList>
  <layer id="l2" type="label"/>
  </layerList>
...
</teiHeader>
<text>
<body>
  <p>Gister-morgen heb ik de proef en den
      brief,<ptr target="n1"/> om 8½u
      met de <hi type="underline">gewone
      </hi> post ontvangen en beide stukken
      met een etiket voorzien-:<layerStart
      id="l1s" layer= "l2"/><print type=
      "newspaper"> REMIS A
      LA POSTE<lb/> l' adresse étant inexacte
      <lb/> ou insuffisante.<lb/></print>
      <layerEnd id="l1e" layer="l2"/>
      W&agrave;t er wel aan de adressen
      mankeerde of te kort was, kan ik
      niet raden.
  ...
  </p>
</body>
</text>
```

8.2 Theoretical difficulties

Even with a technically satisfying solution for the encoding of hierarchical overlaps, the correlations of logical and physical structures need serious theoretical attention. For example, in this edition project, the following occurrence of a shift of physical layers was encountered. On 7 October 1924, Stijn Streuvels received a printer's proof by mail with on the envelope a sticker which said 'Sent back to the post. The address was not correct or not complete'.[49] The next day, the author wrote an answer to his publisher who was the sender of the proof, stuck the label on the letter and commented on it in the text of the letter. He wrote: 'Yesterday morning I received the proof and the letter by regular mail at 8.30 a.m., and both pieces were labelled as such-: [label] I cannot see what was wrong with the addresses or how they were incomplete'.[50]

When transcribing the letter, the key question is whether this external material is part of the letter or not. The logical answer is yes. The author refers to the text on this label, thus incorporating the external object in the logical contents of the letter. Also regarding the chronology of writing, this object forms an organic part of the letter with the previous and next words. It can be assumed that Streuvels first wrote the sentence in which he mentioned the importance of the label, then pasted that fragment on to the letter and then continued writing the letter. Although this sounds chronologically sound, there are some discontinuous aspects involved. The text on the label predates the moment of writing of the letter. In addition, the author of this fragment is not Streuvels himself. Another discontinuity is obviously situated at the physical level.

This example does of course provoque daunting questions about the nature of text, that should deserve thorough theoretical investigation for every encoding endeavour. Such theoretical issues can be illustrated with a possible encoding for this example, according to the minimal layer model

presented above. When the physically distinct layer is defined in the header, the empty elements <layerStart/> and <layerEnd/> can refer to it while signalling its boundaries in the text. Regarding the logical level, the external text cannot be tagged with <add>, as it is as much an addition as each distinct word written by Streuvels. However, the authorial shift for the text printed on the sticker should be articulated. This can be done with the <print> element with a value of 'newspaper' for its @type attribute, which can also be used when newspaper clippings etc. become part of the letter.

9 Computing the Edition: Access

After this overview of specific markup aspects of the DALF textbase and the efforts made to adhere to international standards, a final observation can point out the theoretical similarities between our model and two converging trends in the information sciences, aiming at making accessible archive materials.

If letters are amongst the most important monuments which the individual can leave behind, as we cited Johan Wolfgang von Goethe at the beginning of this paper, striving towards better access to documents is enhancing the chances for a better understanding of our culture. In the last decade, we can see this striving at work, for instance, in two concrete developments in the world of libraries, archives, documentation centres, and museums: the digital library and networked information management. The digital library which Deegan and Tanner define as 'a managed collection of digital objects'[51] can of course itself become a node in a network, as can digital catalogues, digitized (location) registers, etc., but it is mainly aimed to make the digital objects[52] available 'in a cohesive manner, supported by services necessary to allow users to retrieve and exploit the resources just as they would any other library materials'.[53] But where the digital library provides the user access to long-term stable resources in the form of image, text, hypertext or hypermedia, the managed information networks provide the user with information about resources which often do not live in digital form. A relevant example of such network is the MALVINE project[54] (MAnuscripts and Letters Via Integrated Networks in Europe) which aims to 'provide overall access to existing item level description catalogues about modern manuscript holdings via a search engine'.[55] MALVINE enables the user to search and retrieve descriptions of modern manuscripts and letters, both on the collection level and on the item level, from a heterogeneous group of libraries, archives, documentation centres, and museums through one location (URI) and a multilingual interface. At the core of the integrated network lies a common European metadata format based on EAD[56]/XML with which the data from the participating institutions have to comply before the search engine can access it via Z39.50.[57] The ultimate goal in the development of MALVINE is to offer the user harmonized access to heterogeneous databases and the possibility to request digital images of every document.

