Section C: Web Crawling Report

C1. This report details the findings and method of our 'web crawling' exercise, carried out to explore the underlying structure of UK central government's presence on the internet. The aim of this exercise was to identify various characteristics of government websites: their size and nature of content; the ease with which information can be found within sites; their visibility to search engines via patterns of hyperlinks; their deep linking provision and the extent to which they direct users outwards to other information sources.

C2. The data presented here represent a broad overview of a range of metrics, which should be interpreted with caution. There is no standard way to estimate the size of a website, for example; different search engines will give very different estimates of the number of pages contained in any given site. The results provide a picture that is specific to the methods employed and the date when the analysis was carried out. We have been consistent in our methodology, but at each point if we had used one of the many available alternatives we would have obtained different results. Any one measure of the structural properties of a website has limitations[1]. This is why we provide a range of measures, none of which should be interpreted in a normative or prescriptive way. Web crawling is also a laborious process which creates a composite snapshot of a website, each part taken at one point in time. Different websites were crawled at different times which may have affected the results and mean that the data are not comparable in fine detail across sites. We did not crawl all central government sites; we selected the main departmental websites and some other key websites for major agencies, but our analysis omitted the bulk of small and medium sized agencies.

Data collection

C3. Between November 2006 and January 2007 we systematically collected the content and structure of 26 government websites (see Annex C1 below for the full list) with a specialized programme (a 'web crawler') that visits websites and records data in a way that identifies their size, structure and patterns of links between pages. We also used the web crawl to establish other websites that the subject site links out to (external 'outlinks'). In addition we used the application Yahoo 'site explorer' in order to obtain pages that link to the site under consideration (inlinks). We analysed both outgoing and incoming links for their type (for example, commercial or governmental) and country of origin.

Size of government websites

C4. **Figure 1** shows the size of the UK central government websites we crawled and how they are distributed across departments. This Figure includes the websites of all central government departments, the main cross-government sites (such as Directgov and businesslink.gov.uk) and the websites of the Cabinet Office and the Prime Minister. These sites may be broken down into large sites (over 80,000 pages), medium sized sites (between 10,000 and 50,000 pages) and smaller sites (less than 10,000 pages). Of these, over half (54 per

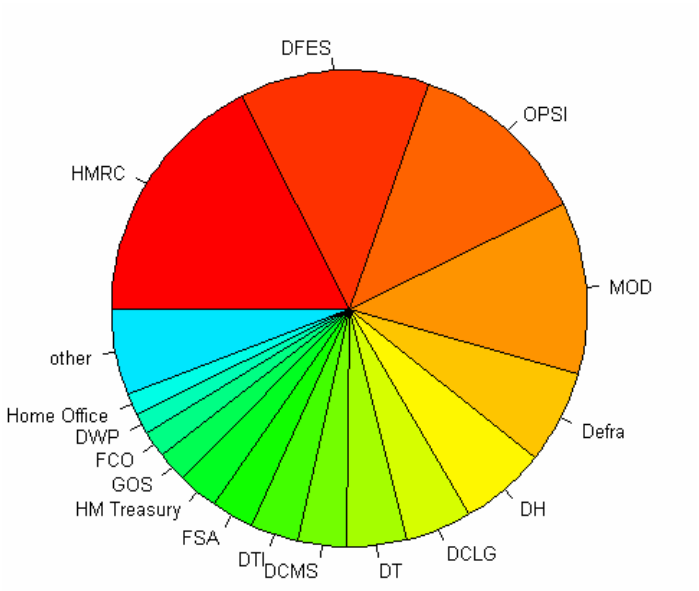
[1] For a full discussion, see Escher, T., Margetts, H., Cox, I.J. and Petricek, V. (2006) 'Governing from the Centre? Comparing the Nodality of Digital Governments'. Paper presented at the Annual Meeting of the American Political Science Association (APSA) in Philadelphia, 31 August-4 September. Available from www.governmentontheweb.org.

cent) belong to four large departmental sites: HM Revenue & Customs, the Department for Education and Skills, the Office of Public Sector Information, and the Ministry of Defence. Although these sites do not compete in size terms with massive media sites such as that of the BBC (14 million pages) they are larger or equivalent to (for example) the sites of larger department stores.

