ECP 2006 DILI 510049

ENRICH

From MASTER to TEI P5

Deliverable number	*WP3-00.1.0*
Dissemination level	*Public*
Delivery date	*2008-03-03*
Status	*Draft*
Author(s)	*Lou Burnard, OUCS and James Cummings, OUCS*

e*Content*plus

This project is funded under the *e*Content*plus* programme, a multiannual Community programme to make digital content in Europe more accessible, useable and exploitable.

Document Version Control

Version	Date	Change Made	Initials
1	2008-02-31	Reformatted to list	SR
0	2008-02-30	Draft answers part	JC
	2008-02-19	First draft for discussion	

Document Review

Reviewer	Institution	Date and result of review
Lou Burnard, OUCS		Document approved

Approved by (signature)	Date

Acceped by at European Commission (signature)	Date

This working document summarizes differences between the DTD originally defined by MASTER, the version of that DTD currently used in the Manuscriptorium, and a DTD derived from the TEI P5 manuscript description module. The source code for these DTDs is readily available online; for convenience, local copies are kept here

The two systems are actually very close and many of the differences between them are systematic, resulting from design decisions taken during the production of TEI P5. There are also a few places where TEI P5 has reorganized and made more consistent structures which were a little ad hoc in MASTER. For a more detailed discussion of the evolution of the TEI P5 proposals, see an article by Matthew Driscoll published in *Digital Medievalist* (2.1, 2006) on this topic, which also describes and illustrates the major differences between P5 and MASTER, together with their rationale.

The main objectives of this document are firstly to facilitate production of a script for automatic conversion of existing MASTER records to TEI P5, and secondly to highlight a number of design questions which need to be resolved before the ENRICH P5 application is finalised. This document includes a summary of the decisions taken concerning these questions at the WP3 first meeting, in Copenhagen, 28 February 2007.

1 Content model changes

1. Many elements in MASTER had a content model of p+. In TEI P5, all such elements have a content model of macro.specialPara. The difference is that there is no need to introduce a `<p>` element to wrap the content of the element. For example, a MASTER record such as

```
<accMat><p>Pastedown on fols 2 and 4</p></accMat>
```

   could appear in TEI P5 simply as

```
<accMat>Pastedown on
fols 2 and 4</accMat>
```

   Elements with this content model are: `<accMat>`, `<acquisition>`, `<additions>`, `<collation>`, `<condition>`, `<custEvent>`, `<decoNote>`, `<foliation>`, `<layout>`, `<musicNotation>`, `<origin>`, `<provenance>`, `<source>` and `<surrogates. The element <watermarks> also had this content model>`, but has been removed (see below).

   Question: when converting from MASTER to P5, should the unnecessary `<p>` be retained or removed?

   Resolution: The ENRICH project decided that where multiple `<p>` elements exist, they should be kept, but where only a single `<p>` element exists in a TEI element where it is optional, then the wrapping `<p>` should be discarded unless it contains attributes recording important intellectual content.

2. Many elements in MASTER had a content model in which p+ was alternated with some other, more specific, elements. In most cases the alternation is exclusive -- that is, the element may contain either p+ elements or more specific elements but not a mixture; in a few cases, the alternation is inclusive -- that is, the element may contain a mixture of paragraphs and more specific elements. Elements which are exclusive in this sense are: `<adminInfo>`, `<custodialHist>`, `<decoDesc>`, `<history>`, `<msContents>`, `<physDesc>`, `<objectDesc>`, and `<layoutDesc>`. The only inclusive elements are `<bindingDesc>` and `<binding>`.

In TEI P5, the direct references to <p> within all such content models have been revised to reference the element classes model.pLike. This makes it possible in a TEI P5 schema to use either <ab> or <p>, if <ab> is available; or indeed to define other generic elements for use here.

Question: should <ab> be used?

Resolution: The ENRICH project did not express a preference for the use of <ab> instead of <p>. They recognised that there was semantic baggage with <p>, they felt it was an unnecessary change.

3. The following elements, which all had a content model of PCDATA in MASTER, in TEI P5 have a content model of macro.xText: <msName> (formerly <altName>), <collection>, <depth>, <height>, <institution>, <locus>, <origDate>, <origPlace>, <repository>, <width>. This change permits the inclusion of the <g> element used in TEI P5 to represent nonstandard characters or variant glyphs in otherwise straightforward text.

