

Investigating the volume and readability of guidance on the government's GOV.UK website

Summary

This paper sets out how we used a technique called “web-scraping” to harvest and analyse textual content from the government's GOV.UK website.

Background

In May 2016, our report *The quality of service for personal taxpayers* (HC 17, Session 2016-17) examined the value for money of HM Revenue & Customs' customer service – including the impact of changes in services on taxpayers. As part of this, we used web-scraping techniques to assess the volume and readability of guidance available on the GOV.UK website, both in relation to personal taxpayers and more generally.

Method

We wrote an application in the programming language python¹ to automatically navigate the GOV.UK website, identify different page types and assess the readability of their content.

The GOV.UK home page describes itself as ‘The best place to find government services and information’. It invites users to navigate according to 16 high-level service categories, e.g., ‘[Education and learning](#)’ and ‘[Crime, justice and the law](#)’. Each service category is then further broken down into subcategories, e.g., ‘[Schools and curriculum](#)’ and ‘[Prisons and probation](#)’. These subcategories list web pages containing help and guidance on specific government services, and interactive self-help tools. Some subcategories are listed as ‘detailed guidance’, and lead to a further, more detailed navigation page, and more technical, in-depth content.

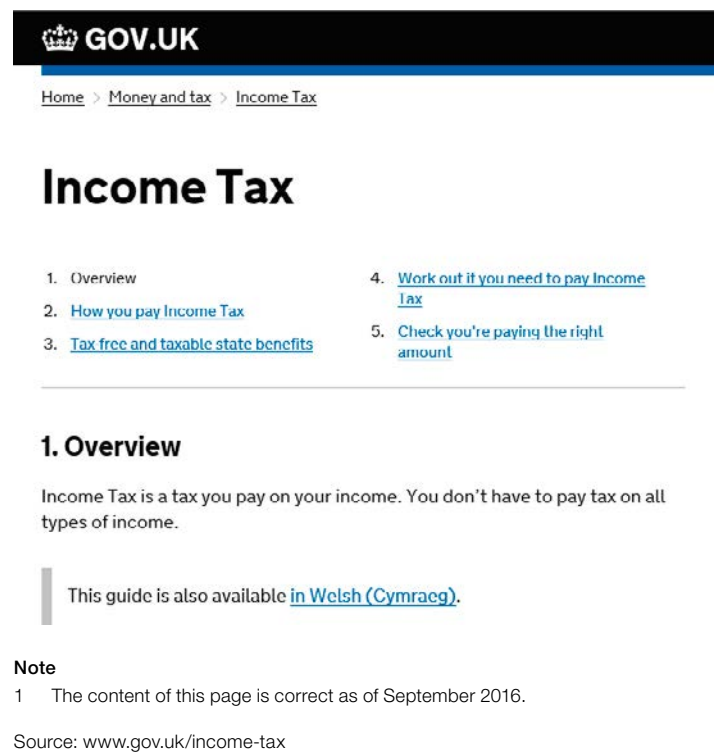
Through our approach, we identified several different page types, which we describe as follows.

- Guidance pages: The simplest content pages, often with multiple parts, containing high-level help and guidance, e.g., on ‘[Income Tax](#)’ (Figure 1)
- Interactives: interactive self-help tools, e.g., ‘[Check if you need to fill in a Self Assessment tax return](#)’
- Detailed topics: navigation pages, leading to more technical content such as Articles, e.g., ‘[VAT](#)’
- Articles: pages attributed to individual government departments, typically containing more in-depth content, e.g., on ‘[Emergency response and recovery](#)’

¹ In addition to python's standard library, we used the packages [pandas](#), [BeautifulSoup](#), [urllib](#), [urllib2](#), [datetime](#) and [matplotlib](#).

Figure 1 Guidance pages

Guidance pages are typically multipart pages, containing high-level help and guidance



Guidance pages and Articles typically contain significant amounts of text content. We harvested this content using our python application, then analysed the text using metrics such as word count and readability.