C5. A further 40 per cent is accounted for by the medium sized sites including the Departments of Health; Environment, Food and Rural Affairs; Communities and Local Government; Transport; Culture, Media and Sport; and Trade and Industry. These sites are of an equivalent size to some UK retailer sites, such as Marks and Spencer (44,000 pages). The central government domain consists of a further 13 medium sized sites, including most of the other main departments, the Food Standards Agency and two cross-government sites, the Government Office Network (a central site for the Government Offices of the English Regions) and businesslink.gov.uk.

C6. The remaining five per cent consists of small sites (less than 10,000 pages) including that of the Department for Constitutional Affairs and Directgov, which at 5,000 pages is smaller than any departmental site. With respect to international comparators, it is also far smaller than the Service Canada site (13,700 pages), although considerably larger than the USA.gov portal, which is a finder site with little content.

Figure 1: Share of UK central government websites across departments



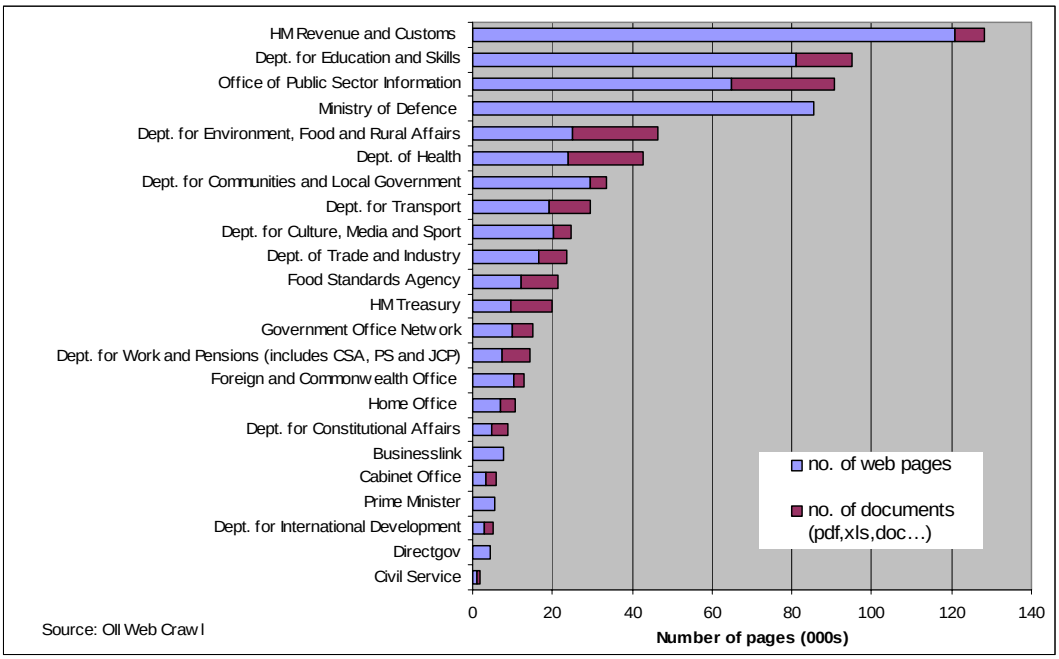
Source: Web crawl of .gov.uk domain.

Note: We did not crawl the NHS sites but only the corporate DH site.

C7. **Figure 2** provides a more detailed overview on size of the individual websites and how many of the pages are documents (such as PDF files, spreadsheets and others). Government sites hold a mixture of information ranging from understandably sizeable documents (often stored in PDF form) reflecting

government's wider responsibility for the stewardship and custody of public records through to information on citizen facing services.

