Tags required

<text> a complete text
Attributes: *id* an ID number that links to the metadata spreadsheet

<p> indicates the start or end of a new para.
Attributes: *align* can be “left”, “center” (note the US spelling) or “right”. Default to left unless specified.

<head> indicates that this text is a heading or subheading. A heading must always be a "paragraph" in its own right – otherwise just use bold, italic or capital letters as appropriate.
Attributes:
level can be “1”, “2”, or “3” 1 is the largest level of heading used in the chapter or similar; 3 means that the heading is in the same font size as normal text (including if it is in bold or italics); 2 is anything in between.
align can be “left”, “center” or “right”. Defaults to “left”

<chap> encloses a chapter or similar section of a document. David Cooper will decide what does and does not constitute a chapter for each document
Attributes:
n the chapter number.
title the chapter title if given

<pb/> indicates a page break
Attributes:
n the number of the page that begins **after** the break… If an unnumbered page is inserted that would disrupt the sequence use n="236.5" (one unnumbered page) or n="27.3" and n="27.6" (two).

<gap/> indicates that there is a gap in the text for something such as an image
Attributes:
desc a (brief) description of whatever it is that is breaking up the text for example "Figure 1: Image of Grasmere" or "Thumbnail map of Derwent Water". Include a note if the picture uses landscape instead of portrait.

<b> bold text

<i> italics

<unclear> the text between the tags is illegible

<poem> indicates the text that follows is a poem. A poem can contain the following tags (which should not be used outside a poem). It must have at least one line but need not have anything else.

<title> the title of the poem. Only used if the title is given

<author> the author. Only used if the author is given

<stanza> encloses a stanza. This only needs to be used if there is more than one verse.

<line> indicates each line

Special symbols: (note that these are not enclosed in <>)

< (less than) **<**
> (greater than) **>**
& (ampersand) **&**
ß (sharp s) Use Word's Insert – Symbols – Symbol (under Latin-1 supplement)
¼, ½, ¾, ⅛ Use Word's Insert – Symbols – Symbol (under Latin-1 supplement or Number forms)

Within attribute values the following also need to be coded: " (double quote) *"* and ' (single quote/apostrophe) *'*

Comments:

Any comments that you need to add to explain what was done and why should be included as follows (Note that this should be done as little as possible and should usually only be for questions that you want to ask us subsequently): **<!--This is a comment-->**

Tables:

If there are any tables they should be included using the XML <table>, <row>, <cell> format. Please ask if you come across any.

Basic rules for text:

The aim is to produce as close to a complete reproduction of the text on the page as it is possible to create including the capitalizations, spelling and grammar. If any of these appear unusual (or even wrong) do not try to correct them. The only exception to this is that hyphens over line ends should not be reproduced – just type the word as normal.

For an em-dash (–) use two hyphens with a space before and after.

Basic rules for tags:

Note: this is a very brief introduction – for more info see Andrew Hardie's *Modest XML for corpora* document.

- Editing should be done in Word with smart quotes switched off (uncheck the box under *Office Button > Word Options > Proofing > Autocorrect options > Autoformat*) and "replace text as you type" also turned off (uncheck the box under *Office Button > Word Options > Proofing > Autocorrect options > Autocorrect*)
- A tag indicates the structure of the document so <head>A Guide to the Lakes</head> indicates that "A Guide to the Lakes" is a heading. The backslash at the end of the second tag means "end of."
- Some tags do not enclose any text, for example, the <gap> tag indicates a gap in the text that contains, for example, an image. Rather than open and close these tags (<gap></gap>) these are simply written as <gap/> indicating that this both opens and closes the tag.
- Tags can also have attributes, for example, <chap n="1" title="Introduction: A Lake District tour"> bla, bla, bla</chap> indicates that this is chapter 1 and its title is *Introduction: A Lake District tour*. Note that the syntax is important the attribute name is not quoted but its value must be. The "end of" tag never contains attributes. Where a tag forms both a start and end tag (for example <pb/>) the syntax becomes <pb ID="1" />. Smart quotes must never appear within tags.
- Tags usually form nested hierarchies, for example, a text will consist of one or more chapters and chapters will have zero or more headings within them. Thus:

```
<text>
        <chap>
        <head> Introduction to the Lakes</head>
        <p>On arriving in the Lake District…</p>
        </chap>
        <chap>
        <head>Moving on</head>
        <p>Today we will go…</p>
        </chap>
</text>
```

An issue that this may cause issues as tags must nest within each other thus:

- Some text in bold and italic is legal because the italic tags nests within the bold tags however Some text in bold and italic because the bold tag is closed before the italic tag that nests within it.
- This is why pages are expressed just as breaks rather than enclosing the page. As long as we could assume a chapter was split into pages which split into paragraphs we would be fine with this:

```
<text>
    <chap>
        <page>
            <p> Some text…
            </p>
            <p>More text…
            </p>
        </page>
        <page>
        More paras
        </page>
    </chap>
</text>
```

- To handle situations such as this we use page breaks instead to mark the start of each new page so the following more realistically reflects the structure that we use

```
<text>
    <pb/>
    <chap>
        <p> Some text which
            <pb/>
            carries on over the page</p>
        <p>More text in a new para</p>
    </chap>
    <pb/>
</text>
```

- Finally note the need to use special characters if any of the following appear in the within attribute values: <, >, &, ', ". Special symbols must be used instead. Also smart quotes must not be used.