Data

In total we automated a trawl of over 18,000 individual pages on GOV.UK.² We were primarily interested in Guidance pages and Articles, of which we identified almost 2,900: 1,350 Guidance pages and 1,550 Articles. A significant portion (36%) of these pages were in the 'Business and self-employed' category ([Figure 2](#)).

We harvested the text content of each page, then assessed its readability by calculating its Flesch Reading Ease:³ a score between 0 and 100,⁴ where 100 is the easiest to read and 0 is the hardest. The score uses the average number of syllables per word (ASW) and the average number of words per sentence (ASL):

$$\text{Score} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}).$$

A score in the range 60-70 is considered to be Plain English.

To take account of formatting, we applied certain adaptations to the text we harvested. In particular, lists are commonly used on GOV.UK, sometimes for large sections of text, but are not punctuated into sentences. For example:

The VAT Return records things for the accounting period like:

- your total sales and purchases
- the amount of VAT you owe
- the amount of VAT you can reclaim
- what your VAT refund from HMRC is⁵

We applied full stops to all list items, considering each as a separate sentence. We also removed navigational items, footers and contact details.

Results

As expected, Guidance pages, which offer high-level information, were shorter than Articles, which offer more in-depth content. The average word count for Guidance pages was 670 words, and many (53%) were shorter than 500 words. Articles were longer, at an average of 1,280 words, although many (57%) were less than 1,000 words ([Figure 3](#)).

Similarly we found that Guidance pages were more accessible than Articles. Over three quarters (79%) of Guidance pages we assessed were Plain English or easier, with a readability score in of 60 or above. Articles were harder to read, with only a quarter (27%) at or above Plain English standard. Instead, Articles mostly scored in the range 40-60, which is considered difficult or fairly difficult to read ([Figure 4](#)).

Our analysis of readability by service category again revealed that Guidance pages were typically Plain English or easier, with pages in the Money and tax category scoring highest ([Figure 5](#)).

Challenges and limitations

We found that this method was a quick and effective way to harvest and analyse significant amounts of GOV.UK content. However, although we trawled systematically through the menus on the GOV.UK home page, we cannot be sure that our application identified all relevant content on the site.

Also, GOV.UK caters for different audiences, from members of the public to professionals, and thus the complexity of its content varies widely. By identifying Guidance pages and Articles, we have attempted to distinguish between different audiences. However there are other page types which we did not consider in this paper, including publications like '[Intellectual property offences](#)'.

The measures of readability that we used only capture certain aspects of what makes a web page readable. GOV.UK does not simply present content as paragraph text: web pages use formatting and layout to make their content easier for readers to digest. As described above (*Data*), we took account of list formatting by punctuating each item. Although this approach was often appropriate, it is likely to have produced inflated scores in some cases.

We found that web-scraping alone is unlikely to provide definitive answers but it can be a useful tool for directing lines of enquiry. In examining guidance for taxpayers, web-scraping found that some self-assessment guidance had reduced by half. This may have reduced the burden on customers or increased customers' research costs if guidance was no longer relevant. We used a survey to test this and found most customers considered the time to complete a self-assessment return was reasonable.

Uses

In our report *The quality of service for personal taxpayers*, we recognised HMRC's strategy to reduce costs by delivering technological improvements, such as increased automation and better online services. As Government moves to offer more services online, it will need to strike a balance between its running costs and costs borne by customers. Government will have to ensure online guidance materials are both accessible and comprehensive. New techniques such as web-scraping can help by allowing large volumes of published material to be analysed quickly and efficiently.

