The making of TEI P5

Christian Wittern
Kyoto University, Kyoto, Japan

Arianna Ciula
King's College, London, UK

Conal Tuohy
New Zealand Electronic Text Centre, Wellington, New Zealand

Abstract

The TEI Consortium has taken on the task of maintaining the *Guidelines for Electronic Text Encoding and Interchange*. This article describes how the latest major revision to these Guidelines was developed over the course of >6 years by the members of the TEI Technical Council and workgroups charged and overseen by the Council and gives background information and reasoning for the decisions taken. Among the new additions for P5, two of the most outstanding, the chapters on Names, Dates, People, and Places and on digital facsimiles are treated in some more detail. The article concludes with a brief account of the decisions made with respect to customization and conformance.

Correspondence:
Christian Wittern, Institute for Research in Humanities, Kyoto University, 47 Higashiogura-cho, Kitashirakawa, Sakyo-ku, Kyoto 606-8265, Japan.
E-mail:
wittern@zinbun.kyoto-u.ac.jp

1 Introduction

The Text Encoding Initiatives (TEI) *Guidelines for Electronic Text Encoding and Interchange* had been developed since 1987 as an international cooperative project under the joint sponsorship of the Association for Computers and the Humanities, the Association for Computational Linguistics, and the Association for Literary and Linguistic Computing. By the late 1990s, it was felt that a more stable organizational structure was needed for the further development and long-term management of the Guidelines, which by that time had gained wide recognition as the foremost standard for the encoding of scholarly relevant textual resources in virtually every field of the Humanities. Institutions with a strong investment in TEI were taking the necessary legal steps and formed the TEI Consortium, which formally started its work in 2001 with the first Members Meeting in Pisa[1] At this meeting, the first members of the TEI Technical Council had been elected by the membership and took office at the beginning of 2002 with the first meeting of the Council held in January of 2002 in London. The agenda for the work laying ahead was set here in a 2-fold way:

- A maintenance release of the current version of the Guidelines (P3),[2] updated to enable users to work with XML tools, but still compatible with SGML was to be released as soon as possible.
- Work on a complete overhaul of the Guidelines, to bring it in line with technical and theoretcial advances influencing the encoding of text, with a license to break compatibility with the existing schema was to begin immediately.

The first of these items was conducted mainly be the TEI Editors Lou Burnard and Syd Bauman. It did not require too much attention from the Council and was completed in June 2002 with the publication of P4. Thus right from the start, the TEI Council was about to focus on the new development that needed to be conducted. From among the work items on the table, two workgroups were charged to

Literary and Linguistic Computing, Vol. 24, No. 3, 2009.
doi:10.1093/llc/fqp017 Advance Access Published on 29 May 2009

Table 1 An overview of all workgroups that contributed to the development of P5

TEI Workgroup on character encoding	2001–05	Christian Wittern
TEI Standoff Markup Workgroup	2002–05	David Durand
TEI Manuscript Description Task Force	2002–05	Matthew Driscoll
TEI META Task Force	2003–05	Sebastian Rahtz
TEI: Names and Places Activity	2006–07	Matthew Driscoll

start work immediately with the limited resources available, others were charged later as the necessity arose and resources allowed. Table 1 gives an overview of all work groups that contributed to the development of P5, together with their period of activity and the chair responsible to oversee the work.[3]

The 6 years of development of P5 can be roughly divided into two periods: In the first few years most of the work was done in the work groups charged by the Council. However, some of the plumbing involved a redesign of much of the underlying ontological structure of the TEI scheme, which made it increasingly necessary for the Council members to be more directly involved in the development process. Thus for about the last 2 years the Council itself became its most active work group. This article will be roughly structured along these lines; the work done in development for P5 will be first covered in terms of work done by the workgroups; followed by the work done mainly within the Council and some *ad hoc* groups mostly composed of Council members. Of these, the Personography task force (contributed by Arianna Ciula) and an *ad hoc* task force for Facsimile encoding (contributed by Conal Tuohy), will be described in some detail below. The work of the manuscript taskforce has already been described by Matthew Driscoll (2006); this will not be taken up again here.

2 TEI Workgroup on Character Encoding

The encoding of characters is a prerequisite for text encoding. In the SGML world, which was the underlying technology for the TEI until the advent of P5, no assumptions could be made about the character encoding used for a given document. Instead, information about the character encoding was part of the SGML prolog, which is used to inform the SGML processing system about some basic parameters of the document to follow. As a result, the TEI needed a mechanism to specify which script was used in a document and also add information about the language of the document and what binary values were used to encode which characters. This information was encoded in a separate, auxiliary document type, the Writing System Declaration (WSD; Chapter 25 of P3 and P4), while a short discussion of Characters and Character Sets was included as Chapter 4.

The task of this workgroup was to update these two chapters in light of the changes required by the switch to XML and other recent developments.[4] One of the major changes thus induced was that XML allows only Unicode (or a subset thereof) to be defined as a document encoding. This made the life for text encoders both easier and more difficult. Easier, because for most purposes the characters needed are now readily available for encoding in texts, more difficult, since many text encoders had so far comfortably lived in a world encodable with 127 ASCII characters and suddenly found they needed to get a grip on the 100,000 characters now at their disposal. But the real difficulty arises for text encoders in need of representation for characters that are not included in even this comprehensive list of encoded characters. The task of the workgroup therefore was also to find a way around this new limitation.

The first meeting was held in Berkeley in October of 2001: this was a joint meeting of members of the WG with invited guests from the Unicode Consortium (Ken Whistler and Rick McGowan), Peter Constable from the Summer Institute of Linguistics (SIL), as well as Richard Cook and Howie Lan from UC Berkeley. The agenda included

a wide range of items from 'Encoding Differences at the Glyph Level' to 'The future of the WSD', 'Glyph Registrar', 'Phonetic Glosses', and the usage of the language attribute in TEI.

