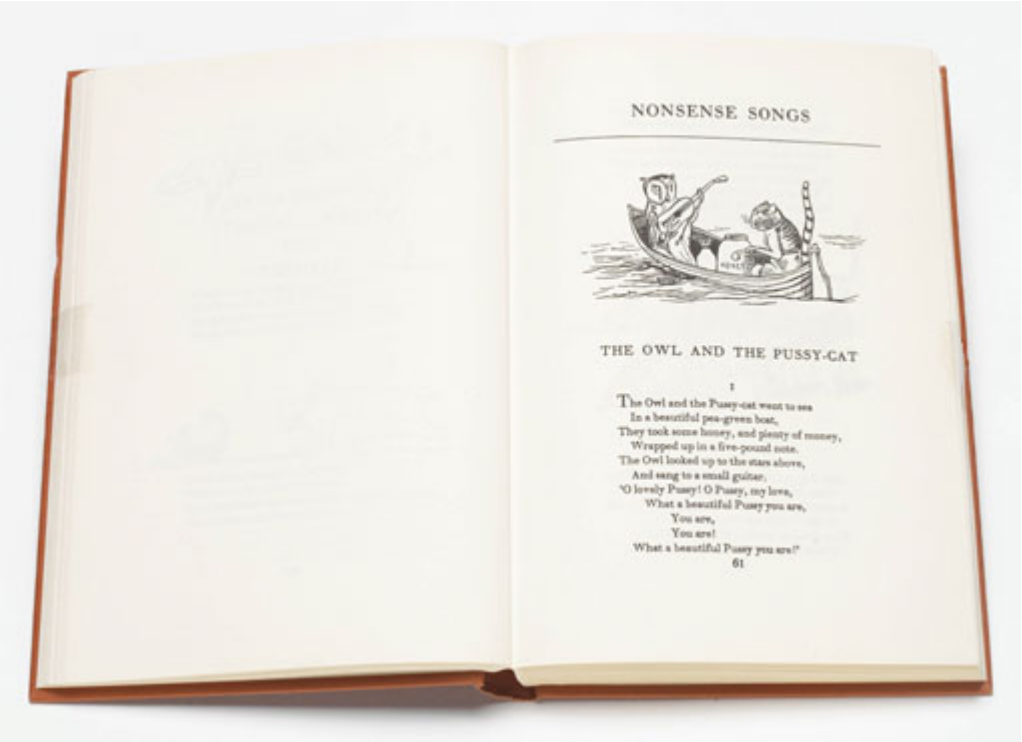
Text Conversion

From source text to screen: the digitisation process

The Chadwyck-Healey publishing team is proud of its reputation for the quality and reliability of its titles. Our primary full-text content in *Literature Online* boasts textual accuracy of between 99.97 and 99.995 per cent. To achieve this level of quality, we have spent over £50 million on digitisation over the last decade.

Selection of texts

Before texts are digitised, our publishing team works in close collaboration with international scholars, bibliographers, prestigious publishing houses (such as Faber & Faber), research libraries, national libraries and other experts to select and source texts. The guiding editorial policy is one of comprehensiveness, inclusion and authority.


