# template extraction

Many documents will have boilerplate headings and text that we aren’t particularly interested in.

## On the effectiveness of templates

It is likely that pieces of metadata are likely to be contained in headings, like the one above. There may also be text, such as dates, mentioned in the body, like the following:

Tuesday, **January 24, 2017**

### Tabular data

We may also find text of interest in table form, such as this. The user might also add extra text and change formatting – eg below

|  |  |
| --- | --- |
| **Title** | Mr |
| **Full Name:** | Joe Bloggs |
| **Address:** | 1, First Street,  London. |
| **DoB** | 1/1/1970 |

**Description:**

The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. The quick brown fox jumps over the lazy dog. Antarctica is a funny place.

|  |  |
| --- | --- |
| **Title** | Mrs |
| **Full Name:** | Joanne Bloggs |
| **Address:** | 2, First Street,  London. |
| **DoB** | 1/1/1971 |

**Description:**

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum. England is a funny place.

Finally, there may be information in the footer, as at the bottom of this page.