One of the major areas of disagreement in the discussions at the Berkeley meeting and in the following deliberations of the working group was the status of Unicode in respect to TEI. Specifically, if the markup in TEI documents could affect the meaning of Unicode characters. Since the semantics of the characters are defined by the Unicode standard, there was a strong opinion that this could not be affected in any way by markup constructs. On the other hand, since Unicode only defines an abstract character that could be represented by a broad range of glyphs, it was argued, markup constructs could meaningfully give more precision to the definition of characters by selecting only a subset of the possible glyphs, or indeed provide a graphic representation of a character encoded in the document without trespassing on Unicode territory.

As a result of these discussions, the recommendations provided by the working group as included in P5 provide a largely parallel structure for the description of characters (<char>) and glyphs (<glyph>), both wrapped into a <charDesc> element. The former is to be used when referring to a character, which is not seen as a glyph variant of an existing character, but rather a character that has not been defined hitherto. The latter is to be used when some additional annotation for a character that does exist in Unicode is required, for example in a dictionary entry where two different shapes of a character are discussed and it is essential that they are shown in a specific way. In spite of the name of the element, it should also be used in cases where additional information about an existing character, for example its pronunciation, sorting behaviour, or other properties have to be specified.

Another important outcome of this workgroup was the 'war on attributes'. At some point in the discussions it was realized that the constructs suggested for describing characters or annotating characters would break down for textual content that resides in attributes, since markup constructs are not allowed as content of attributes. Furthermore, and even more severely, other important descriptive means, for example the ability to express the language used for a chunk of textual information, is also not available for text that is given as an attribute value. These considerations led the workgroup to request that the TEI Council seek measures to overcome this problem. After some initial reservations, the Council decided to adhere to this and the reconsideration of all attribute values in the TEI encoding scheme was taken up as part of the META activity (see below *4.2 Datatypes*).

3 Standoff Markup Activity

Standoff Markup is another area where much development occurred with the advent of XML and related recommendations and specifications by the World Wide Web Consortium (W3C). The term 'Standoff Markup' served within the TEI Council as a rather loose denomination for everything to do with the need to point into some other part of the current or a different document, including out-of-line markup (that is markup that is living outside the document, but using some kind of pointing facility to address locations or spans within the document) via extended pointers and linking. In addition to looking into how to bring the TEI recommendations in line with recent technical and theoretical developments, the group was also charged with looking into particular practices of TEI users in specific areas, like applied linguistics representing multiple levels or layers of analysis of a corpus or attempting integrated analyses of multimodal communication (speech, writing, gesture, movement etc.), historians creating and archiving multimodal resources (oral history, video, transcripts), philologists recording multiple interpretations of a single written text, translators working with multilingual corpora, and investigate what standards would be applicable and recommendable in these specific areas.[5]

The TEI had already developed a fairly sophisticated pointing facility, the TEI Extended Pointer syntax, described in TEI P3.[6] This was taken up, together with concepts from the HyTime specification for SGML, for the XLink and XPointer specification developed by the W3C and the TEI now in

turn needed to integrate that new specification into its Guidelines, together with XPath for simple adressing into a document. As one of the results of the work done in this workgroup, some constructs for TEI pointers have been re-expressed using the XPointer facilities and registered as a XPointer scheme with the W3C,[7] thus giving a good example of how development efforts between the TEI and the larger markup communities represented in the W3C have mutually inspired and reenforced each other. All this resulted in a fairly major overhaul of the text that used to be in Chapter 14 of P3 and P4.

The full adoption of the recommendations of this workgroup also required other radical changes to the TEI schema. For one thing, where P3 and P4 make extensive use of the linking mechanism built into SGML (and still available in XML) between an element given a specific value on an attribute of type ID and another element linking to that element by virtue of having an attribute of type IDREF with the same value; P5 has almost completely abandoned that mechanism in favour of using URI (Universal Resource Identifiers[8]). A major drawback of this decision is that validation of such links is not provided for free by the XML parser anymore: instead it has to be built into a customized validation process for example using XSLT. On the other hand, the old mechanism was only available within a single XML document and could not be applied equally to document collections. In the light of the advance of XML databases and large document collections, it was considered an issue that would diminish in importance for future applications, while the advantages of using the now well understood linking mechanism of the World Wide Web, and many standards built on this, seemed to be more important.

Some of the discussions of this work group led to elaborate drafts, for example Corpus Applications[9] (which have not been merged into the text of P5, but do contain valuable examples of how to represent corpus annotations, encode parallel text, and point into video streams, among others. While this is a draft paper with some sections not fully fleshed out, text encoders interested in these topics might still benefit from looking at the examples and the accompanying discussions.

4 META Task Force

In early 2003, it became clear that a complete rewrite of the way the TEI Guidelines had been formally expressed and that a major revision to the class system was necessary to achieve the goals of improving extensibility and customization. A new activity was started to oversee this, with members recruited from the Council and additional invited technical experts. The task this group was charged with was as follows:

- The ODD ('One Document Does it all', the format used internally to derive the TEI DTDs) format should be revised to be entirely independent of the SGML/XML notation for DTDs.
- The ODD format and the current TEI DTD for tag documentation (TSD) should be combined into a single standard TEI tagset useable as a pizza topping.
- The new notation should be based on one of the XML schema languages.
- Additional processors should be created to make not only XML DTDs (and possibly SGML as well), but also at least one XML schema format.
- Where possible, data types should be converted to use the datatype library of the W3C.
- The Pizza Chef should be rewritten to allow user choice of DTD or schema output.
- The output should be a new version of the TEI Guidelines source and associated tools.[10]

In short, this meant that not only the underlying language used for expressing the TEI schema in machine readable form, but also the units of organization for the Guidelines and its internal structure had to be revised—all this without too much changes to the user-visible part of how texts are represented.