4. In MASTER, the <msHeading> element was used to provide a brief description or characterization of a manuscript. In P5, this has been replaced by a generic <head> element, which contain most -- though not all -- of the child elements previously allowed within <msHeading>: specifically, it does not permit <author>, <respStmt>, or <textLang>, although a very large number of phrase level elements are permitted. Driscoll points out that most of the components of <msHeading> are better provided elsewhere in the record, and that the heading could simply be untagged text.

   In the Manuscriptorium document on Technical Compatibility of Metadata, however, amongst other rules about minimal compatibility requirements, the recommendation is made that <msHeading> should include as a minimum <title>, <author>, <origdate>, <textLang> and (optionally) <note>. This is not enforced by any DTD however.

   Question: Should the schema enforce the Manuscriptorium recommendations for <head>? if so, should we rename the element?

   Resolution: It was felt by a few members of ENRICH that the need for an existing, possibly structured, brief description was beneficial even if this resulted in a duplication of information. The ENRICH project decided that <head> should not be renamed for the ENRICH schema, and could act as a secondary location to record summary/overview metadata, but that the proper places in the header will be used in preference by processing software. Thus, if a <msHeading> exists which contains unique information, this will be copied to the correct places, and transformed to a valid <head>, but it no such <msHeading> exists, then it will not be created. If information exists in the proper locations and a <head> then the proper locations should always be used when providing the information. The project accepts that <name type="author"> is a good enough replacement for <author> in such a field. However, the project wonders if <lang> could be used instead of the missing <textLang> that is unavailable at this point, or if <textLang> could be added to **model.pPart.msdesc** and thus be available in **macro.paraContent**

Other miscellaneous content model differences are listed below:

1. the <dimensions> element in TEI P5 may contain zero or one <height>, <width>, and <depth> (in that order) rather than any number of such elements in any order.

2. the element <accMat> becomes a sibling of <additional> rather than a child of it

3. the element <remarks> within <adminInfo> is replaced by a reference to the class model.notelike
4. the element <overview> is not permitted within <bindingDesc>: if one is provided, its contents will be converted to <p> elements.
5. the element <overview> is replaced by the element <summary> when it appears within <msContents>
6. the element <langUsage> is replaced by <textLang> when it appears within <msItem>
7. the element <q> is replaced by <quote> when it appears within <msItem>; <cit> (grouping a quote with an attribution) is also possible.
8. the element <msContents> may contain an optional <textlang> element
9. a new element <msItemStruct> may appear within <msContents> as an alternative to <msItem>
10. the <handDesc> (formerly <msWriting>) element now contains an exclusive alternation of model.pLike elements with <handNote> (formerly <handDesc>) elements, rather than an inclusive one.
11. The <support> has been given a simpler content model (macro.specialPara) and the optional specialist elements (<overview> and <watermarks>) it contained are no longer available within it.

2 Elements renamed

A small number of elements were renamed at TEI P5. The following table lists them.

P4 Name	P5 Name	comment
msDescription	msDesc	"desc" is used as a suffix throughout P5
altName	msName or altIdentifier	altName in Master is ambiguously used both for an alternate name and for an alternative identifier; in P5 the two usages are distinguished
decoration	decoDesc	For consistency
handDesc	handNote	For consistency
overview	summary	only within msContents
msWriting	handDesc	For consistency
msHeading	head	
idno (at the start of msPart)	altIdentifier	other uses for idno remain unchanged

3 Elements added

A number of new elements were added as a consequence of the restructuring of the MASTER <physDesc> element: specifically <objectDesc>, <supportDesc>, <sealDesc>, and <layoutDesc>. The existing <support>, <extent>, <foliation>, <collation> and <condition> elements, formerly direct children of <physDesc>, are now grouped within the new element <supportDesc>. The existing <layout> element is now wrapped within the new grouping element <layoutDesc>. These two elements now constitute the new <objectDesc> element, which is a child of <physDesc>,

along with <accMat>, <additions>, <bindingDesc>, <decoDesc> (formerly <decoration>), <handDesc>, <musicNotation>, and the new <sealDesc> element.

