# TGCW45
# Proposals for amendments to OUP transfer format

Dominic Dunlop

Draft of 22nd February, 1993

In the light of experience gained since mid-1992 at OUCS this paper suggests a number of minor changes to TGCW33, *BNC data capture: OUP format definition for text hand-over to OUCS*. It is OUCS' opinion that these changes will simplify processing, particularly of the unpublished material to be collected by Chambers and captured by OUP.

In the following, section numbers refer to those of TGCW33.

## 2 — General file conventions

- TAB characters are used in very few texts — 65 of 1,433 received to date from OUP. Since TABs are stated to be equivalent to spaces, we suggest that they are translated to spaces before texts are sent to OUCS.
- The initial `<head>` tag in a text is stated to contain "some identifying title". In our experience, for books it often contains the author name as well as the title. It would be more helpful to us if only the title appeared at this point.

## 3.1 — Basic structure: written text

- Where a book is divided into parts containing chapters, `<div1>` is almost always incorrectly used for both structural units. We suggest either that, in such texts, `<div1>` is retained for parts and `<div2>` used for chapters; or that `<div0>` is used for parts and `<div1>` is retained for chapters. (Either is acceptable, although we would prefer the latter, as it is consistent with what we have been doing.)
- OUP's data capture format strongly suggests to its users that `<hd1-4 ...>`, the marks corresponding to CDIF's `<div1-4>`, should only be used where there is an associated heading. It is often helpful to use these marks even where there is no heading — for example, to separate several readers' letters on the same topic. Data capture staff should be encouraged to do this. (See also 3.2 below on `<head>`.)

- There is a tendency for captured material to contain a <div*n*><head>... </head> sequence for every headline. This is incorrect, and introduces spurious extra structural levels where multiple headlines introduce a single piece of text. See the discussion of <head> in 3.2 below, which shows how multiple <head>s may follow a single <div*n*>.
- The use of <div1–4> in material captured from periodicals and ephemera is very inconsistent, both within individual texts and between texts. While the intent is that these marks should show the structure of a text, their use seems more often to be related to the typeface used for the corresponding headline than to anything else. The process of retrofitting consistency has been very expensive in terms of OUCS' time. More guidance to data capture staff is needed. While TGCW46, *Proposed Guidelines to Keyboarders for Magazine Capture* (forthcoming from OUCS) addresses this issue, OUP must take responsibility for the production of tutorial material and staff training derived from it.

## 3.2 — Paragraph-level elements

- In some cases, direct speech, conventionally represented with one paragraph per speaker turn, does not have sufficient <p> tags applied. OUP should review imaginative material for this type of error.
- The laxity of the current description of <head> can cause considerable problems in converting to CDIF, which has a much tighter specification. We strongly suggest that the description is amended to read as follows:

> A <head> may appear only immediately after the beginning of a marked structural division (<div0–4>), or immediately after the beginning of a <list> or <poem>. A single <head> or a sequence of <head>s can appear at any of these locations. Where an article has multiple headlines (including bylines), multiple <head>s must be used. In all other situations where a <head> might seem appropriate, a caption (<ct>) should be used instead. If a caption appears between two headlines in the source text, it should be captured after the sequence of <head>s. (Since the position of captions relative to surrounding text is not indicated by mark-up, moving a caption in this manner is acceptable.)

The bulk of uses of <head> in received material satisfy this specification. Those which cause problems tend to look like

**BOMB BLAST**
Two held
*Staff reporter*

This must be captured as

```
   <head>BOMB BLAST</head>
<head>Two held</head>
<head>Staff reporter</head>
```

It is not acceptable to capture it with line breaks or blank lines separating the parts: section 2 states that one or more line break is functionally equivalent to a single space — a rule which, when acted upon by subsequent reformatting, runs the separate headlines together.

- It should be made clear that <p> tags may be used inside <ct> (caption) if the material inside the caption is divided into paragraphs. In such cases, the first thing inside <ct> must be <p>. As a special case, where a caption has a heading followed by body text, the heading should be marked as the first <p>, the body text as the second and subsequent <p>s. (This is necessary because CDIF has no mechanism for giving headings to captions.)
- While no examples have been seen to date, the use of <p> and <div*n*> tags within <poem>s has no analogue in <cdif>, and should, we suggest, not be permitted.
- The <table> tag is seldom used (approx. 40 texts), and, where it is used, its contents seldom follow the specification given. We suggest that no tabular material is captured unless it can be easily represented as a <list> (or, if it is considered valuable enough, as nested <list>s). (See also &table; below.)

## 3.3 — Addition, deletion and regularization

- <del desc='Pages *m*–*n* omitted' cause='sampling strategy'> should be used at the beginning or end of a sample to list omitted pages where the beginning or end of a part or chapter of a book has not been captured because to have done so would have made a sample too long. If OUP desires, OUCS will accept a "pseudo entity", &samp*m*–*n*;, as a shorthand for such <del> tags.
- TGCW33 states that <del> should be used in preference to <note> to mark deleted material in the middle of sentences. In fact, it would be desirable if it were always used at any point: <note> is not used for this purpose by CDIF.
- Entities &addr;, &figs;, &illus;, &name;, &table; and &tel; may be used as a shorthand for the deletion of addresses, figures, illustrations or pictures, names, tabular material, and telephone numbers respectively.
- While TGCW33 says that some words marked with <sic> may actually not be in error in the original, higher standards of marking should be enforced if possible. For example, we find it hard to imagine why every occurrence of *analyse* and *Argentinian* in one series of texts was marked with <sic>.

## 3.4 — Quotations and highlighted phrases

- The rendition attribute to <hi> should be mandatory.

## 3.5 — Miscellaneous elements

- The definition of the <pb> tag implies, but does not specify, that the value given to the n attribute is the number of the new page. This should be made clear. Perhaps 30% of texts with page numbers have n attributes off by one because the number of the old page is given.
- The <salute> tag has been seen in only eight received texts. Unless it is considered necessary for unpublished material, it should be dropped.

## 4 — Special and non-ASCII characters

- Nested quoted material currently results in errors in the use of normalized quotation marks. The software which inserts these marks should be reviewed. (Already in hand.)
- The hyphenation-removal software (currently under review) should be amended so as correctly to handle words hyphenated across page breaks.
- The sequences ' –' at the end of a line, and '–' at the beginning of a line should be checked: they generally indicate an em-dash which should be replaced by an entity reference.
- Backslash (\) is not allowed in content, but sometimes appears. We have never found a case where this represented an actual backslash in the source text. We suggest that OUP examines all instances of literal backslashes in outgoing material, and uses the entity \ if a backslash really appears in the source text.
- It is quite common for the degree sign (°) to appear incorrectly as a superscript zero (&sup0;).
- The entity &formula; is under-used: many formulæ whether mathematical, chemical or atomic, remain in some texts, generally with transcription errors. Since transcription and correction are time-consuming, all formulæ in text should be replaced with &formula;.