Example from Harrier Martineau's *Guide to Windermere*

<text id="test">

<gap desc="The first x pages are front matter and have not been digitised" />

<pb n="3" />
<chap title="GUIDE TO WINDERMERE">
<head level="1" align="center">GUIDE TO WINDERMERE</head>
<gap desc="symbol" />

A few years ago there was only one meaning to the word WINDERMERE. It meant a lake lying among mountains, and so secluded that it was some distinction even for the travelled man to have seen it. Now, there is a Windermere Railway Station, and a Windermere post office and hotel; -- a thriving village of Windermere and a populous locality. This implies that a great many people come to this spot; and the spot is so changed by their coming, and by other circumstances, that a new guide book is wanted; for there is much more to point out than there used to be; and what used to be pointed out now requires a wholly new description. Such new guidance and description we now propose to give.

The traveller arrives, we must suppose, by the railway from Kendal, having been dropped at the Oxenholme Junction by the London train from the south, or the Edinburgh and Carlisle train from the north.

<pb n="4" />

The railways skirt the lake district, but do not, and cannot, penetrate it: for the obvious reasons that railways canning traverse or pierce granite mountains or span broad lakes. If the time should ever scome when iron roads will intersect the mountainous parts of Westmorland and Cumberland, that time is not yet; nor is it in view, -- loud as have been the lamentations of some residents, as if it were to happen to-morrow. No one who has ascended Dunmail Raise, or visited the head of Coniston Lake, or gone by Kirkstone to Patterdale, will for a moment imagine that any conceivable railway will carry strangers over those passes, for generations to come. it is a great thing that stream can now convey travellers round the outskirts of the district, and up to its openings. This is not effectually done: and it is all that will be done by the steam locomotive during the lifetime of anybody yet born. This most important of the openings thus reached is that of WINDERMERE.

...

<pb n="5" />

[etc, etc]

<pb n="9" />

with rocks, which afford as fine a station as the summit of Elleray for a view of the entire lake and its shores.

<head level="3" align="center">BOWNESS</head>

Is the port of Windermere. There are the new steamboats put up;...

[etc, etc]

<pb n="12" />
house, under its canopy of tall sycamores, and with its pebbly beach, is immediately opposite; ...

On a high wall by the road side, immediately

<pb n="12.3" />
<gap orient="l" desc="Image entitled 'WINDERMERE FROM NEAR STORRS.'" />

<pb n="12.6" />
<gap desc="Blank page with symbol" />

<pb n="13" />
before reaching the gate of Rayrigg, the stranger will be struck with the variety of ferns.

[etc, etc]

<pd n="18" />
boats and safe enough, ... at breakfast and dinner.

<gap desc="Symbol" />

</chap>

<pb n="19" />
<chap title="FIRST TOUR.">
<head level="2" align="center"> FIRST TOUR.</head>
<head level="3" align="center">BY NEWBY BRIDGE AND ULVERSTONE TO FURNESS ABBEY, RETURNING BY CONISTON, HAWKESHEAD, AND THE FERRY</head>

Of the three extended tours...

[etc, etc]

</chap>
<pb n="59" />

<chap title="EXCURSIONS.">
<head level="1" align="center">EXCURSIONS TO AND FROM KESWICK.</head>
<pb n="60" />

<head level="1" align="center">EXCURSIONS.</head>

<gap desc="horizontal line" />

Note. -- The asterisks (*) at the beginning of paragraphs denote objects at the left-hand side of the road, and the figures the distance in miles from the start point
<gap desc="horizontal line" />

<head level="3">FROM THE SWAN INN GRASMERE TO KESWICK.</head>
¼ Tollbar. -- From this point the road rises in a steep though gradual ascent to an elevation of 720 feet.
¼. Fairfield and Seat Sandal.
¼. *Helm Crag -- A singularly-shaped hill, affording from its summit a delightful prospect. The curious appearance presented by its rugged apex has given rise to some fanciful comparisons. Seen from one part of the valley it strikingly resembles a lion couchant, with a lamb lying at its nose: from another, an old woman cowering. Wordsworth in his "Johanna," designates it as

<poem>
<line>"That ancient woman seated on Helm Crag."</line>
</poem>

An again, in the "Waggoner," thus alludes to this singular appearance, giving, as will be seen, a companion to the Ancient Woman.
<poem>
<line>"The Astrologer, sage Sidrophel,</line>
<line>Where at his desk he nightly sites,</line>
<line>Puzzling on high his curious wits;</line>
<pb n="61" />
<line>He, whose domain is held in common,</line>
<line>With no one but the ancient woman:</line>
<line>Cowering beside her rightful cell,</line>
<line>As if intent on magic spell.</line>
<line>Dread pair, that, spite of wind and weather,</line>
<line>Still sit upon Helm Crag together!</line>
</poem>

2½. Dunmail Raise. -- This celebrated pass admits the traveller into Cumberland. A Cairn, or pile of stones, is said by tradition, to have been raised here, in the year 945, by Edmund, the Anglo Saxon King, in commemoration of a victory gained over Dunmail, the British King of Cumbria. The British King was slain here, and his territory given to Malcolm, King of Scotland. Part of this cairn still remains.
<poem>
<stanza>
<line>"They now have reached that pile of stones</line>
<line>Heap'd over brave king Dunmail's bones:</line>
<line>He who once held supreme command,</line>
<line>Last King of rocky Cumberland;</line>
<line>His bones, and those of all his power</line>
<line>Slain here in a disastrous hour."</line>
</stanza>
-- <author>WORDSWORTH.</author>
</poem>
[etc, etc]

</chap>
</text>