4.1 ODD and Modularization

Early on, the group decided to use RelaxNG fragments to express the TEI schema at the lowest level, since this was not only the language with the richest expressive capability, but it also had the necessary

tools to derive the other formats as required. At the same time, Sebastian Rahtz developed also roma, the first ODD processor.[11] Instead of the pizza model[12] of core, base and toppings that had been used in earlier TEI versions, the new TEI schema is now organized in a set of modules that can be (mostly) independently chosen in addition to the core modules.

In order to organize the more than 500 elements used in the TEI schema into a comprehensible system, but also to group them together by similarity of semantics, structure, or location and make modification and customization possible, most TEI elements are defined through classes rather than directly. Attributes commonly used together are also grouped into classes. Classes can contain other classes. Elements inherit the properties of the classes of which they are members.

4.2 Datatypes

Another task of this group was to rationalize datatypes of attributes. In previous versions of the TEI, most attributes had simply been declared as having CDATA content, even for cases like dates, where the prose of the Guidelines did in fact require specific restricted values. This made it practically impossible to verify with the tools available for processing of SGML text that a given instance did indeed conform to the intended value. Since XML has a rich set of datatypes that can be used to restrict values of attributes (or elements) according to predefined values, it was natural to extend these definitions to the TEI schema.

Early on the work group looked at all attribute values and tried to identify those that could possibly hold free text content, for example the <sic> element, which can in P4 contain the corr attribute, which holds a representation of the element content of <sic> as it appeared in the text. For example:

```
for his nose was as sharp as a pen, and
<sic corr="a' babbled">a Table</sic>
of green fields.
```

This allows arbitrary textual content for the attribute, something that P5 was trying to avoid as much as possible. The equivalent expression in P5 looks like this:

```
for his nose was as sharp as a pen, and
  <choice>
    <sic>a Table</sic>
    <corr>a'babbld</corr>
  </choice>
of green fields.
```

This has the added advantage of not prioritizing one representation over the other: the need for correction is identified and a correction supplied, but allowing the encoder to simply record the observed facts.

Underlying this modification, however, there is a fundamental change in the relationship between textual content and markup that has caused considerable debates in the TEI Council. Up to and including the text model used in P4 of the Guidelines one could argue that the characters that appear on the page of a printed book would be what becomes the textual content of elements in an encoded text, whereas the markup itself, confined to the realm of angled brackets and carefully cordoned off from the text itself, could be seen as commenting, annotating, or otherwise speaking about this textual content. If one were to produce a text-only version, the instructions to produce this were as simple as taking away all of the markup contained in angled brackets which would thus restore the original pure text.[13] In P5, on the other hand, there is not such an easy distinction between textual content and the markup operating on it; the markup has to be interpreted, processed using the correct semantics to produce a version that shows the text only; this process would have to know, for example that it should take either the content of <sic> or of <corr>, but not both—which one is chosen will depend on the context in which the processing takes place and not on a naïve view of the markup of a text. In a way, this reflected a more mature and sophisticated view of the relationship between markup and text, which benefitted from the years of experience with text encoding. Consequently the TEI Council did not feel the need to preserve a now outgrown distinction that had in fact been flawed from the beginning.

Most datatypes in the TEI schema are defined using the datatypes provide by the W3C XML Schema recommendation.[14] In the same way as the model classes, they are defined with some indirection, to allow users to extend or otherwise modify the range of allowed values permitted through these datatypes. For the datatypes for temporal values and durations, however, the W3C has been 'inspired' by those defined in ISO 8601, without completely following it.[15] More precisely, not all datatypes covered by ISO 8601 are present in the W3C standard and not all lexical representations are included as datatypes. The ISO standard, for example, allows specifying dates and durations with a precision that can be dependent on the requirements of a user by omitting some digits to the right, while the W3C datatypes require in most cases conformance to a stricter precision. For example, using the W3C datatypes, one would have to say:

```
He arrived around <time when="12:00:00">noon</time>
```

whereas a more accurate encoding that does not introduce unwarranted precision would be permitted using the ISO datatypes:

```
He arrived around <time when-iso="12">noon</time>
```

The advantages of using the datatypes of the W3C (which are well supported in existing validation tools) were weighed against the ISO datatypes' greater freedom and aptness for the encoding of historical documents. As what would be of use to a particular project would likely differ depending on the goals of the project, the Council decided to allow two different sets of attributes for the encoding of temporal information.

5 Persons and Places

The chapter now entitled 'Names, Dates, People, and Places' gives recommendations on the encoding of names and of data about names. While there was a similar chapter 'Names and Dates' among the additional tagsets of P3 and P4, the material contained therein has been considerably revised and extended. Aimed at the application of text encoding to a range of fields that can span from history to geography, from onomastics and toponomastics to biography and prosopography, this chapter also deals with the definition and standard expression of temporal dimensions.

5.1 Main extensions and common model

Between January 2006 and May 2007, the TEI Council chartered a small workgroup on the encoding of names and places which formulated new material that eventually formed part of this chapter. The chair of this group, Matthew Driscoll, reported on the rationale and the initial stage of this development—which at the time already included a solid model for the encoding of person names and data about persons—during the TEI day in Kyoto in 2006.[16]

In the final phase of the development of P5, the chapter has been extended further to allow for:

- refined encoding of text where there are references to persons and places;
- integration with and creation of data structures related to persons and places (e.g. authority files, biographical and prosopographical structures, gazetteers);
- representation of canonical information about names (of specific interest for onomastic and toponomastic studies);
- a coherent model across diverse data.

The first point is directly related to the last one. Indeed, once the model adopted for persons had been enlarged and tested, its scheme—with appropriate variations—was extended to places and to the representation of organizations (such as businesses or institutions, racial or ethnic groupings, or political factions).

To give an example: information about people, places, and organizations, of whatever type, is understood to comprise a series of statements or assertions that are considered part of one of these three main conceptual groups:

- characteristics or traits which do not, by and large, change over time;

- characteristics or states which hold true only at a specific time;
- events or incidents which may lead to a change of state or, less frequently, trait.

