TGCW23
Sample CDIF-encoded texts from OUP
Preliminary OUCS findings

Dominic Dunlop

7th January, 1992

1 Introduction

This report summarizes the changes which were found necessary in order to make thirteen sample corpus texts provided by OUP to OUCS parse according to the draft CDIF DTD described by TGCW18. (Owing to time constraints, it was not possible to check all the thirty-two texts provided by OUP at the beginning of December.)

The page layout is not all that it could be: forgive me for not having the time to spend on convincing LaTeX to do exactly what I want!

2 Problems with CDIF

2.1 Amendments to existing draft CDIF

In the course of the investigation, a small number of changes had to be made to the draft CDIF of TGCW18:

- As an alternative to a `<label>`, an `<enum>` was allowed ahead of each `<item>` in a `<list>`. On reflection, the existence of the `<label>` tag in TEI P1 as a means of labelling list items probably means that there is no need to use `<enum>` for this purpose in CDIF; for the moment, however, `<enum>` is being used.
- The `<ct>` (caption) tag as added. We decided to do this at the task group C meeting of 23rd August, but I had omitted to update the CDIF DTD until now.
- Some changes were made to the description of front matter: *crystals* (notably `<list>`s were allowed as content; the requirement for a title page in front matter was removed. Both these changes seemed reasonable, given the type and level of mark-up which we can expect in incoming texts.
- The `<hi.b>` and `<hi.e>` tags introduced as a temporary expedient during early work on *The Wimbledon Poisoner* were removed again.
- Entities `&ft;` and `&ins;` were added to handle single and double quotes used for *feet* and *inches* respectively. (SGML public entity sets do not contain these symbols.)

2.2 The impact of a new CDIF DTD

Towards the end of the processing of the OUP texts, work started on the development of a new DTD for CDIF. This is to be a purpose-built DTD, replacing the current prototype, which closely follows the TEI's *toy2* DTD, and so carries along a fair amount of excess baggage not required by CDIF. The existing prototype is, as we have discovered, also ambiguous, and rather more permissive than OUCS considers desirable. Work on the replacement is not complete: the results presented here refer only to the earlier prototype. It may be that the replacement CDIF DTD will highlight additional problem areas in incoming

texts. However, it will also fix some existing problems, such as the current inability to have a caption (`<ct>`) as a floating element within a paragraph (`<p>`).

3 The incoming texts

The following texts, designated by the six-character filenames assigned by OUP, were brought into line with CDIF:

AidFct — a collection of five factsheets etc. on AIDS published by a Christian charitable organization.

AidLft — a collection of seven leaflets etc. on AIDS, all published by the same organization as is responsible for the contents of `AidFct`.

Author — *Authors*, Karl Miller (OUP, 1989). A sample from the middle of a relatively academic text on the attributes of authors and the process of authorship.

Belief — *The Tragedy of Belief*, John Fulton (OUP, 1991). A sample from the middle of an account of "the role of religion in the division and conflict in Ireland." Again, a relatively academic work.

BigGls — *The Big Glass*, Gabriel Josipovici (Carcanet Press, 1991). A complete and somewhat "arty" book describing the creation of *Large Glass: The Bride Stripped Bare by her Bachelors* .

ECrime — *A Classic English Crime*, ed. Tim Heald (Pavilion, 1990). A sample from the end of a selection of short stories, written mainly in the mode of Agatha Christie by a variety of current authors.

FilmTV — *Film & Television*, September, 1991. "Your essential free guide to the festival", a tabloid newspaper-style giveaway distributed at the 7th Birmingham international film and TV festival.

Inside — *An Inside Job*, Malcolm Young, (OUP, 1991). A sample from the beginning of a middle-brow insider's view of modern policing.

LCohen — *Leonard Cohen, Prophet of the Heart*, L. S. Dorman and C. L. Rawlings (Omnibus Press, 1990). Sample from the beginning of a middle-brow biography of the Canadian poet by two British writers.

Pursue — *The Pursuit of Mind*, ed. Raymond Tallis and Howard Robinson (Carcanet Press, 1991). Sample from the beginning of a volume of academic papers by a number of authors on the subject of the "current philosophy of mind."