A new <altIdentifier> element replaces <idno> when this appears at the start of a <msPart> element.

A new <msItemStruct> element is available as an alternative to <msitem>: it contains the same child elements, but constrains their order and cardinality.

Question: should we constrain the order in which child elements of <msItem> are presented? should we use <msItemStruct> to enforce this?

Resolution: The ENRICH project decided that it was better to use <msItem> element and remove <msItemStruct> entirely from the ENRICH schema. It also felt that it should not constrain the order of child elements.

TEI P5 provides a small number of new elements which have no counterpart in MASTER:

- A new <stamp> element is available to record information transcribed from a stamp or similar device.
- A new <watermark> element is available to record information transcribed from (or descriptive of) a watermark or similar device.
- A new <filiation> element is available to discuss the manuscript's relationship with other manuscripts.

Question: if these new elements are included in the ENRICH DTD, how should we identify their content in existing elements?

Resolution: These elements should be included in the ENRICH schema for those creating new records, but the ENRICH project reached no consensus on how one should identify such information in existing content as no standard method to delineate it was followed in MASTER.

4 Elements removed

The MASTER elements <overview>, <paratext>, <remarks>, and <watermarks> are all removed from TEI P5. They should all be replaced by <p> elements.

The element <form> is removed from the manuscript module of TEI P5 (there is a TEI element of this name in the dictionary module). Its function is served by the form attribute of the new <objectDesc> element.

5 Attribute changes

For reasons elaborated in Driscoll's article, the attributes type, status, and dateAttrib, available for the elements <msDescription> and <msPart> in the MASTER DTD, are removed in the P5 specification. The type value in MASTER should be transferred to the form attribute on the new <objectDesc> element, if present. (It is however an open question whether other typologies might be useful in ENRICH). The values for both status and dateAttrib proposed in MASTER duplicate information provided elsewhere in the manuscript description and may therefore be ignored.

The following elements from the TEI P5 manuscript description module are typed, that is, they inherit a type and subtype attribute from the class att.typed. It would be useful to agree on which of these should have closed sets of values, and what those value lists should be: <accMat>, <altIdentifier>, <custEvent>, <decoNote>, <explicit>, <filiation>, <finalRubric>, <head>, <incipit>, <msName>, <quote>, <region>, <rubric>, <seal>, <settlement>, <stamp>

Similar considerations apply, of course, to many other elements which might be included from the core, transcription, and names and dates modules, notably <name> (for which MASTER already proposes a rather limited value list (other|org|place|person|female)).

The same applies to those elements in MASTER which have a locally defined type attribute, viz <dimensions>, <ptr>, <head>, <note>, <rubric>, <msName> (formerly <altName>)

Question: for which elements should closed value lists be defined? How should existing data be brought into line with those value lists?

Resolution: The ENRICH project felt that where closed or semi-closed value lists could be easily defined, they should be to encourage consistency, where such lists might prove extremely controversial or had very large number of values they should not be closed. Where we can find possible common values for a (semi-)closed list we should propose it and allow them to disagree if desired.

The targets attribute on <locus> is renamed as target in TEI P5. However, it is debatable whether this should be used, given the availability of the new facs attribute in TEI P5.

Question: should the global facs attribute be used more widely in ENRICH?

Resolution: The ENRICH project decided, where appropriate information is available in the legacy content, the new TEI facsimile elements and/or attributes should be used where possible.

At TEI P5 there are a number of changes in the globally available attributes. These changes affect all elements. Briefly:

- id becomes xml:id
- lang becomes xml:lang and must specify a valid language identifier (as further defined in P5)
- rendition is available as a more precise alternative to rend.
- TEIform is no longer used

There are also, of course, several changes in non-specifically manuscript elements. The attributes of <ptr> and <ref> are different, for example. The content model and attributes of <change> have changed. The 'mirror tags'<corr> and <sic> etc. have been replaced by <choice>.

The specialist attributes technique, quality, figurative on <decoNote> are not available in TEI P5. Non-default values for these might be used to develop a typology for the type attribute. (Manuscriptorium additionally has illustrative and size attributes for this element which will need to be handled in the same way).