With respect to epistolary editing, this integration of the content oriented approach of the digital library and the pure cataloguing interest of managed information networks such as MALVINE has rarely been put into practice, and could be the topic of another paper. Therefore, more research should be done on the function of e.g. archival descriptions in metadata[58] for both the textbase and the generated spin off products, as well as for the integrated network providing access to that material.

10 Conclusion

In order to profit from all advantages electronic editions of modern correspondence material offer the editor, researcher, or the wider audience, a formal framework must be agreed on for the description, transcription, and encoding of the documentary source material. Only from a rich textbase of encoded correspondence material, several spin off products can be extracted and realized, such as scholarly editions, reading texts, indexes, catalogues, calendars, regests, polyfunctional research corpora etc. This formal framework should cater for both the detailed documentation of metadata about

the archival finding and the detailed description (caption) of its contents. This way the resulting textbase could integrate the functionalities of the digital library and managed information networks in offering the user all sorts of (user) generated views on the material or 'products'. In developing such a formal framework, much effort should go to its documentation. The resulting document, guidelines, guide to good practice, or technical files will be the starting point for every debate, improvement, or training, and will facilitate a consistent input in the textbase and a continuous assessment of its contents. DALF could be a good attempt at reaching these goals.

Notes

All URLs last accessed 22 May 2002.

1 **Kline, M.-J.** (1998). *A Guide to Documentary Editing*. 2nd edn. Baltimore and London: The Johns Hopkins University Press, p. 70.

2 Eide mentions in this respect that the Norwegian *Documentation Project*, which is digitizing literary texts and editions of correspondence material, was originally meant 'to create a corpus to be used in the creation of a national Norwegian literary dictionary' [**Øyvind Eide.** (2003). Putting the dialogue back together: re-creating structure in letter publishing. *Computers and the Humanities*, **37**(1): 67].

3 For the codicological ramifications, see **Shillingsburg, P. L.** (1996). *Scholarly Editing in the Computer Age: Theory and Practice*, 3rd edn. Ann Arbor: The University of Michigan Press, p. 16. Goethe's remark remains a touchstone: 'Briefe gehören unter die wichtigsten Denkmäler, die der einzelne Mensch hinterlassen kann' cited in **Buschmeier, G.** (1999). 'Die Auswertung von Komponistenbriefen in deutschen Forschungseinrichtungen' (summary in English: 'Evaluation of Composers' Letters within German Research Institutions.') *Info RISM*, **10.** http://rism.stub.uni-frankfurt.de/–inforism/InfoRISM_Vol_%2010.htm (accessed 22 May 2002).

4 See: Centrum voor Teksteditie en Bronnenstudie <http://www.kantl.be/ctb/>. Koninklijke Academie voor Nederlandse Taal- en Letterkunde <http://www.kantl.be>. All information, encoding guidelines and DTDs can be downloaded from the project webpage <http://www.kantl.be/ctb/project/dalf/>. The Royal Academy of Dutch Language and Literature has since its founding in 1886 had an interest in the edition of letters. The first concrete plan for an analogue textbase of correspondence material was proposed in 1948: see **Schmook, G.** (1949). Pleidooi voor de uitgave van de brieven van negentiendeeuwse Vlaamse figuren. *Verslagen en Mededelingen van de Kon. Vlaamse Academie voor Taal- en Letterkunde*, **1949**: 23–42.

5 **Flanders, J.** (1998). Trusting the electronic edition. *Computers and the Humanities*, **31**(4): 305.

6 **Piez, W.** (2001). Beyond the 'Descriptive vs. Procedural' distinction. *Markup Languages*: *Theory & Practice*, **3**(2): 141.

7 'Markup that describes the structure and other attributes of a document in a non-system-specific manner, independently of any processing that may be performed on it. In particular, SGML descriptive markup uses tags to express the element structure' [**Goldfarb, C. E.** (1990). *The SGML Handbook*. Oxford: Clarendon Press, pp. 137 and 262]. 'Descriptive markup describes/characterizes/identifies a text component/feature/part'; 'procedural markup invokes/specifies/commands a formatting/rendering procedure/effect/process/action' [Renear, A. (2000). The descriptive/procedural distinction is flawed. *Markup Languages: Theory & Practice*, **2**(4): 412].