SoVery — *So Very English*, ed. Marsha Rowe (Serpent's Tail, 1991). Sample from the end of a compilation of medium-radical short stories, plays and poetry by various contributors centered on "the vagaries of English life".

Weight — KTG *Know the Game — Fitness with Weights*, Alan Fleet, (A & C Black, 1991). Complete illustrated booklet on the subject of weight training.

Woodwk — *Woodworker*, August, 1991 issue (Argus Specialist Publications). Complete editorial material from popular woodworking magazine.

	AidFct	AidLft	Author	Belief	BigGls	ECrime
<chp>[1]	–	–	•	•	–	•
<ct>	•	•	–	–	–	–
<div0>	•	•	•	•	•	•
<div1>	•	•	•	•	–	–
<div2>	•	•	–	•	–	•
<div3>	•	•	–	•	–	•
<enum>	•	•	–	•	–	–
<fo>[1]	–	–	–	–	–	•
<head>	•	•	•	•	•	•
<hi>	•	•	•	•	•	•
<list>	•	•	–	•	–	–
<note>	–	•	–	–	–	•
<p>	•	•	•	•	–	•
<pb>	•	•	•	•	•	•
<poem>	–	–	•	–	–	•
<q>	–	–	•	•	–	•
<sic>	–	•	•	–	–	–

Table 1: Features tagged by OUP (first six documents)

	FilmTV	Inside	LCohen	Pursue	SoVery	Weight	Woodwk
<chp>[1]	–	•	•	•	•	–	–
<ct>	•	–	–	–	•	•	•
<div0>	•	•	•	•	•	•	•
<div1>	•	–	•	•	–	–	•
<div2>	•	•	•	•	•	•	•
<div3>	•	•	–	•	•	•	•
<enum>	–	–	–	–	–	–	–
<fo>[1]	–	•	–	•	–	–	–
<head>	•	•	•	•	•	•	•
<hi>	•	•	•	•	•	–	•
<list>	–	–	–	–	–	–	–
<note>	•	•	•	•	–	•	–
<p>	•	•	•	•	–	•	–
<pb>	•	•	•	•	•	•	•
<poem>	–	–	•	•	•	–	–
<q>	–	•	•	•	•	–	–
<sic>	–	–	–	–	–	–	–

Table 2: Features tagged by OUP (remaining seven documents)

4 Features of the incoming texts

Tables **??** and **??** list all tags appearing in the texts received from OUP[2], and the processed texts in which they appear. Some notes are in order:

- Where the `<head>` and `<note>` tags are used only in connection with header-related functions, their use is not shown by the tables, which are concerned with tag usage in the body of the text.
- The `<chp>` and `<fo>` tags are not defined by CDIF, but appear in received texts.
- On examination, many of the texts proved to contain features provided for by the data capture markup (*Codes for Freelancers*, October, 1991 (TGCW04 annex)), but which were not actually marked up. (Lists and cover pages are cases in point.) This is probably to be expected: we cannot expect that, because *Freelancers* provides for the tagging of a particular feature, all instances of that feature will be tagged.

Table **??** shows CDIF tags which do not appear in any of the incoming texts, together with the probable reason for their omission.

5 Problems with the texts

There was considerable variation in the level, completeness and consistency of markup both between and, to a smaller extent, within, texts. For example, some texts had no paragraphs (`<p>`s) marked, while, in others, not all paragraphs were marked.

Table **??** shows how long it took me to edit each text into a state in which it parsed without error using the `vm2` parser and the amended prototype CDIF DTD. These times include time required for the creation of `awk`, `sed` and `sh` scripts as necessary. The word counts were obtained with the UNIX command line

`stripsgml <` *filename* `| wc -w`

where `stripsgml` is a simple-minded program written in Icon, and `wc` is the standard UNIX word-counting program, which counts sequences of non-whitespace separated by whitespace. As orthographic word counts, the numbers are slightly high because:

- `stripsgml` is not quite as thorough as it might be in removing document prologue information;
- The vestigial header remains in the text;
- Non-word entities such as `&emdash;` remain in the text; and
- The content of `<note source=ed>` tags remains in the text.

.