The class attribute on <msItem> in MASTER could supply multiple pointers but in TEI P5 may supply only one value (though since this is defined as data.code, the change is probably invisible).

The langKey attribute on <textLang> is renamed mainLang. It and otherLangs now have values which have the same constraints as xml:lang.

All datable elements are now treated in the same way. In addition to the notAfter and notBefore attributes which the MASTER dtd provided, the attributes when (for an exact point date), from (for an exact point start date), and to (for an exact point end date) are now available. These attributes must all supply dates in a valid W3C format (ISO is also an option). An additional attribute period may be used to indicate a named period as an alternative. The following elements in the manuscript description module are members of this class: <acquisition>, <binding>, <custEvent>, <origDate>, <provenance>, <seal>, <settlement>, <stamp>.

It would be desirable to agree on a common way of normalizing all dates, choosing a subset of these attributes.

Resolution: The ENRICH project decided that where feasible all of the new TEI P5 dating attributes should be used, (i.e. when, from and to in addition to notBefore and notAfter and these should be correctly regularised to W3C format.

In MASTER, the elements <acquisition>, <custEvent>, <origDate>, <binding>, <origin> and <provenance> all included two additional attributes (certainty and evidence) to indicate the degree of reliability associated with a dating or event. In TEI P5 the equivalent attribute class (att.editLike)

provides attributes cert (equivalent to certainty), resp, evidence and source. Of the elements listed, only `<origDate>` and `<origin>` are members of this class in TEI P5.

Question: Are these indications of uncertainty needed for `<acquisition>`, `<custEvent>`, `<binding>`, or `<provenance>`? Which of the available attributes should be used?

Resolution: The ENRICH project feels that the new cert should be used for the old certainty and other TEI attributes used where possible. The ENRICH project would strongly suggest that the TEI include these attributes on the suggested elements.

At TEI P5, a new module provides for very detailed and consistent annotation of names of places and persons, and of the events and properties associated with them. Some of the attributes used in MASTER have changed as a consequence of this. In particular: on `<name>`, `<country>`, `<institution>`, etc. key is renamed as ref, and points to a `<person>` or `<place>` element as appropriate. A new key attribute is used to provide a coded value for such an entity (as an alternative). The full, reg and role attributes are no longer available on these elements. reg is not available on `<origPlace>` either.

The measurement elements `<dimensions>`, `<width>`, `<height>`, and `<depth>` in MASTER have attributes units and scope. In TEI P5, units is renamed unit, scope is retained unchanged, and a new attribute quantity may be used to supply the actual measurement, as an alternative to including it within the content of the element.

It would be desirable to present all measurements uniformly: which method should be adopted?

Resolution: The ENRICH project has decided that where possible all measurements should be converted to millimeters, and where possible this should be stored as a quantity and unit. The ENRICH project felt that the current provisions of `<dimensions>` was limited, and perhaps needed a shape attribute or similar which would then change the semantics of child elements. (i.e. shape='circle' means that `<width>` means 'diameter'). But the ENRICH project believes the TEI should revisit the content model of `<dimensions>` to record the dimensions of more oddly-shaped objects.

At TEI P5, the schema must use the TEI namespace.

6 The Manuscriptorium DTD

The manuscriptorium DTD currently combines three DTD fragments:

- Elements from MASTER, discussed in the previous section of this document.
- About 150 Other TEI Elements, chiefly from the Core, Header, and Names and Dates modules.
- 302 elements from elsewhere, some (but by no means all) of which appear to duplicate functions provided by existing TEI elements

Question: how many of these elements are needed in the ENRICH dtd?

Resolution: These metadata categories are part of the internal manuscriptorium format in the conversion from MASTER to TEI P5, WP3 should not worry about them. Where TEI P5 provides some of this information, such as with the `<facsimile>` element, then Manuscriptorium will attempt to change its internal processes to use that information. It agrees that it will not throw away any incoming metadata fields. However, for the time being, it will be maintaining the use of MASTER inside its MASTER+ container and retro-converting the TEI P5 documents back to MASTER as an internal format for searching, etc. The TEI P5 record will, of course, always be available for downoad. But for the purposes of WP3, we should not concern ourselves with worrying about storing camera metadata, etc.