TCGW59
CDIF mechanisms available for the representation of anonymized information

Dominic Dunlop

Draft of 15th September, 1993

1 Introduction

In the unpublished material being collected for the BNC by Chambers, and captured into electronic form by OUP, there is a requirement for the anonymization of information which, if left intact, would allow private individuals, confidential information, and, to a lesser extent, organizations to be identified. In an ideal world, it would be desirable to assign unique identifiers to each distinct anonymized entity in the corpus, in order that references to the entity could be tracked within and across corpus texts.[1] While CDIF could, with a small, TEI-conformant, extension, accommodate a scheme representing this level of detail, the corpus project does not have the resources to do this on behalf of contributors to the unpublished corpus, nor is it likely that contributors themselves would have the motivation (or the tools) to provide the necessary mark-up themselves.

Consequently, sections **??** and **??** of this paper discusses a variety of anonymization mechanisms available in CDIF as currently defined. Unpublished texts passed by OUP to OUCS should represent anonymization using these mechanisms. Chambers may wish to pass texts obtained in electronic form to OUP marked up in some alternative but consistent form which may easily and mechanically be translated to one of the representations shown in this paper. Section **??** presents an example of such a format. Chambers and OUP should agree on the representation to be used in for particular classes of material obtained in particular forms from particular sources. Additionally, where a provider of material in electronic form agrees to anonymize it themselves prior to passing it to Chambers, Chambers should formulate instructions on the marks which should be put into the material in order to anonymize it.

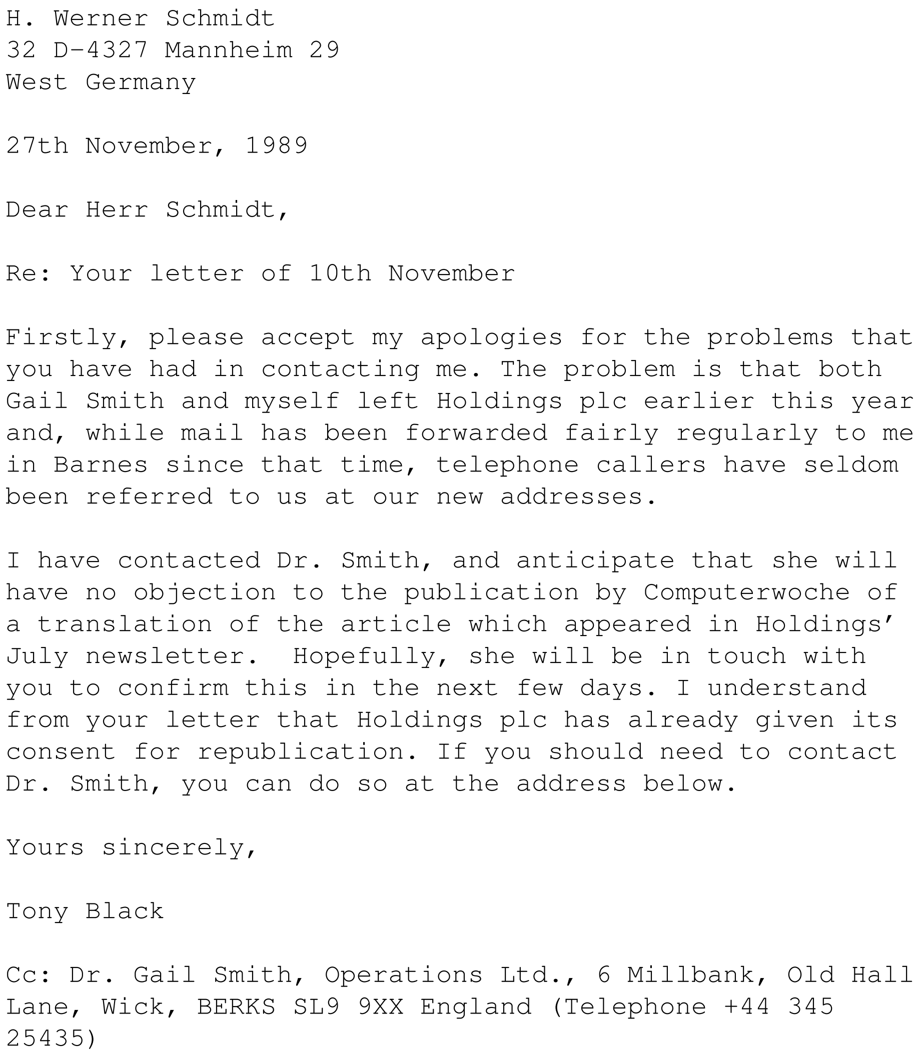
The introduction of anonymization may have implications for the word-class tagging of BNC texts. This issue is discussed in section **??**.