Tables **??** and **??** show the changes which were found necessary in order to make the texts parse according to prototype CDIF. Table **??** also shows the category into which I consider the problem falls:

system Systematic miscoding, where some tag or entity name is consistently spelled incorrectly, or used in the wrong context;

parse Parse errors, where the incoming text is supposed to be in CDIF format, but does not parse when run through an SGML markup validator;

mistag Mistagging, where features are incorrectly marked with the same tag as other, correctly-marked features; and

[1]This tag is not defined by CDIF.

[2]Three tags are not listed, as they do not appear in the thirteen texts processed: `<hd>`, `<np>`, and `<table>` — although they do appear in other texts. Since none of these is defined by CDIF, all must be considered erroneous.

[3]There is a short play in `SoVery`, but it is not part of the captured sample.

Tag	Function	Notes
`<abbrev>`	Abbreviation	??
`<add>`	Addition	??, ??
`<back>`	Back matter	??, ??
`<corr>`	Correction	??
`<date>`	Date	??
`<del>`	Deletion	??
`<div4>`	Smallest text subdivision	(none)
`<epigraph>`	Epigraph	??
`<foreign>`	Foreign word(s)	??
`<item>`	List item	??, ??
`<l>`	Verse line	??
`<norm>`	Normalization	??
`<propname>`	Proper name	??
`<s>`	Text segment	??, ??
`<salute>`	Salutation	??
`<signed>`	Signature	??
`<stanza>`	Verse stanza	??
`<trailer>`	Trailing material	??, ??
`<u>`	Utterance	??, ??, ??

1. No encoding for this feature in *Freelancers*.
2. No suitable material captured, as far as I can see.
3. Correction and normalization are not allowed by *Freelancers*. (See, however, `<sic>`.)
4. The marking of deleted material is one of the functions of the *Freelancers* `<note>` tag.
5. See first item in §**??**.
6. See *Fix incorrectly tagged* `<list>`*s,* `<poem>`*s* in §**??**.
7. Subsequently to be tagged automatically.
8. Material is either omitted or tagged as sub-`<div>`.
9. Not relevant — except in the case of scripts[3]?

Table 3: Tags not appearing in incoming texts

File	Words	Fix time	Comments
AidFct		1:00	Necessary to tag much free-floating text.
AidLft	6,717	2:00	Fixing of <list>s time-consuming.
Author	41,182	2:30	Genuine problems with unbalanced <hi>s and dropped or corrupted text; self-made problems with quote marks. Also carries overhead of developing global change script for all texts.
Belief	7,956	1:00	(None)
BigGls	45,549	0:25	Text consists of just one long paragraph (which was untagged). Problems with unclosed or doubly-closed <hi>s fixed by reference to book.
ECrime	36,454	2:10	Much free-floating text. Difficult <poem>[4]and strangely-marked footnotes, together with some problems arising from data capture errors. (Also unmatched quote marks.)
FilmTV	14,930	0:30	Text had no <p>s.
Inside	43,137	1:30	Problems with unbalanced <hi>; incorrect nesting of s and use of <div3>; footnotes required attention. (Also unmatched quote marks.)
LCohen	41,813	2:30	Many <poem>s; strange and sometimes unbalanced use of <hi>; questionable use of <div2> for epigraphs, part titles etc.. (Plus problems with quote marks.)
Pursue	41,483	1:00	Unbalanced <hi>s; incorrectly tagged footnotes.
SoVery	35,834	1:30	No <p>s tagged; <poem>s required clean-up; Unbalanced <hi>s. (Also unmatched quote marks.)
Weight	10,611	0:10	Some <p>s unmarked.
Woodwk	31,929	0:30	No <p>s tagged.
	319,919	16:50	

Table 4: Time to fix each text (hours:minutes)

typo Typographical errors due to rekeying or scanning. This class of errors can include unintentionally omited text, and runs of garbage characters in captured text.

— see TGCM22, *Task Group C minutes, 10th December, 1991*, §4.1. Where the class is marked with an asterisk, the problem is associated with an abortive attempt to transduce quote marks into tagging, and can probably be ignored — see *Replace* `‘...&quot;` *with* `<q>...</q>` *throughout* below.