8 **Renear, A.** (2000). The descriptive/procedural distinction is flawed. *Markup Languages: Theory & Practice*, 2(4): 413. Mood is defined as: 'whether markup describes something, or requests processing,' and domain as 'the sort of thing being described, or requested' (p. 417).

9 This example fills the empty slot in Renear's classification of authorial markup (Renear, p. 417, Fig. 4).

10 We cannot say that the author/editor *makes* the words into a lemma or a reading (which would mean they were in the logical domain), for these are constructs of the mind in the context of reading several documentary sources together.

11 See in this respect the Archive/Museum model in **Vanhoutte** (2000). See note 14.

12 See *Dancing with DALF: Towards a Digital Archive of Letters written by Flemish authors and composers in the 19th and 20th century*, paper presented at the ACH/ALLC conference, New York, 13 June 2001.

13 See **De Smedt, M. and Vanhoutte, E.** (2000). *Stijn Streuvels. De Teleurgang van den Waterhoek: Elektronisch-kritische editie/Electronic-critical Edition*. Amsterdam: Amsterdam University Press. For the TEI (Text Encoding Initiative) see <http://www.tei-c.org>.

14 **Vanhoutte, E.** (2000). Where is the Editor? Resistance in the Creation of an Electronic Critical Edition. In Deegan, M. Anderson, J. and Short, H. (eds), *DRH 98: Selected Papers from Digital Resources for the Humanities 1998*. London: Office for Humanities Communication, pp. 171–83. This paper is available online at <http://www.hb.se/bhs/ith/1-99/ev.htm>.
15 http://mep.cla.sc.edu/.
16 http://mep.cla.sc.edu/mepinfo/mep-info.html.
17 See the text at <http://mep.cla.sc.edu/mepinfo/MEP-Docs/proptoc.htm>.
18 The ten optional elements are: <preparedBy>, <prepDate>, <copyright>, <permissions>, <docAuthor>, <docTitle>, <docDate>, <dateline>, <sourceDesc>, and <idno>.
19 All citations in this paragraph taken from <http://mep.cla.sc.edu/MepGuide.html>.
20 'Image Editions presenting images of original documents together with retrieval and search tools and supplementary material; [. . .] live Text Editions presenting transcriptions of original documents together with retrieval and search tools and supplementary material; [. . .] combined Editions presenting the documents in both image and live text form together with retrieval and search tools and supplementary material; [. . .] transitional Editions presenting page images of letterpress editions made accessible through live indices and search engines; or page images of typeset volumes with live transcriptions of documents not published in the letterpress volumes linked with live indices and search engines.'
21 'MASTER is a European Union funded project to create a single on-line catalogue of medieval manuscripts in European libraries. This project has developed a single standard for computer-readable descriptions of manuscripts. It has created software for making these records, and tested the standard and the software on descriptions of some 2000 manuscripts. Many of these records will be mounted in a single networked catalogue, available to everyone. MASTER is funded under the Framework IV Telematics for Libraries call.' Online at <http://www.cta.dmu.ac.uk/projects/master/>. See Section 7 below for detailed information about the influence of MASTER on DALF.
22 See **Kline** (1998) for a concise treatment of documentary editing (see note 1).
23 See **Van Hulle, D.** (1999). *Textual Awareness. A Genetic Approach to the Late Works of James Joyce, Marcel Proust, and Thomas Mann*. Ph.D. thesis, University of Antwerp; (published in 2004 by Michigan University Press).
24 **Scheibe, S.** (1988). Some notes on letter editions: with special reference to German writers. *Studies in Bibliography*, **41**: 137.
25 '[G]eschrift in de vorm van een toespraak, tot een of meer afwezige personen gericht om hem (hen) iets te doen weten, gesloten en van een adres voorzien verzonden' [**Geerts, G. and Heestermans, H.** (eds). (1989). *van Dale Groot Woordenboek der Nederlandse Taal*, Elfde herziene druk. Utrecht/Antwerpen: Van Dale Lexicografie, p. 448].
26 This leads to the question whether an unsent letter still qualifies as a letter. The edition of the correspondence of Charles Darwin, for instance, includes all drafts, even when it is uncertain whether a final version has ever been sent. Cf. **Burckhardt, F.** (1988). Editing the correspondence of Charles Darwin. *Studies in Bibliography*, **41**: 149–52.