2 The web-content described in this report was harvested on 15 October 2015, and all data is correct as at that date.

3 Flesch R (1948), *A new readability yardstick*, Journal of Applied Psychology 32: 221-233.

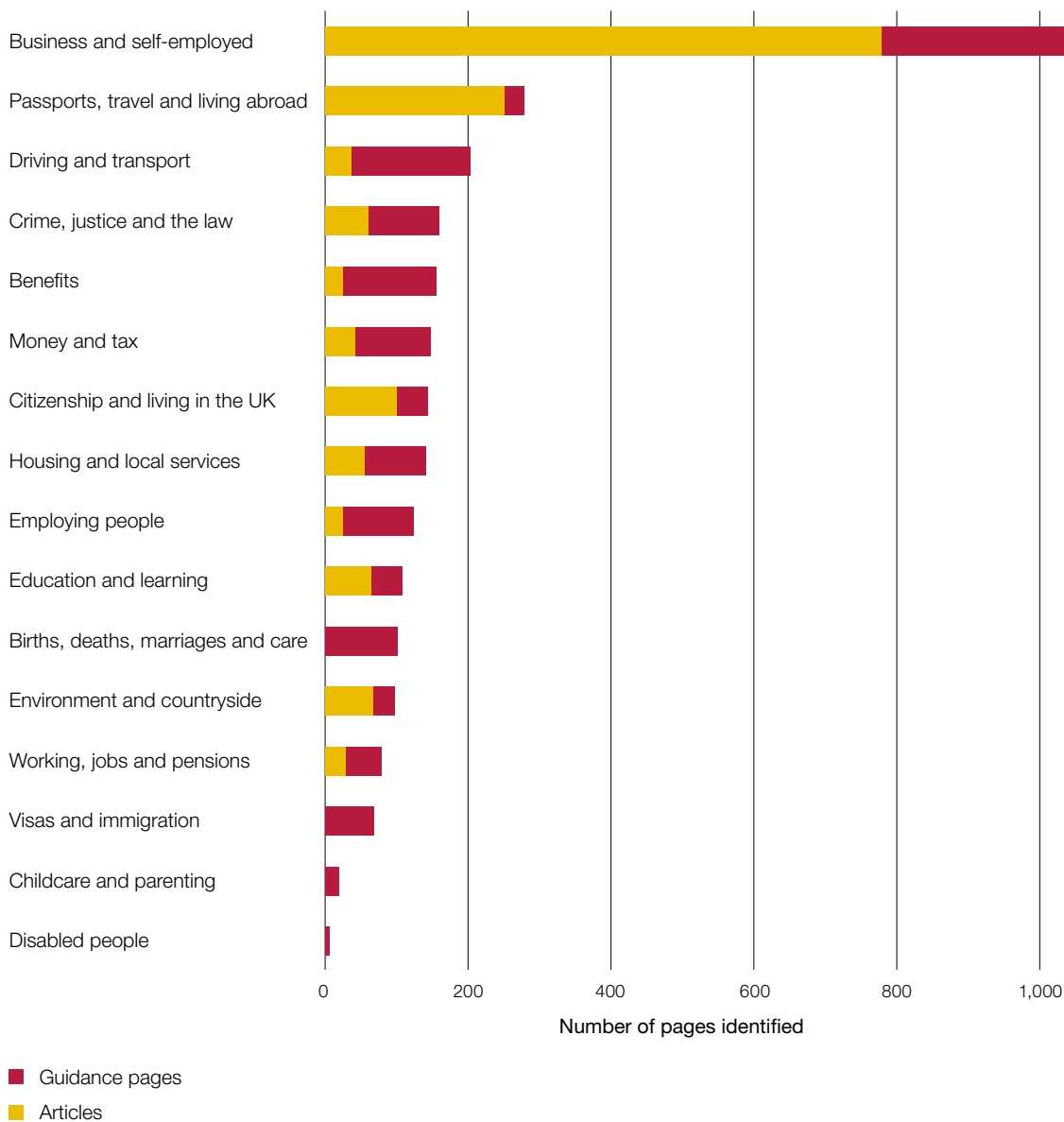
4 Scores outside of this range are possible, but rare (for example due to a very small number of words). Only scores in the range 0-100 are considered in our analysis.

5 www.gov.uk/vat-returns

Figure 2

Number of pages identified by service category

We identified almost 2,900 content pages, over 1,000 (36%) of which were in the category 'Business and self-employed'.