The notes which follow describe the background to each of the changes:

Add missing header & text structures As received, none of the texts correctly specified the DTD of which it was an instance, nor did they have correctly formatted header information. Pending a full definition of the contents of the CDIF header, this problem was overcome simply by citing a small dummy header document in each text.

Add <p> tags for paragraphs quoted with <q>s Where texts contained paragraphs quoted with the <q> tag (block quotations), the <p> was often omitted. This is incorrect: without the <p> tag, the quotation appears to be part of the preceding paragraph.

Add <p>s (sometimes missing) Several texts contained blocks which were clearly paragraphs, but which were not tagged as such, while other apparently similar blocks were tagged. An attempt was made to fix this up by hand, but nothing like full coverage was obtained. (See also *Tag untagged text blocks*.)

Add name attributes for s CDIF requires that at least the first at a particular level in a document has an n attribute (n=chapter, n=subsection etc.). This information was added by hand,

[4]Ostensibly by Robert <sic>Bums</sic> (my tags).

	Type	AidFct	AidLft	Author	Belief	BigGls	ECrime
Add missing header & text structures	system	•	•	•	•	•	•
Add <p> tags for paragraphs quoted with <q>s	system	–	–	–	•	–	•
Add <p>s (sometimes missing)	mistag	–	•	–	–	–	•
Add name attributes for s	system	•	•	•	•	•	•
Close unclosed <q>s resulting from conventional typography	system	–	–	•	•	–	•
Close <p>s ahead of <list>s	system	•	•	–	•	–	•
Correct incorrectly tagged footnotes	system	–	•	–	–	–	•
Correct unbalanced <hi> tags	typo	–	–	•	–	•	–
Correct unbalanced ‘ and "	typo*	–	–	•	•	•	•
Correct ’ to ‘ or "	typo*	–	–	–	•	–	•
Correct ‘ or " to ’	typo*	–	–	–	•	•	•
Delete incorrect <enum>s from <list>s	mistag	•	•	–	–	–	–
Delete remaining <chp> tags	system	–	–	•	•	–	•
Delete spurious <hi>s at line ends	system	–	–	–	–	–	–
Divide into <front> and <body>	mistag	–	–	–	–	–	–
Enclose floating <hi>s in <p> etc.	parse	•	•	–	–	–	–
Fix incorrectly nested </hi></q>	parse	–	–	–	–	–	–
Fix incorrectly nested s	parse	•	•	–	•	–	–
Fix incorrectly tagged <list>s	parse	•	•	–	•	–	–
Fix incorrectly tagged <poem>s	parse	–	–	–	–	–	•
Fold rendered= from <hi> to adjacent tag if appropriate	mistag	•	•	–	–	–	•
Notice omitted, mis-scanned or garbled text	typo	–	•	•	•	–	•
Notice <pb>s appear to mark end of page, not start	system	–	–	–	–	•	•
Replace rendering= to rendered= throughout	system	•	•	•	•	•	•
Replace strange “entities” referencing footnotes	system	–	•	–	–	–	•
Replace ∧ with & throughout	system	•	•	•	•	•	•
Replace &degree; with ° throughout	system	–	–	–	–	–	–
Replace % with % throughout	system	•	•	•	•	•	•
Replace ‘..." with <q>...</q> throughout	system*	•	•	•	•	•	•
Tag untagged text blocks	parse	•	–	–	–	•	•

Table 5: Amendments required (first six documents)