NONSENSE SONGS

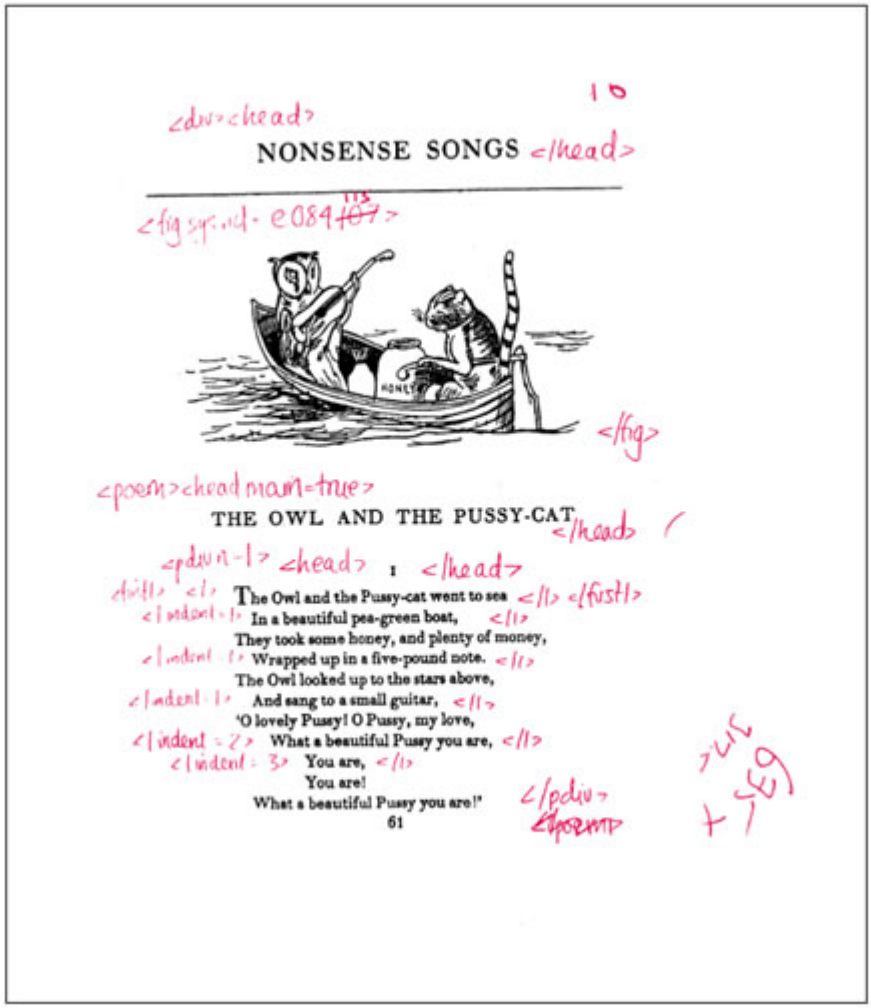
THE OWL AND THE PUSSY-CAT

I

The Owl and the Pussy-cat went to sea
In a beautiful pea-green boat,
They took some honey, and plenty of money,
Wrapped up in a five-pound note.
The Owl looked up to the stars above,
And sang to a small guitar,
'O lovely Pussy! O Pussy, my love,
What a beautiful Pussy you are,
You are,
You are!
What a beautiful Pussy you are!'
61

Encoding and indexing

Once primary and secondary materials have been selected, the conversion process from original printed book to full-text database is labour-intensive, lengthy and expensive. First, copies of original documents are marked up by an editorial team for encoding in Standard Generalised Mark-up Language (SGML).



SGML encoding of original texts allows works to be divided into content elements - such as chapter headings, paragraphs, footnotes, endnotes and illustrations - and recognised accordingly. For example, elements of dramatic works distinguished by the encoding scheme include scene, act, speaker, stage instructions and lists of characters. Marking up texts provides a route through vast amounts of data, enabling users to conduct different searches ranging from simple keyword searches to advanced searches combining a number of different data fields. SGML encoding also allows highly sophisticated indexing of information.

Bibliographic acknowledgements are included, which means that the electronic version of the text can be cited in research papers and publications. Copyright material is clearly marked as such.