1.1 Sample material

Examples will show how the business letter shown in figure **??** may be anonymized.[2] In order to avoid clutter, most figures show only tagging relating to anonymization: tagging for paragraphs and headings is not shown.

[1] The anonymization applied to participants in the written corpus allows speakers explicitly to be tracked in this manner, and usually permits the resolution of references in speech to other participants by name.

[2] The letter is genuine, but has already been anonymized through the substitution of invented names.



H. Werner Schmidt
32 D-4327 Mannheim 29
West Germany

27th November, 1989

Dear Herr Schmidt,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that you have had in contacting me. The problem is that both Gail Smith and myself left Holdings plc earlier this year and, while mail has been forwarded fairly regularly to me in Barnes since that time, telephone callers have seldom been referred to us at our new addresses.

I have contacted Dr. Smith, and anticipate that she will have no objection to the publication by Computerwoche of a translation of the article which appeared in Holdings' July newsletter. Hopefully, she will be in touch with you to confirm this in the next few days. I understand from your letter that Holdings plc has already given its consent for republication. If you should need to contact Dr. Smith, you can do so at the address below.

Yours sincerely,

Tony Black

Cc: Dr. Gail Smith, Operations Ltd., 6 Millbank, Old Hall Lane, Wick, BERKS SL9 9XX England (Telephone +44 345 25435)

Figure 1: Example of a letter

&addr;	An address
&anon;	General-purpose anonymization (see text)
&date;	A date
&desig;	A personal designator (vice chancellor, lord mayor, etc.)
&fax;	A facsimile number
&name;	A person's last or full name
&namef;	A female person's last or full name
&namem;	A male person's last or full name
&namen;	A person's nickname
&num;	A number, such as an account number or bank sort code
&org;	An organization (company, charity etc.) name
&place;	A place name
&prod;	A product or trade name
&publ;	A publication name
&tel;	A telephone number (also facsimile, telex etc.)
&telex;	A telex number

Table 1: CDIF marks for anonymization

2 Simple anonymization

CDIF provides a selection of marks which provide for anonymization by replacing all references to objects of a particular type with the same type-specific mark. The marks[3] are shown in table **??**. By use of these marks, all information which might allow individuals and organizations to be identified, either directly or by inference, may be removed from a text. Where both more and less precise marks are provided — as in `&name;` `&namef;`, `&namem;` and `&namen;`; `&tel;`, `&fax;` and `&telex;` — the less precise forms may be used as an alternative to the correct more precise forms. Precision is recommended if resources allow.

`&anon;` should be used only if none of the other marks is suitable. As a contrived example, the possibly-actionable `I've heard that the best-selling drug for ulcers kills you` might be anonymized as `I've heard that the best-selling drug for &anon; kills you`. Alternatively, OUCS can provide for further marks, should these prove necessary.