	FilmTV	Inside	LCohen	Pursue	SoVery	Weight	Woodwk
Add missing header & text structures	•	•	•	•	•	•	•
Add <dp> tags for paragraphs quoted with <q>s	–	•	•	–	–	–	–
Add <p>s (sometimes missing)	–	–	•	–	–	•	–
Add name attributes for s	•	•	•	•	•	•	•
Close unclosed <q>s resulting from conventional typography	–	•	•	–	•	–	–
Close <p>s ahead of <list>s	–	–	–	–	–	–	–
Correct incorrectly tagged footnotes	–	•	–	•	–	–	–
Correct unbalanced <hi> tags	–	•	•	–	•	–	–
Correct unbalanced ‘ and "	–	•	•	–	•	–	–
Correct ’ to ‘ or "	–	•	•	–	–	–	–
Correct ‘ or " to ’	–	•	•	–	–	–	–
Delete incorrect <enum>s from <list>s	–	–	–	–	–	–	–
Delete remaining <chp> tags	–	•	•	•	•	–	–
Delete spurious <hi>s at line ends	–	–	–	•	–	–	–
Divide into <front> and <body>	•	–	–	–	–	–	•
Enclose floating <hi>s in <p> etc.	–	–	–	–	–	–	–
Fix incorrectly nested </hi></q>	–	•	–	•	–	–	–
Fix incorrectly nested s	–	•	–	•	•	–	•
Fix incorrectly tagged <list>s	–	–	–	–	–	–	–
Fix incorrectly tagged <poem>s	–	–	•	•	•	–	–
Fold rendered= from <hi> to adjacent tag if appropriate	•	•	•	•	–	–	•
Notice omitted, mis-scanned or garbled text	–	•	•	–	•	–	–
Notice <pb>s appear to mark end of page, not start	–	•	–	•	•	–	–
Replace rendering= to rendered= throughout	•	•	•	•	•	•	•
Replace strange “entities” referencing footnotes	–	•	–	•	–	–	–
Replace ∧ with & throughout	•	•	•	•	•	•	•
Replace &degree; with ° throughout	–	–	–	–	–	•	–
Replace % with % throughout	•	•	•	•	•	•	•
Replace ‘..." with <q>...</q> throughout	•	•	•	•	•	•	•
Tag untagged text blocks	–	–	–	–	•	–	–

Table 6: Amendments required (remaining seven documents)

attempting to make the description as appropriate as possible while not spending too much time on the task.

Close unclosed <q>s resulting from conventional typography The underlying problem is that, in a series of paragraphs quoted with quote marks, conventional typography dictates that an extra quote mark appears at the head of the second through last paragraphs. Simple-minded translation to <q> and </q> tags results in as many unclosed elements as there are paragraphs after the first. This is actually not a big problem in the source texts, although there are a few instances: I made my life more difficult by experimenting with transducing all quote makes into <q> tags. (See *Replace* `‘..."` *with* `<q>...</q>` *throughout.*)

Close <p>s ahead of <list>s In CDIF, *crystals* do not close the current paragraph. In most cases, it is necessary explicitly to put a paragraph end tag ahead of a <list>[5].

Correct incorrectly tagged footnotes Where footnotes were tagged with the <fo> tag, this was changed to <note type=footnote>.

Correct unbalanced <hi> tags; delete spurious <hi>s at line ends There were quite a few instances of <hi> tags which were not closed. (CDIF requires explicit end tags.) One document apparently suffered from a processing error which resulted in spurious </hi> end tags at line ends.

Delete incorrect <enum>s from <list>s Where <enum>s were used, I sometimes judged their use inappropriate: for example when they had the content • rather than an enumerator. This issue needs to be addressed by the revised CDIF.

Delete remaining <chp> tags The incoming documents contained spurious <chp> tags. These were deleted, their <n> attribute (if any) being rolled into the corresponding tag.

Divide into <front> and <body> The text from the newspaper and the magazine clearly fell into front matter and body text, so explicit tagging was added to show the division.

Enclose floating <hi>s in <p> etc.; fold rendered= from <hi> to adjacent tag if appropriate In CDIF, text elements tagged with the <hi> tag may only appear inside other elements such as paragraphs: unenclosed <hi>s floating about at the level are unacceptable. In some cases this was fixed by adding <p> tags (see *Tag untagged text blocks*); in others by deleting the tag and moving its rendered attribute to a suitable adjacent tag.

Fix incorrectly nested </hi></q> In a few cases, sequences such as <q> <hi> ...</q> </hi> appeared. This is not legal SGML. (Attempts to transduce quote marks into tags exacerbated this problem.)

Fix incorrectly nested s Several texts had <div2>s which were not enclosed in <div1>s. A few had <div3>s not enclosed in <div2>s. Neither is allowable in CDIF.

Fix incorrectly tagged <list>s, <poem>s Where <list>s and <poem>s were tagged, their content was not tagged correctly. Basically, lists consist of <item>s and poems of <l>s. In both cases, a variety of frills may be added.