The data about named entities—persons, places, or organizations—can therefore be modelled following this tripartition.

A common model is also used to associate different instances of a named entity, both to group entities through the use of list elements (e.g. <listPerson>, <listPlace>) and to express explicit relationships between entities by introducing the <relation> element.

```
<listPerson>
<person xml:id="jsbach">
<persName>Johann Sebastian Bach</
persName>
</person>
<person xml:id="cdbach">
<persName>Catharina Dorothea</persName>
</person>
<person xml:id="ghbach">
<persName>Gottfried Heinrich</persName>
</person>
<listPerson>
<!-...->
<relationGrp type="children" subtype="first-
marriage">
<relation name="parent" active="#jsbach"
passive="#cdbach"/>
<!-...->
</relationGrp>
<relationGrp type="children" subty-
pe="second-marriage">
<relation name="parent" active="#jsbach"
passive="#ghbach"/>
<!-...->
</relationGrp>
```

5.2 Data structures related to persons and places: potential for sharing data

The concepts of separation and association between names and corresponding referents, being persons or places or any other named entity, are at the core of this chapter.

Following the general principle adopted in P5 of using URIs to make explicit connections between textual components, the attribute ref (reference), member of the attribute class created for linking names and their referents (att.naming), can be used to refer to the formal definition of a named entity in the following way:

```
<name ref="#jsbach" type="person">Johann
Sebastian Bach</name> was a prolific German
composer...
<!-...->
<person xml:id="jsbach">
  <persName>
    <forename type="first">Johann
      </forename>
    <forename type="middle">Sebastian
      </forename>
    <surname>Bach</surname>
  </persName>
</person>
```

This application of so called stand-off markup, as mentioned before, is particularly useful not only when the objective of a specific encoding strategy is to study individuals and places mentioned in a certain text, but, in general, when the corpus of texts in question is large enough to justify the creation of separate structures to contain data about entities only mentioned—eventually with different referential strings—in the core texts.

This seemed to be a suitable response to a concrete exigence of an expanding TEI community. Indeed, as the practice of TEI encoding spreads and the amount of text encoded in TEI XML becomes more substantial, when different texts refer to or make assertions about the same set of entities, the integration, and sharing of data among projects could be achieved in practice (e.g. in the case of places, through shared gazetteers encoded in TEI).

The occurrences of names in the text can also point to separate dedicated structures not necessarily maintained in TEI (e.g. biographical database and ontologies with data stored in a different format than TEI XML); but, thanks to the refinement of the TEI encoding model, the mapping and

integration between these external structure and the TEI texts is facilitated.[17]

5.3 Interdisciplinarity and interpretative focus

The TEI data structures for places, persons, and organizations have been devised in such a way to take into consideration issues that lie at the heart of the humanities, such us the historical dimension of names and their correspondent referents.

Indeed, the definition of places, for example, can be modeled at the intersection between history and geography, taking into account issues such as the standard expression of physical locations [see for instance the introduction of the element <geo> to express latitude and longitude with the recommended *datum* World Geodetic System, (WGS84)] and the encoding of geographical features as well as the historical and cultural interpretation related to the existence of a particular place.

In this respect, responsibility and uncertainty about the sources that provide assertions about named entities can be expressed by using the members of the attribute class att.editLike, such as resp.

```
<org type="tribe" resp="#herodotus">
  <orgName>The Maxyans</orgName>
  <country>Libya</country>
  <desc>According to Herodotus, they were a
    west Libyan tribe who said that they were
    descended from the men of Troy.</desc>
</org>
```

5.4 Refined expression of date and time

In this context, the refined expression of date and time is of particular importance.

Thanks to the mapping to W3C and ISO date formats (this is discussed in more detail above, see *4.2 Datatypes*), automatic processing and validation of expression of dates and times are now supported. The attribute class att.datable.w3c provides attributes for normalization of elements that contain dated or datable events using the W3C datatypes. Time periods and relative chronology can also be defined.

```
<place xml:id="leipzig-univ">
  <placeName>University of Leipzig
    </placeName>
  <!-...->
  <event type="opening" notBefore=
    "1409-09-09">
    <desc>The <foreign xml:lang="la">Alma
      mater Lipsiensis</foreign> opened in
      1409, after it had been officially endorsed
      by Pope Alexander V in his Bull of
      Acknowledgment (on September 9 of
      that year).</desc>
  </event>
</place>
```

5.5 Definition for a canonical name or name part

To allow for the representation of canonical information about names as mentioned above, the occurrence of a named entity could also point to the formal definition of the name itself, in which case the attribute nymRef (reference to the canonical name or <nym>) can be used.

```
<placeName nymRef="#leipz">Leipzig
  </placeName>
<!-...->
<nym xml:id="lipsk">
  <form>Lipsk</form>
  <etym>From <lang>Slavic</lang>; it means
  <gloss>settlement where the lime trees
  stand</gloss>.</etym>
</nym>
```

This chapter leverages the legacy of previous proposals by expanding and extending the encoding models that have proved to be successful, taking on board, at the same time, the challenge of supporting the new needs of the TEI community, such as the creation of common authority files of named entities.

Moreover, following the approach adopted in the rest of the guidelines, this chapter gives firm recommendations on the encoding of names, dates, and data on persons, places and organizations, leaving room, at the same time, to alternative strategies (such as for instance the possibility of expressing a geopolitical structure either with hierarchical nesting of elements or with the use of explicit <relation> elements) that could suit different project-specific aims.

6 Facsimile Module

The development of the TEI facsimile module was a conscious effort to synthesize a number of existing encoding techniques which had been independently developed within and outside of the TEI community. These different encoding practices had all sprung from particular cases with their own requirements and none were sufficiently general to deal with other cases. Given the increasing number of projects making their own *ad hoc* customizations for facsimiles, it was high time that a standard was developed and promulgated, both to reduce the amount of repeated customization work, and to ensure interoperability between projects. The development of this module has been led by Conal Tuohy and Dot Porter, with active input from other members of the Council.