Note

1 The data underlying this chart was harvested on 15 October 2015.

Source: National Audit Office analysis of GOV.UK content

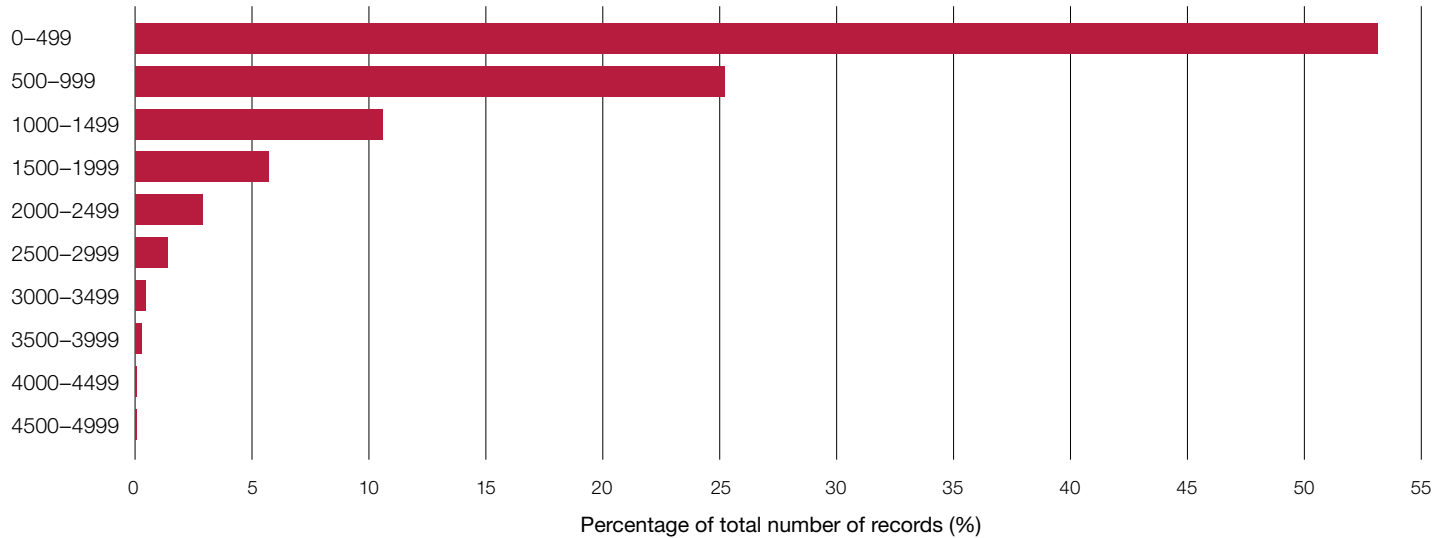
Figure 3

Guidance pages and Articles by word count

Articles are typically longer than Guidance pages.