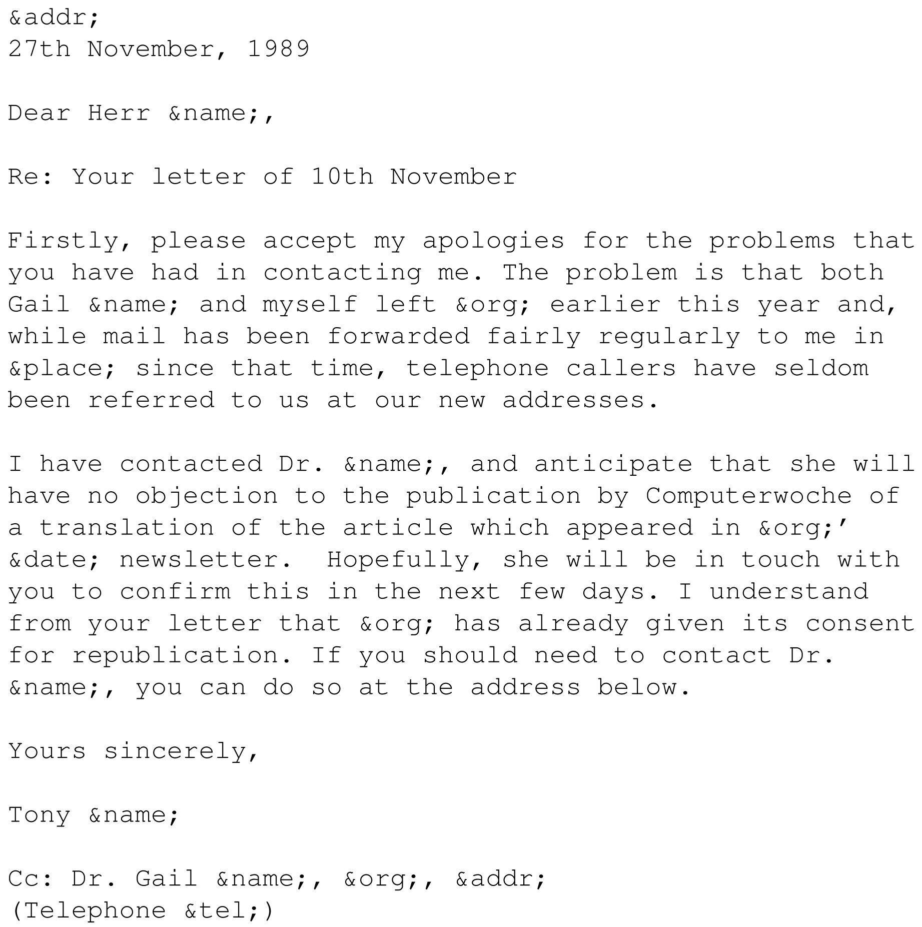
Figure **??** shows the example letter anonymized using simple marks.

Because just one distinct mark is used to replace, for example, all organizations, the distinction between them is lost. Note, however, that not all dates, and not the whole of each person's name has been replaced: no more information has been replaced than is necessary to perform the anonymization. This leaves in place some contextual information which corpus users may find helpful. (Although it is tricky to work out that `Gail &name;` and `Dr. &name;` refer to the same person.)

2.1 Simple treatment of headers and trailers in letters

The beginning and end of a letter or memo generally provide a great deal of personal information about the text's originator and recipient, but are seldom linguistically interesting. As such, they are prime targets for anonymization.

[3] Considerations of limits in SGML parsers suggest that the names of the marks should consist of eight or fewer characters; short names are also convenient when keyboarding material



&addr;
27th November, 1989

Dear Herr &name;,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that you have had in contacting me. The problem is that both Gail &name; and myself left &org; earlier this year and, while mail has been forwarded fairly regularly to me in &place; since that time, telephone callers have seldom been referred to us at our new addresses.

I have contacted Dr. &name;, and anticipate that she will have no objection to the publication by Computerwoche of a translation of the article which appeared in &org;' &date; newsletter. Hopefully, she will be in touch with you to confirm this in the next few days. I understand from your letter that &org; has already given its consent for republication. If you should need to contact Dr. &name;, you can do so at the address below.