The purpose of the facsimile module is to allow encoders to record images of pages of a text, to identify arbitrary rectangular regions on those pages, to annotate those regions, and to link them with corresponding sections of the transcript.

6.1 Use of non-TEI markup

Some TEI projects have resorted to other markup languages to encode facsimiles, primarily METS (Metadata Encoding and Transmission Standard),[18] and SVG (Scalable Vector Graphics).[19]

METS documents can be used as a form of stand-off markup, containing references into a TEI document and into a set of images. SVG has been used within TEI documents, by customizing the TEI schema to include SVG elements. Within the SVG markup, interesting areas of an image can be defined as <rect> or <view> elements and these can be linked to TEI elements using the standard linking mechanisms provided in TEI.

These approaches can work well for individual projects: but both METS and SVG provide far more expressive power than is actually needed for a TEI facsimile edition. For many projects these approaches would be overkill; this is especially true of METS as it is necessary to record facsimile information in a separate file from the TEI. The overly expressive nature of SVG and METS also provides too many different ways that a facsimile might be encoded. It was felt that a much simpler mechanism within the TEI markup language itself would better serve to coordinate encoding practices and promote interoperability within the TEI community.

6.2 Conceptual model

At first glance the task of facsimile encoding seems rather simple, but there are a number of complex cases:

- Some projects have just one image per page. Other projects need to encode several different images of the same page taken using different photographic techniques, e.g. in ultraviolet light in order to reveal details not normally apparent. Similarly, some projects need to encode high-resolution images of certain particularly interesting details of a page. So for a given page there may be a set of images which may have different resolutions and which may overlap in arbitrary ways.
- Some projects will include images and transcript, whereas other projects may include only images without a transcript at all.
- Some projects will align entire images with transcripts page by page, whereas other projects will align images of individual words with transcripts of those words.
- Some TEI documents (variorum editions) are transcriptions of several different editions of a single text. In these documents, a given section of transcripts could be aligned with multiple images, taken from the different editions.

Clearly, to adequately support all of these requirements, the facsimile schema would have to be fairly complex. To ensure that the quite common simpler cases could also be encoded without undue complexity, the schema would also need to include syntactic shortcuts. For instance, one fairly common practice which is supported by the new facsimile module has been to customise the "<pb>" (page break) element to add an attribute whose value is a pointer to an image file.

The facsimile module therefore defines:

- a conceptual model of text bearing objects (for example pages of a book), rectangular areas on those objects, image files, and how they all align;
- a verbose syntax which expresses all those concepts explicitly, and;
- shortcuts in which certain relationships are left implicit.

6.3 Facsimiles, surfaces, and zones

6.3.1 *Facsimile elements*

Facsimile information is encoded within a new top-level <facsimile> element which can appear after a <teiHeader>. The element can used instead of, or in conjunction with, a <text> containing a transcript of the source document: hence a TEI document may now consist of just bibliographic metadata and a set of images, without a transcript. This might be useful for documents which are purely graphical, without any text at all, but more importantly it allows for a workflow in which a purely image-based TEI document is produced first, and a transcription added later. This marks a departure from the model of TEI P4 and earlier, where a text consisted only of transcribed characters; this new model allows for either both representations to stand side by side or for one of them (<text> or <facsimile>) to represent the whole text.

6.3.2 *Surfaces*

The <surface> element represents a page, or any other inscribed surface.

An encoder can choose to encode a double-page spread as either a single <surface> or as two distinct <surface> elements each representing a single page.

A <surface> element contains information about any rectangular regions of interest on that surface, and any graphics which depict it. A <surface> may bear attributes, which align it with particular areas of interest on that surface.

As a shortcut, in place of <surface> elements, a <facsimile> element may contain a sequence of <graphic> elements, each of which is assumed to depict an implicit <surface>. This shortcut syntax may be convenient for encoding simple cases, but it is not adequate for recording more than one image of each page, or recording where the margins of the actual page appear in the scanned image.

6.3.3 *Zones*

A <zone> element is used within a <surface> to identify a rectangular region of the surface.

The region which the <zone> covers is given by the bounding box attributes, and is expressed in the same coordinate space as that used by the parent <surface> element. If a <zone> and its parent <surface> have the same bounding box values, then the <zone> corresponds to the entire area described by the surface.

The rectangular areas identified by <zone> elements can serve two purposes.

- A zone can be used for analytical purposes (such as to represent the area in which a particular word appears), in which case the <zone> can be linked to the transcription of the word.
- A zone can also be used as a container for one or more <graphic> elements, in which case the <graphic> elements are all considered to exactly cover the area specified by the <zone> element.

As a shortcut, a <graphic> may be contained directly within a <surface>, without specifying a <zone>. In such a case the <graphic> is assumed to exactly depict the entire <surface> without any margin.

A <zone> may also contain a gloss or description which may be useful for certain types of image annotation projects

6.4 Coordinate spaces

The alignment of images, pages, and transcriptions presupposes that they share a common coordinate space in which their location can be expressed. The issue arose in deciding the units of this coordinate

```
<TEI>
<!-- TEI header omitted -->
  <facsimile>
  <!-- coordinate space of the whole photograph -->
    <surface ulx="0" uly="0" lrx="1024" lry="684">
      <desc>A photograph of the left hand side of the beginning of the
        typoscript scroll for <title>On the road</title> by Jack Kerouac,
        photographed by Thomas Hawk.</desc>
      <graphic url="ontheroad1.jpg"/>
      <zone xml:id="firstline" lrx="137" lry="78" ulx="1002" uly="115">
      <desc>Beginning of the first line of text</desc>
      </zone>
      <zone xml:id="secondline" lrx="136" lry="138" ulx="1010" uly="161">
         <desc>Beginning of the second line</desc>
      </zone>
    </surface>
  </facsimile>
  <text>
    <body>
      <p><seg facs="#firstline">I first <sic>met</sic> met Neal not long
                  </seg><gap/>
      <seg facs="#secondline">serious illness that I won't b</seg></p>
    </body>
  </text>
</TEI>
```