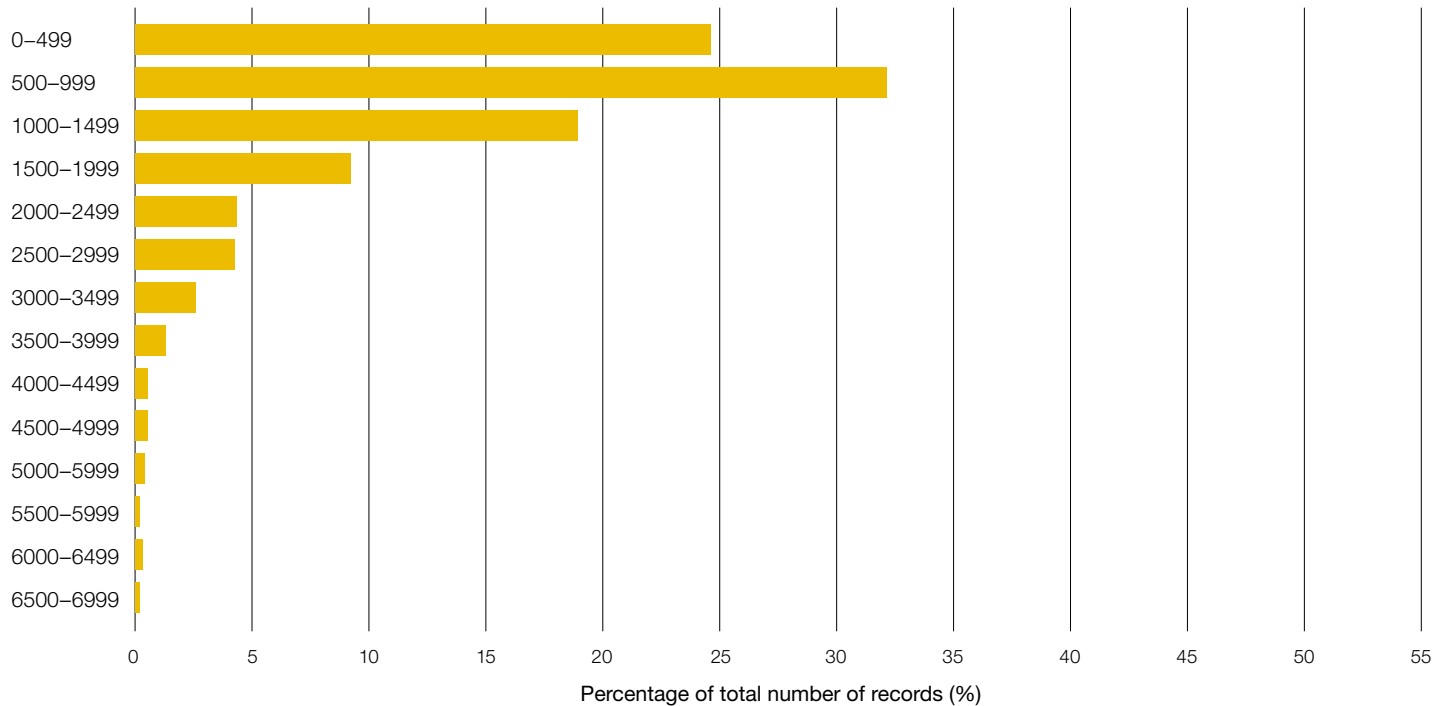
Guidance pages

Word count (bin)



Articles

Word count (bin)



- Guidance pages
- Articles

Notes

- 1 We have excluded the small number of articles with 7,000+ words.
- 2 The data underlying this chart was harvested on 15 October 2015.