27 In a formal DTD representation they would be defined as attributes of the element <written.message>.
28 Mathijsen's complete definition: 'A letter is a text which is intended to maintain or establish contact between the author and a non-anonymous person or a number of non-anonymous persons who are connected to each other and which is originally not intended for publication. The use of a form of address is characteristic, as is the presence of an opening and a signature. The text is meant to be sent or given to the one to whom it is addressed' ('Een brief is een tekst die gericht is op het in stand houden of leggen van contact tussen schrijver en een niet anoniem persoon of een aantal niet anonieme personen die met elkaar in betrekking staan en die in eerste instantie niet geïntendeerd is voor openbaarmaking. Gebruik van de aanspreekvorm is kenmerkend, evenals de aanwezigheid van een aanhef en een ondertekening. De tekst is bedoeld ter verzending of overhandiging aan degene aan wie hij gericht is'). See **Mathijsen, M.** (1991). Het Dilemma van de brievenediteur. *Gezelliana*, **1992/1**: 16.
29 Scheibe, 'Some notes on letter editions', 138 (see note 24); Mathijsen, 'Het dilemma van de brievenediteur', 16 (see note 28).
30 **Sulloway, F. J.** (1983). Further remarks on Darwin's spelling habits and the dating of beagle voyage manuscripts. *Journal of the History of Biology*, **16**: 361–90; and '**Burckhardt** (1988 see note 26), esp. 152,' have demonstrated the importance of Darwin's correspondence for a correct reconstruction of the chronology of his theories.
31 The suggestion by Winfried Woesler not to include typewritten or computer printed material in a letter

edition cannot be held valid any more in an era where the computer has for a great deal replaced the pen: '*Drucksachen*, *Computerbriefe*, vervielfältige *Rundschreiben* werden in der Regel nicht aufgenommen' [**Woesler, W.** (1988). Vorschläge für eine Normierung von Briefeditionen. *Editio*, **2**: 9].

32 Paradoxically it is exactly this quality of privateness which sometimes obstructs the publication of the edition. When dealing with modern correspondence material which is still under copyright, the estate often uses the privateness of the letter as an argument not to grant permission for publication or use in an edition.

33 'Ein Brief ist ein Verkehrsschriftstück, das nicht zu einem Aktenschriftstück geworden ist'[**Schmid, I.** (1988). Was ist ein Brief? Zur Begriffsbestimmung des Terminus 'Brief' als Bezeichnung einder quellenkundlichen Gattung. *Editio*, **2**: 4].

34 'Bei allem überlieferten Schriftgut haben wir es mit Zeugnissen der Vergangensheit zu tun, für deren quellenmäßige Nutzung es nicht unwichtig ist, aus welchen Motiven heraus sie entstanden sind. Eines dieser Motive ist das der Mitteilung' [**Schmid, I.** (1988). Was ist ein Brief? Zur Begriffsbestimmung des Terminus 'Brief' als Bezeichnung einer quellenkundlichen Gattung. *Editio*, **2**: 3].

35 'The ADDRESSER sends a MESSAGE to the ADDRESSEE. To be operative the message requires a CONTEXT referred to ('referent' in another, somewhat ambiguous nomenclature), seizable by the addressee, and either verbal or capable of being verbalized; a CODE fully, or at least partially, common to the addresser and addressee (or in other words, to the encoder and decoder of the message); and, finally, a CONTACT, a physical channel and psychological connection between the addresser and the addressee, enabling both of them to enter and stay in communication' [**Jakobson, R.** (1971). Linguistics and Poetics. In Sebeok, T. A. (ed.), *Style in Language*. Cambridge, MA: M.I.T. Press, p. 353].

36 Cf. **Geeraerts, D.** (1989). *Wat er in een woord zit. Facetten van de lexicale semantiek*. Leuven: Peeters, pp. 130–54.

37 **Sperberg-McQueen, C. M. and Burnard, L.** (eds). (2002). *TEI P4 Guidelines for Electronic Text Encoding and Interchange: XML-compatible Edition*. Oxford, Providence, Charlottesville, and Bergen: The TEI Consortium. See also <http://www.tei-c.org/P4X/>.