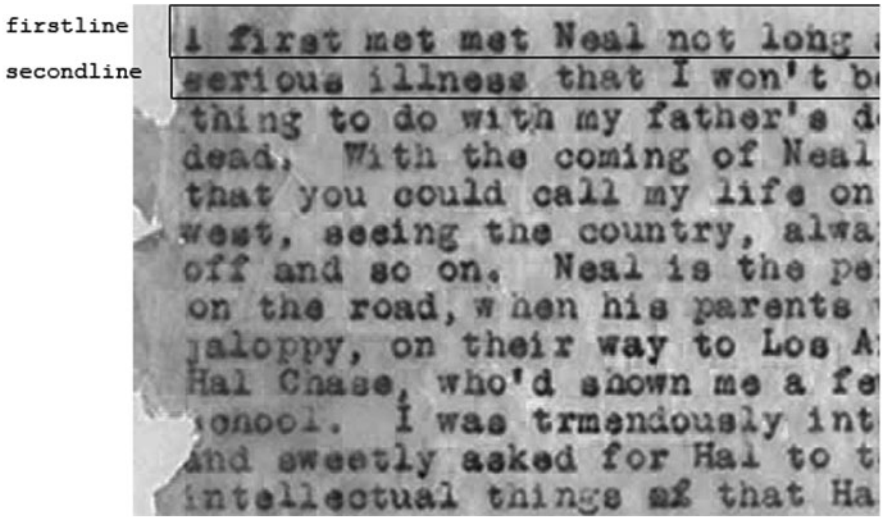

Fig. 1 Extended example of using the facsimile module to align a digital facsimile with a transcription of a text. Photo copyright by emdot, released under a Creative Commons licence at www.flickr.com/photos/emdot/4876145/.

space should be. Should the areas be expressed in physical units such as millimeters? Or might encoders prefer to take measurements by counting pixels in their scanned images? Because either option might make sense in some circumstances, it was decided to leave the units of the coordinate space undefined. In order to function, all that's required is that the units are used consistently within a given <surface>. For example, if a page is specified as 300 U high, then a graphic covering the top half of the page would be 150 U high, whether the measurements were taken in pixels or in millimeters.

One draft of the schema allowed for <zone> elements to recurse, and for the coordinate space of each <zone> to be treated as a transformation of the coordinate space of its parent. This would have allowed for a <zone> to be used to zoom in on part of a <surface> and define a more fine-grained unit of measurement for that part of the <surface>. This feature was considered too complex and of insufficient utility and was therefore abandoned.

6.5 Aligning transcription and facsimile

An encoded facsimile, with surfaces and zones, may then be linked to the transcription.

The facsimile module introduces a new global attribute facs, which allows any textual element to point to the facsimile element to which it corresponds. In its most articulated form, the referenced element would be a <zone> representing an area on a particular page. In a shorthand form, the facs attribute could point to a <surface>, which would imply that the element bearing the facs attribute corresponds to the entire page described by the surface.

By convention, when a (page break) element has a facs attribute, the referenced facsimile is understood to be an image of the entire page which begins at that point.

6.6 Encoding references to regions

The facsimile markup required a way to refer to a particular rectangular area of a graphic. This is typically expressed by giving the coordinates of two opposite corners of a box (a so-called bounding box) which encloses the area. There were several possible ways in which the bounding boxes could be encoded. We first explored the idea of using URIs with fragment identifiers to point to parts of an image. This turned out to be unsatisfactory, and in the end we defined a class containing four attributes, ulx, uly, lrx, and lry, denoting the *x* and *y* coordinates of the upper left and lower right corners of the bounding box.

7 URI Fragment Identifiers

A URI referring to a document may optionally include a fragment identifier suffix. This is the portion of the URI following the # symbol. A fragment identifier on the end of a URI refines the scope of the URIs reference to just a part of the larger information resource which the URI refers to. A fragment id might refer to an element or set of elements in an XML document; a rectangle within an image; or a specific sequence of frames within a video file. Because different types of media can be fragmented in various ways, each media type is associated with a particular syntax for fragment identifiers.

In particular, two fragment identifier syntaxes were examined: The syntax defined for SVG documents, and that defined for MPEG documents.

7.1 SVG

The fragment identifier syntax for SVG includes the XPointer facilities available for XML documents in general. For example, a URI can directly refer to an element by id (the so-called 'bare name' shorthand syntax).

```
graphic.svg#footnote-1
```

Another way to refer to part of an SVG graphic is to use the svgView XPointer scheme in which an identifier can specify any of a variety of different paramaters, including a scale factor, the coordinates of a rectangular region, transformations such as rotation, shearing and stretching, and even identify a specific element within the SVG document to highlight.

```
graphic.svg#svgView(viewBox(0,0,200,200);view
Target(footnote-1))
```

7.2 MPEG-21 Part 17

In part 17 of the MPEG-21 standard,[20] the Motion Picture Experts Group has defined another fragment identifier syntax for referring to parts of multimedia resources. The syntax includes the facility to refer to a rectangular area, like so:

```
media.mp4#mp(region(rect(20,20,40,40)))
```

Unfortunately, at present the syntax is only applicable to resources in MPEG format (although the standard suggests that the syntax could be extended to cover other media types in future).

7.3 Problems with URI fragment identifiers

Although URI fragment identifiers had some attractive features (not least that they would have made a concise alternative to defining new XML markup), they were eventually rejected because of the limitation that each URI fragment identifier syntax applies only to a specific set of media types. This limitation is part of the definition of URIs.