Re-keying and scanning

Once texts have been marked up in SGML, they are usually manually re-keyed. Different methods are used, depending on the format and condition of the original volume. Texts are either double-keyed by two different operators and the resulting versions compared by computer programs for differences, or they are re-keyed once and compared to a version of the text generated by Optical Character Recognition (OCR) software. Re-keying allows us to preserve all idiosyncrasies of spelling, punctuation and page layout in the original texts; it is this attention to detail that enables scholars to depend on *Literature Online* as a reliable resource for serious research.

owl_and_pussycat.txt

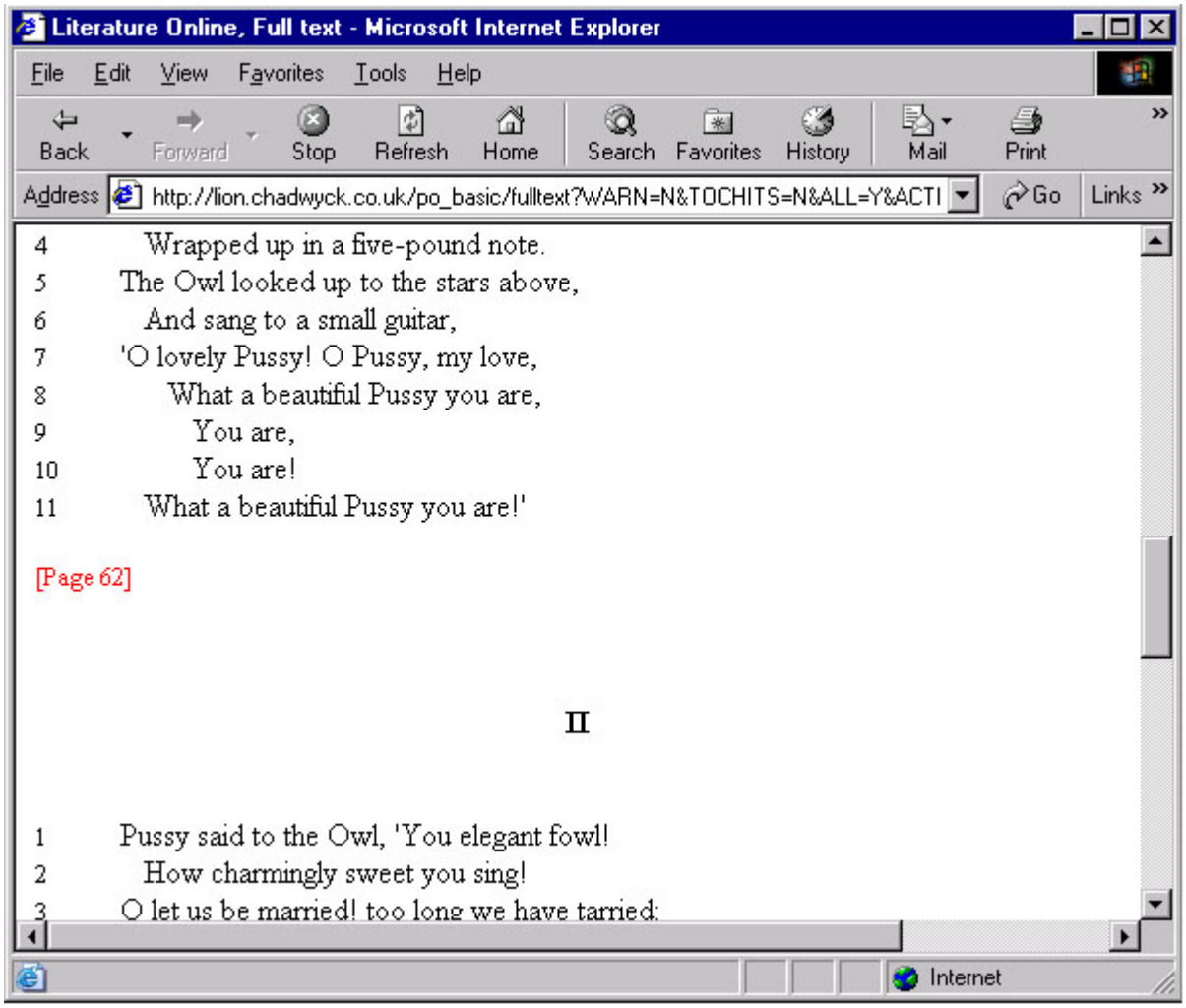
```
<div4>
<poem>
<comhd4><idref>Z400647393</idref><somhead><somauth>Lear, Edward, 1812-1888
<sombiog>2774</sombiog></somauth>
<mainhead>THE OWL AND THE PUSSY&hyphen;CAT</mainhead>
<attbytes>3Kb</attbytes><volhead>The Complete Nonsense of Edward Lear:
edited and introduced by Holbrook Jackson (1998)</volhead></somhead>
<reflink>000469</reflink>
</comhd4>
<attribs><attpoet>Lear, Edward, 1812-1888</attpoet><attidref>LL4006023
</attidref>
<attrhyme>y</attrhyme></attribs>

<newatts><attidref>LL4006023</attidref><attdbase>English Poetry 2nd Edition
</attdbase><atttitle><mainhead>THE OWL AND THE PUSSY&hyphen;CAT</mainhead>
<volhead>The Complete Nonsense of Edward Lear: edited and introduced by
Holbrook Jackson (1998)</volhead></atttitle><attsize>3Kb</attsize>
<attperi></attperi><attpubn1>1998</attpubn1><attpubn2>1998</attpubn2>
<attview></attview><attview><engcorp2></engcorp2></attview>
<attautid>2774</attautid><attgenre>Nonsense poem</attgenre></newatts>

<div5>
<comhd5><idref>Z500647394</idref>I <attbytes>1Kb</attbytes>
<reflink>000469</reflink></comhd5>
<first1><l ln="1">The Owl and the Pussy&hyphen;cat went to sea</l>
</first1>
<l ln="2        ">&indent;In a beautiful pea&hyphen;green boat,</l>
<l ln="3        ">They took some honey, and plenty of money,</l>
<l ln="4        ">&indent;Wrapped up in a five&hyphen;pound note.</l>
<l ln="5        ">The Owl looked up to the stars above,</l>
<l ln="6        ">&indent;And sang to a small guitar,</l>
<l ln="7        ">&lsquo;O lovely Pussy! O Pussy, my love,</l>
<l ln="8        ">&indent;&indent;What a beautiful Pussy you are,</l>
<l ln="9        ">&indent;&indent;&indent;You are,</l>
<l ln="10      ">&indent;&indent;&indent;You are!</l>
<l ln="11      ">&indent;What a beautiful Pussy you are!&rsquo;</l></div5>
</div4>
```