Notice omitted, mis-scanned or garbled text With some texts, it was clear that parsing problems were due to problems with data capture. The errors were either corrected — often with reference to the original text — or worked around.

[5] In the case of an “in-line” list such as (a) this one or; (b) another one, it would not be appropriate to end the paragraph. No such lists were tagged, however.

Notice <pb>s appear to mark end of page, not start While the current version of *Notes for freelancers* (TGCW04) specifies that the <pb> tag appears ahead of the material captured from the corresponding page, in many (possibly all) texts it appears that the tag follows all the material. This systematic error could be corrected automatically (although this would be difficult with material from magazines and newspapers), but I didn't bother.

Replace rendering= with rendered= throughout The name of the rendered attribute was consistently misspelled [6].

Replace strange "entities" referencing footnotes Some texts used symbols such as &sup28; to reference footnotes. These were transduced automatically into references such <note type=footnote n=28>, since SGML does not allow entities to be invented on the fly.

Replace &degree; with °, % with %, ∧ with & throughout The names of some entities were consistently misspelled.

Replace ‘...” with <q>...</q> throughout; correct unbalanced ‘ and "; correct ’ to ‘ or ", ‘ or " to ’ I experimented with transducing the normalized open and close quote marks in the incoming text into <q> and </q> respectively. While some documents parsed almost correctly following this change, many had considerable problems. Among the sources of these problems are: conventional typography for multiple quoted paragraphs (see above); the susceptibility of quote marks to scanning errors; incorrect guesses by the heuristic used to normalize quote marks; systematic errors in the transcription of quote marks in some texts; and, in a couple of cases, errors in the original printed material. All in all, it appears impractical for us to attempt to tag quoted material by interpreting quotation marks.

Tag untagged text blocks Several documents had much or all of their text floating freely. In general, CDIF does not allow this, requiring that text is contained (for example) in tagged paragraphs. In some cases the problem was fixed by hand, in others automatically with ad hoc tools.

6 Recommended actions

Referring to table **??**, my processing throughput averages 20,000 words per hour. Reckoning my loaded costs at around £20 per hour, this equates to £1 per thousand words, or £90,000 and 112 Dominic-weeks for the whole written Corpus. Admittedly, this figure is inflated, perhaps by 30% or so, by my experimentation with the processing of quote marks. Even discounting this, and allowing for improved performance with experience, the resulting order of attention is clearly impractical. In order to cut the cost of accession to the Corpus, the correctness of the markup on incoming texts must be improved.

The thirty types of problem encountered fall into the groups shown in table **??**, which also shows how each class of problem might be addressed.

With improvements in the transduction software — for example, by using an SGML-aware translation system[7], most of the problems encountered in this study can be eliminated or, at least, made less demanding of human attention for their resolution. It seems that there is also a need for improvements in at least two areas of data capture: the consistent marking of paragraphs, and the balancing of font-change information. In the case of the latter a text editor which could give an impression of the typeface implied by the font-change markup would probably be of great assistance to the data capture staff[8].

OUCS and OUP should work together to address the issues raised by this study.

[6] Its spelling may yet change again to rend= in revised CDIF.

[7] XTRAN, which OUCS is licenced to use, is such a system. The cost of licencing it for use by OUP would, however, be exorbitant. It should also be possible to build such a system on the basis of the SGML engine which underlies the vm2 parser.

[8] It might be possible to get some way towards this with WordPerfect macros. A syntax-directed editor capable of displaying type in a number of styles, renditions, or colours would be better for the job, however.

Type	Count	Action
Systematic	14	Should be possible to fix with relatively simple changes to *Freelancers* → CDIF transduction software.
Parse	6	Probably requires more complex changes to transduction software (for example, keeping track of context).
Mistagging	4	Requires transduction software changes and, in some cases, more accurate tagging during data entry.
Typographical	2	Cannot be eliminated entirely; however, the cost of typographical errors which result in unbalanced `</hi>` tags is high, so these must be caught and corrected during data capture if at all possible.
Quote mark	4	These errors can be ignored.

Table 7: Error types and proposed action