Yours sincerely,

Tony &name;

Cc: Dr. Gail &name;, &org;, &addr;
(Telephone &tel;)

Figure 2: Letter with simple anonymization

```
Dear Herr &name;,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that
you have had in contacting me. ...

...

Yours sincerely,
```

Figure 3: Letter with opening and closing matter removed

Figure **??** shows how headers and footers may be excised. Salutations are preserved, since they may be of interest if the relationship between originator and recipient is known. (See also next subsection.) The date of the letter has been excised along with the recipient address, as this is likely to happen in practice if some mechanical procedure is used to delete everything prior to the word `Dear`. (Or prior to the last "short paragraph" ahead of the first "long paragraph", or some other heuristic.)

3 More detailed anonymization

3.1 Provision of contextual information for material

It is likely to be helpful for corpus users to know the context in which an unpublished work was generated. For example, the relationship between writer and audience is almost certain to affect the register of the language used. However, texts are unlikely to contain information of this type, since writer and audience generally know their relationship without the text having to state it. If the information is to be provided, it must be added. Generally only the originator, or a colleague of the originator, can do this: BNC staff will not be aware of the relationships involved, or of other contextual information.

The information must be provided in a manner which makes it clear that the words in the description of context are not part of the original text. The CDIF editorial note fills this rôle. Context may provided in an editorial note ahead of the text, as figure **??** shows. The date of the letter has been preserved here, as it may be useful in placing the correspondence in context in a series of communications between the same parties.

3.2 Provision of information about anonymized objects

The simple scheme of §**??** does not allow the representation of contextual information about the anonymized items. For example, if a text refers to more than one anonymized organization, the mark `&org;` is used to represent the names of all of them, so making it difficult either to determine how many organizations are being referenced, or which organization is being referenced in a particular context.[4] To overcome this problem, some means of replacing sensitive information with an anonymized description is required.

[4] Where a text names more than one person, some contextual information may be preserved by anonymizing only last names: for example, `Gail Smith` becomes `Gail &name;`.

```
<note type=ed ed=auth>Letter to German inquirer not
previously known to author</note>
27th November, 1989

Dear Herr &name;,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that
you have had in contacting me. ...

...

Yours sincerely,
```

Figure 4: Letter with prefactory note by anonymizer

The empty CDIF `<omit>` tag fulfils this function. Its `desc` attribute carries the anonymized description; the `cause` attribute a reason for the the omission (`anonymization` in this case); and the `ed` attribute an identifier for the person or organization responsible for omitting the replaced material. (Both `desc` and `cause` should be specified; `ed` is optional.)

Figure **??** shows the example letter anonymized using `<omit>` tags. Note that `<omit>` is used only where there is, in the opinion of the editor, scope for confusion; the simple marks (`&date;`, `&place;`...) continue to be used in other situations.

The result of anonymization using `<omit>` is likely to be both verbose and repetitive. A means of avoiding these factors during data capture is presented in the next subsection.

3.3 Example data capture format

Figure **??** presents a simple, non CDIF-compliant, data capture format which may easily be transformed to the CDIF-compliant form shown in figure **??**. A program, written, for example, in `perl`, can use leading lines beginning with digits followed by a closing square bracket to populate a look-up table. Instances of digits enclosed in square brackets in the body of the text may then be replaced by an `<omit>` element with its `desc` attribute provided by the relevant entry in the look-up table. For example, `[2]` in figure **??** is transformed to `<omit desc='former employer' cause=anonymization>` in figure **??**.

4 Anonymization in the context of a CDIF text

4.1 Replacement text for anonymization marks

The marks shown in table **??** are SGML *entity references*, which are replaced, on processing by SGML-aware software with their expansions. Table **??** shows the marks and their replacements.