The main problem was that many popular image formats including TIFF, JPEG, PNG, etc, do not as yet have any defined fragment identifier syntax. The possibility also existed that a variety of different fragment identifier schemes might eventually come into use for different image formats, and this variability of syntax was felt to be a weakness; an encoding scheme that could be used consistently regardless of image format would be easier to use in the long run. This effectively ruled out the use of URI fragment identifiers, and implied that rectangular bounding boxes would have to be expressed in some kind of XML markup.

8 Attributes for Expressing Bounding Boxes

Both HTML and METS use a single attribute called coords to express a rectangular bounding box. The value of the coords contains four numbers, which expresses the x and y coordinates of opposite corners of the bounding box.

This approach was tried in the facsimile module, but was replaced with a more verbose syntax in which each of the four numbers is represented by a distinct attribute. The coordinates of the upper-left corner of the bounding box are given by the attributes ulx and uly, and the coordinates of the lower right corner are given by lrx and lry.

Although the coords syntax used in HTML and METS had the advantages of familiarity and also concision, the alternative syntax with individual attributes had the advantage that the single-valued numeric attributes are easier to process automatically, and particularly to validate with schemas.

9 Customization and Conformance

One of the important goals in revising the existing Guidelines and developing P5 had been the simplification of the customization process which was difficult to understand and tedious to implement in previous versions. The TEI Council accordingly spent considerable time in designing and implementing the new class-based modular infrastructure building blocks. In this new system, customization is not an afterthought but is required for virtually every use of the TEI. While the TEI continues to offer the widely used TEI Lite subset of frequently used elements, it is now clearly labelled as mostly an example of customization best practice and further adaption to the specific needs of a project is encouraged and made easy.

Having spent the time and effort to create this system, the Council next came to realize that there was some need to define how documents created based on the schemas compiled using this infrastructure could be said to conform (or not conform) to the TEI abstract model as laid out in the prose of the Guidelines. Without such a classification, it would be very difficult to exchange documents that claim to be valid TEI documents and the purpose of creating a standard like the TEI Guidelines would be defeated. This objective goes back to the first of the Poughkeepsie Principles: to 'provide a standard format of data interchange in humanties research'.[21] On the other hand, one might ask, 'Does every serious application of the

TEI for the scholarly encoding of pre-existing documents require customization?' So the challenge was to find ways to encourage experimentation with new features that are not currently accounted for in the TEI schema, but conceivably may be added at some point in the future (at which point it would be good to be able to draw on such experience) but at the same time ensure that the TEI schema itself was stable enough to be used as a format for blind interchange.

As a result of these discussions, documents created according to the TEI schema are now distinguished along the following criteria:

- TEI recommended practice: Applicable to either conformant or conformable documents that follow recommended practice where the Guidelines recognize several ways to encode a feature, but one is marked as preferable.
- TEI Conformant: Criteria for conformance are listed in 23.3. Conformance of the Guidelines; they include validation against a schema derived from the Guidelines, conformance to the TEI abstract model, correct use of the TEI namespace and proper documentation of the schema through an ODD file.
- TEI Conformable: A document that can be transformed algorithmically to a conformant document by a supplied programme.
- TEI Extension: A document that is valid against a schema that has been derived from an ODD file with proper documentation, but that contains additional distinctions representing concepts not present in the TEI abstract model.

While these distinctions are vaguely ordered in descending order of compliance within the letter of the Guidelines, this is a purely formal statement that should not be seen to imply a judgement about the scholarly content of a given text. A given document should, however, declare its membership of one of these categories and adhere to it.

Maybe the single most contentious of these rules was the requirement to 'properly use the TEI namespace', which in fact requires that any extension that changes a content model (for example of a to allow non-tesselation, or to allow <head> at other places than the beginning) will have to place this changed document into its own namespace so that it becomes <my:div>. This has, of course, rather severe consequences for the processing of such documents.

10 Looking Ahead

The development of TEI P5 took much longer than probably most of those involved expected. Since early 2005, the development sources were kept on the open source portal Sourceforge <http://sourceforge.net/projects/tei/> for all interested users to peruse; there have also been frequent development releases. This removed some of the pressure to ship a product, and provided better ways for the community to participate in development efforts since those interested in experimenting with new features could do so using the development versions and give immediate feedback. Since there were occasionally incompatible changes, inevitably users started to ask on the TEI list 'what happened to (attribute/element X)?' or 'Element/attribute X disappeared, what am I supposed to do now?'. Some of these changes were intentional and the explanation was usually given within a short time, others were bugs that were duly fixed in subsequent versions. On the other hand, larger projects could not afford to base efforts of many man-years on a moving target that could change shape at the next release.

At the TEI members meeting 2006 in Victoria, the TEI Council therefore announced its commitment to ship the 1.0 version of P5 in time for the meeting of the following year. At the same time, principles for the further development had been laid out.[22] Frantic months of work followed that announcement and all existing work items had to be scrutinized and pruned for those ready for inclusion in P5 1.0 and another growing list of feature requests and items that had to be scheduled for P5 1.1. or later. With the Council members spread over a time difference of 19 h from Wellington (New Zealand) to Lethbridge (Canada), the sun literally did not set on working council members and everybody had to become accustomed to awake to a new pile of unread messages in the Council folder. This inevitably lead to compromises. Moreover, some of

the items could not be completed; a planned more thorough revision of how text critical variants are represented, for example, had to be postponed and the revision limited to the elements for editorial invention, which needed clarification.

Development of P5 will continue and oversights and bugs will have to be addressed. It is to be hoped that the effort that went into readying the schema and the documentation for publication was well spent and will continue to provide a major reference point to all users of the TEI schema. This will also make it possible for ongoing translation efforts to continue without the fear that a chapter translated today will be completely changed tomorrow and thus enable the TEI community to grow beyond the current strongholds in Europe and North America.

References

Biron, P. V. and Malhotra, A. (eds) (2004). *W3C XML Schema Part 2: Datatypes Second Edition*. World Wide Web Consortium. http://www.w3.org/TR/xmlschema-2/ (accessed 10 March 2009).