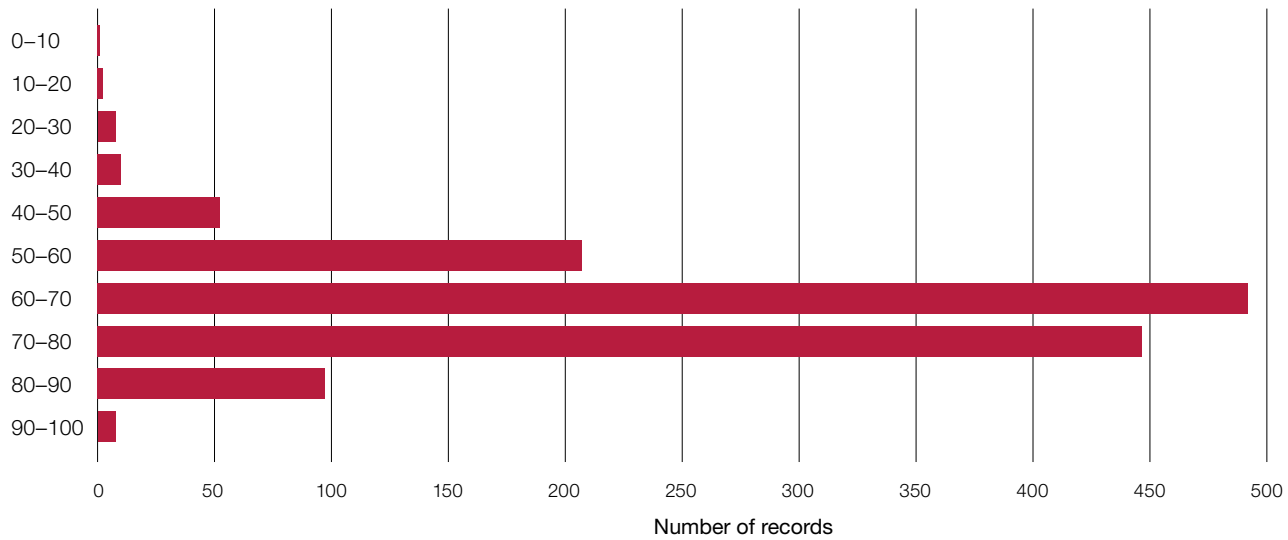
Figure 4

Guidance pages and Articles by readability

The Flesch Reading Ease score assesses readability on a scale of 0 (hardest) to 100 (easiest), where 60-70 is equivalent to Plain English. Guidance pages are typically Plain English or easier, while Articles are more difficult to read.

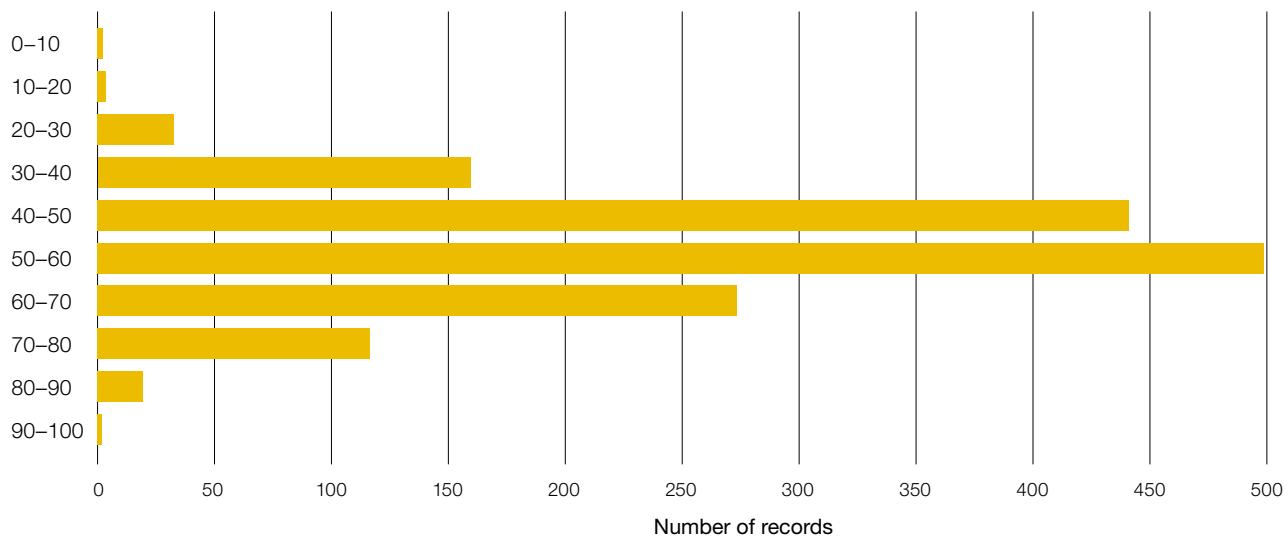
Guidance pages

Flesch reading ease (bin)



Articles

Flesch reading ease (bin)



- Guidance pages
- Articles

Note

1 The Flesch Reading Ease of a piece of text is a score in the range 0-100, with 100 being the most readable and 0 being the least readable.

Source: National Audit Office analysis of GOV.UK content

Figure 5

Readability scores by service category and by page type

The average readability score for Guidance pages is Plain English or easier (60+) for all service categories. Articles are more difficult to read, with average scores in the range 40 to 60.

Average Flesch reading ease



- Guidance pages
- Articles

Notes

- The size of the point represents the number of pages, given in brackets.
- We only display service categories with at least 50 pages.

Source: National Audit Office analysis of GOV.UK content