The replacements are `<omit>` elements like those shown in §**??**, but provide generic, rather than specific, information, and, because they have no `ed`

```
<note type=ed ed=auth>Letter to German inquirer not
previously known to author</note>
27th November, 1989

Dear Herr &name;,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that
you have had in contacting me. The problem is that both
<omit desc='colleague at former workplace'
cause=anonymization> and myself left <omit desc='former
employer' cause=anonymization> earlier this year and,
while mail has been forwarded fairly regularly to me in
&place; since that time, telephone callers have seldom
been referred to us at our new addresses.

I have contacted <omit desc='colleague at former
workplace' cause=anonymization>, and anticipate that she
will have no objection to the publication by
Computerwoche of a translation of the article which
appeared in <omit desc='former employer'
cause=anonymization>'s &date; newsletter.  Hopefully, she
will be in touch with you to confirm this in the next few
days. I understand from your letter that <omit
desc='former employer' cause=anonymization> has already
given its consent for republication. If you should need
to contact <omit desc='colleague at former workplace'
cause=anonymization>, you can do so at the address below.

Yours sincerely,
```

Figure 5: Anonymized letter with contextual information

```
1] colleague at former workplace
2] former employer
<note type=ed ed=auth>Letter to German inquirer not
previously known to author</note>
27th November, 1989

Dear Herr &name;,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that
you have had in contacting me. The problem is that both
[1] and myself left [2] earlier this year and, while mail
has been forwarded fairly regularly to me in &place;
since that time, telephone callers have seldom been
referred to us at our new addresses.

I have contacted [1], and anticipate that she will have
no objection to the publication by Computerwoche of a
translation of the article which appeared in [2]’s &date;
newsletter.  Hopefully, she will be in touch with you to
confirm this in the next few days. I understand from your
letter that [2] has already given its consent for
republication.  If you should need to contact [1], you
can do so at the address below.

Yours sincerely,
```

Figure 6: Suggested data capture format

&addr;	<omit desc=address cause=anonymization>
&anon;	<omit desc=unspecified cause=anonymization>
&date;	<omit desc=date cause=anonymization>
&desig;	<omit desc=’personal designator’ cause=anonymization>
&fax;	<omit desc=’facsimile number’ cause=anonymization>
&name;	<omit desc=’last or full name’ cause=anonymization>
&namef;	<omit desc=’female last or full name’ cause=anonymization>
&namem;	<omit desc=’male last or full name’ cause=anonymization>
&namen;	<omit desc=nickname cause=anonymization>
&num;	<omit desc=number cause=anonymization>
&org;	<omit desc=’organization name’ cause=anonymization>
&place;	<omit desc=’place name’ cause=anonymization>
&publ;	<omit desc=’publication name’ cause=anonymization>
&prod;	<omit desc=’product or trade name’ cause=anonymization>
&tel;	<omit desc=’telephone number’ cause=anonymization>
&telex;	<omit desc=’telex number’ cause=anonymization>

Table 2: CDIF marks for anonymization

```
27th November, 1989

Dear Herr ,

Re: Your letter of 10th November

Firstly, please accept my apologies for the problems that
you have had in contacting me. The problem is that both
Gail and myself left earlier this year and, while mail
has been forwarded fairly regularly to me in since that
time, telephone callers have seldom been referred to us
at our new addresses.

I have contacted Dr. , and anticipate that she will have
no objection to the publication by Computerwoche of a
translation of the article which appeared in '
newsletter.  Hopefully, she will be in touch with you to
confirm this in the next few days. I understand from your
letter that has already given its consent for
republication. If you should need to contact Dr. , you
can do so at the address below.

Yours sincerely,

Tony

Cc: Dr. Gail , ,
(Telephone )
```

Figure 7: Anonymized letter with tagging discarded

attribute, do not identify an editor.

4.2 Word-class tagging considerations

If an anonymized CDIF text is passed through an SGML processor, or if SGML entities and tags are simply stripped from the text, the marks used to replace anonymized words disappear. Figure **??** shows the effect of this on the simple anonymization first seen in figure **??**. It is likely that a parser, such as the CLAWS parser used at Lancaster on BNC texts, will produce poor results when words are omitted in this manner. Lancaster should consider whether it is possible to give CLAWS information along the lines of *omitted word; probably a proper noun* on encountering `<omit>` elements resulting from anonymization.[5]

[5] Lancaster will not see the marks shown in table **??**: these will be expanded to `<omit>` elements before texts are forwarded to Lancaster by OUCS.