38 Chapter 18 of TEI P4 concludes with a hint at future expansion of the TEI tagset in this respect. Perhaps projects like MASTER or the current DALF project, explicitly focusing on the encoding of primary manuscript material, can contribute with insights, or with stimulation of the public debate on this matter.

39 http://www.tei-c.org/pizza.html.

40 The only reason for not including the 'verse' base tagset is that the presence of <lg> and <l> elements in the core tagset was deemed sufficient to encode poetry that may occur within letters.

41 http://listserv.brown.edu/archives/tei-l.html.

42 A full technical discussion can be found in Arjan Loeffers' answer on 28 October 1997 to Andreas Nolda who formulated nearly the same problem we encountered. These messages can be found in the TEI List archive.

43 **Thompson, H. S. and McKelvie, D.** (1997). *Hyperlink Semantics for Standoff Markup of Read-only Documents*, *Proceedings of SGML Europe '97*, Barcelona, May 1997, online at <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>.

44 **Durusau, P. and O'Donnell, M. B.** (2001). Implementing Concurrent Markup in XML. *Extreme Markup*. Montreal, online at <http://www.sbl-site2.org/Extreme2001/Concur.html>.

45 **Huitfeldt, C.** (1998). *MECS: A Multi-Element Code System*. Online at <http://helmer.hit.uib.no/claus/mecs/mecs.htm>. **Huitfeldt, C. and Sperberg-McQueen, C. M.** (2001). *TexMECS. An Experimental Markup Meta-language for Complex Documents*. Online at <http://www.hit.uib.no/claus/mlcd/papers/texmecs.html>.

46 See **Bjørnestad, B. K. and Meurer, P.** (2002). A Computational Model for MLCD. Presentation at AHC/ALLC. Tübingen [typescript], 15.

47 **Durusau, P. and O'Donnell, M. B.** (2002). *Concurrent Markup for XML Documents*. Presentation for 'XML Europe'. Online at <http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/03-03-07/03-03-07.pdf>. See further <http://www.sbl-site2.org/Overlap/>.

48 **Durusau, P. and O'Donnell, M. B.** (2002). *Coming Down From The Trees: Future of the Evolution of Markup?* Paper presented at the Extreme Markup Languages conference, Montreal. Online at <http://www.sbl-site2.org/Extreme2002/JITTs.ppt>.

49 'Remis à la poste. L'adresse étant inexacte ou insuffisante.'

50 'Gister-morgen heb ik de proef van den brief, om 8½ u met de gewone post ontvangen en beide stukken met een etiket voorzien-: [label] Wàt er wel aan de adressen mankeerde of te kort was, kan ik niet raden.'

51 **Deegan, M. and Tanner, S.** (2002). *Digital Futures: Strategies for the Information Age*. London: Library Association Publishing, p. 241.

52 'Although a large proportion of a digital library's collection comprises materials that are born digital, such as e-journals, internet resources, databases, and so on, there are many resources that are not originally created in digital form, but are digitized in order to include them in a digital library's collection' [**Chowdhury, G. G. and Chowdhury, S.** (2003). *Introduction to Digital Libraries*. London: Facet Publishing, p. 103].

53 Deegan and Tanner, *Digital Futures*, p. 241 (see note 51).

54 See http://www.malvine.org

55 **Weber, J.** (2002). Malvine, Leaf and Kalliope: Some Co-operation Models. In Dongelmans, B. A. Leerintveld and van der Weel, A. (eds), *Digital Access to Book Trade Archives: Papers of the 2001 Conference in The Hague*. Leiden: Academic Press, p. 50.

56 Encoded Archival Description, online at <http://lcweb.loc.gov/ead/>.

57 'Z39.50' refers to the International Standard, ISO 23950:'Information Retrieval (Z39.50): Application Service Definition and Protocol Specification', and to ANSI/NISO Z39.50. The Library of Congress is the Maintenance Agency and Registration Authority for both standards, which are technically identical (though with minor editorial differences). The standard specifies a client/server-based protocol for searching and retrieving information from remote databases. Online at <http://www.loc.gov/z3950/agency/>.

58 The specific (technical) problem of integrating EAD encoded finding aids with TEI encoded transcripts and digitized images of archival material is the concern of the LEADERS project based at University College London (Linking EAD to Electronically Retrievable Sources: Online at <http://www.ucl.ac.uk/leaders-project/>).