Further verification

The next stage in the text conversion process is thorough proofreading of the converted texts against the original source material by our editorial teams. All SGML coding is also checked manually and by computer programs. Data is then passed to our software team for building into a searchable database and extensive product testing before being loaded online.



Many other electronic publishers use texts that have simply been scanned and run through OCR software programs that recognise and convert text from printed documents into computer code. Although OCR can be reliable when used on recent high-quality paper documents, no OCR software can guarantee perfect accuracy or eliminate the need for manual clean-up and proof-reading of texts by an editor for common mistakes, such as the letters 'P' and 'R' being misread by the software.

Preserving printed heritage for the 21st century

'Dirty ASCII' - standard OCR-generated text in American Standard Code for Information Interchange format that has not been proofread - is offered by many electronic publishers. Although you can search dirty ASCII text, the result will not be as accurate as searching keyed full text. For some texts in *Literature Online*, we have included high-resolution scans of page images as an additional feature, allowing users to consult a facsimile of the original printed text; however, this is always accompanied by re-keyed text rather than ASCII text.

Accuracy and authority

Scholars rely on Chadwyck-Healey databases to provide accurate search results and to deliver the same text as the original source material. Our text conversion processes ensure that our 99.97 to 99.995 per cent accuracy rate is not compromised and that our reputation for authoritative information and deep archives of primary material remains strong.