Burnard, L. (2000). Text Encoding for Interchange: a new Consortium. *Ariadne*, 24 < http://www.ariadne.ac.uk/issue24/tei/intro.html> (accessed 10 March 2009).

Cummings, J. (2007). The text encoding initiative and the study of literature. In Schreibman, S. and Siemens, R. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell, pp. 451–76. http://www.digitalhumanities.org/companion/DLS/ (accessed 10 March 2009).

Driscoll, M. J. (2006a). P5-MS: A general purpose tagset for manuscript description. *Digital Medievalist*, **2**(1). http://www.digitalmedievalist.org/journal/2.1/driscoll/ (accessed 10 March 2009).

Driscoll, M. J. (2006b). XML markup of biographical and prosopographical data. In Wittern, C. (ed.), *TEI Day in Kyoto 2006*. Kyoto: Institute for Research in Humanities, Kyoto University, pp. 75–83. Available from http://coe21.zinbun.kyoto-u.ac.jp/tei-day/TEIDay Kyoto2006.pdf (accessed 10 March 2009).

Eide, Ø. and Ore, C.-E. (2007). Mapping from TEI to CIDOC-CRM: will the new TEI elements make any difference? Paper presented at the TEI members meeting 2007 at the University of Maryland in College Park, Maryland, USA.

Eide, Ø. and Ore, C.-E. (2008). TEI and cultural heritage ontologies. Paper presented at the Digital Humanities conference 2008 at the University of Oulu, Finland.

Sperberg-McQueen, C.M. and Burnard, L. (eds) (1994). *Guidelines for Electronic Text Encoding and Interchange. (TEI P3)*. Chicago and Oxford: Text Encoding Initiative.

Sperberg-McQueen, C.M. and Burnard, L. (eds) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange. XML Version*. Oxford, Providence, Charlottesville, Bergen: Text Encoding Initiative Consortium.

TEI Consortium (eds). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

TEI Council (s.d.). TCW09: Backward Compatibility and the Maintenance of the Text Encoding Initiative Guidelines. http://www.tei-c.org/Activities/Council/Working/tcw09.xml (accessed 10 March 2009).

TEI EDP1. Design Principles for Text Encoding Guidelines. Version of 9 January 1990. http://www.tei-c.org/Vault/ED/edp01.gml (accessed 27 January 2008).

Vanhoutte, E. (2004). An introduction to the TEI and the TEI Consortium. *Literary and Linguistic Computing*, **19**(1): 9–16.

Notes

1 For more information about the background of the TEI see TEI.Consortium (eds), ''Historical Background.'' Guidelines for Electronic Text Encoding and Interchange. (Version 1.0. Last updated on 28 October 2007) http://www.tei-c.org/release/doc/tei-p5-doc/en/html/AB.html (accessed 1 May 2009) ABTEI and Vanhoutte (2004).

2 Releases of the Guidelines are referred to by their 'proposal' number; P3 had been the first version to be released in print, in April 1994.

3 In addition to the work groups mentioned in Table 1, there have been workgroups charged by the Council that did not provide input to P5, either because the charge was for a different type of work, or because the recommendations were not ready; these are the TEI Task Force on SGML to XML Migration (founded by a NEH grant), chaired by Christine Ruotolo and active from May 2002 to January 2005, and the TEI Physical Bibliography Work Group charged in May 2004 (this workgroup was first chaired by Terry

Catapano, later Murray McGillivray agreed to take over this task).

4 The original charge, as well as working papers and other material can be found at the WG's webpage http://www.tei-c.org/Activities/Workgroups/CE/ (accessed 1 May 2009).

5 The original charge, as well as working papers and other material can be found at the WG's webpage http://www.tei-c.org/Activities/Workgroups/SO/ (accessed 1 May 2009).

6 See *TEI P3*, Chapter 14.2 Extended Pointers, p. 405.

7 See the XPointer registry at http://www.w3.org/2005/04/xpointer-schemes/ (accessed 1 May 2009), specifically the schemes for left, match, range, right, and string range.

8 This is a more general term than URL (Universal Resource Locator) as used to address webpages, which also includes URN (Universal Resource Name) references like for example isbn for the ISBN numbers of books.

9 See http://www.tei-c.org/Activities/Workgroups/SO/sow05.xml (accessed 1 May 2009).

10 Adopted from the charge document at http://www.tei-c.org/Activities/Workgroups/META/ (accessed 1 May 2009).

11 It should be noted here that roma is the name of the commandline tool used for local processing of ODD files, whereas Roma is the web application that does pretty much the same thing, but also allows users to interactively generate and modify the customization files that serve as input to the ODD processor to drive the generation of TEI validators.

12 See Burnard 2000 on the meaning of the underlying metaphor and how this models the construction of TEI schemata.

13 It goes without saying that this is exaggerated to make the point; even in P4 there were traces of what would become much more common in P5; for example in the markup dealing with text-critically collating multiple versions of a text. In such cases the single stream of characters that constitutes the text could not be restored by uncritically removing the markup. On this point see also Cummings 2007, p. 467ff.

14 W3C XML Schema. http://www.w3.org/TR/xmlschema-2/ (accessed 1 May 2009).

15 For more information on this, please refer to the specification *Section D ISO 8601 Date and Time Formats* at http://www.w3.org/TR/xmlschema-2/#isoformats (accessed 1 May 2009).

16 See Driscoll (2006b).

17 See Eide and Ore (2007). For a proposal of how to facilitate the mapping even further see also Eide and Ore (2008).

18 Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/ (accessed 1 May 2009).

19 Scalable Vector Graphics. http://www.w3.org/Graphics/SVG/ (accessed 1 May 2009).

20 MPEG-21 Standard. http://www.chiariglione.org/mpeg/index.htm (accessed 1 May 2009).

21 See TEI EDP01.

22 See TEI Council (s.d.).