Figure 2: Size and content of government websites

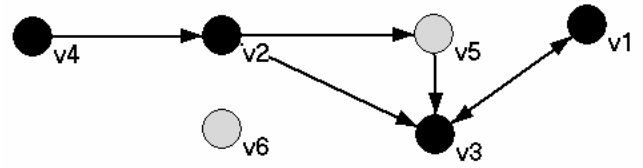


Note: Documents and PDFs count as one page

The navigability of government websites

C8. Mathematical graph theory and social science network analysis have established a range of measures to describe the structural properties of networks which have been applied to websites. These structural metrics provide some indication of how easy a website is likely to be to navigate. That is, it provides information about the likelihood of there being a connection between any two randomly chosen documents on that site in a way that allows navigating from one page to another by following hyperlinks. In other words, is there a path between document A and document B – and how long is it? **Figure 3** illustrates the concept.

Figure 3: Illustration of simple graph with shortest path between v4 and v1. Note that there exists another path from v4 to v1 via v5 which is not the shortest path and that there is no path back from v1 to v4. There are no paths at all for v6.



This leads to two main characteristics that indicate the navigability of websites: (i) is there a path between pages on a site?; and (ii) typically, how long is that path?

C9. **Figure 4** reports the percentage of the government websites we have examined that form a 'strongly connected component', that is, all those pages within a website for which there exists a path. So no matter on which of those pages a user is, they can reach any other page in that component simply by following hyperlinks. In contrast, pages that are not in the strongly connected component are in the OUT component, meaning there is a path from the pages in the strongly connected component – but not back again. This is usually undesirable as this is a dead-end for a user navigating the site – especially if they have arrived on this page from a search engine.

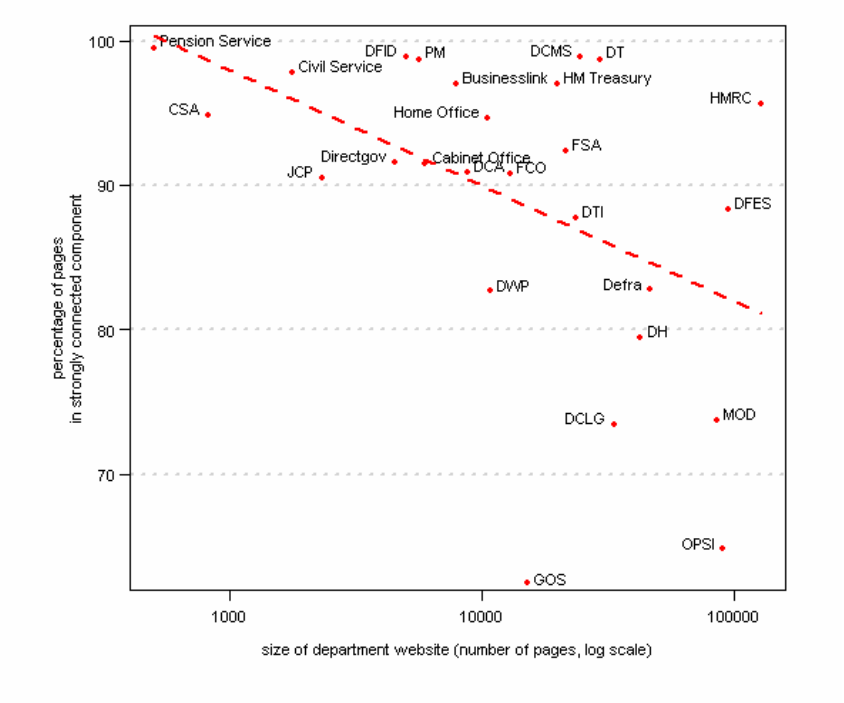
Figure 4: Proportion of UK central government websites represented by 'Strongly Connected Component' (SCC)

Website	per cent
The Pension Service	99.5
Department for Culture, Media and Sport	98.9
Department for International Development	98.9
Department for Transport	98.7
Prime Minister's website	98.7
Civil Service	97.8
businesslink.gov.uk	97.0
HM Treasury	97.0
HM Revenue & Customs	95.6
Child Support Agency	94.8
Home Office	94.7
Food Standards Agency	92.3
Direct.gov.uk website	91.6
Cabinet Office	91.5
Department for Constitutional Affairs	90.9
Foreign and Commonwealth Office	90.8
Jobcentre Plus	90.5
Department for Education and Skills	88.4
Department of Trade and Industry	87.7
Department for Environment, Food and Rural Affairs	82.8
Department for Work and Pensions	82.7
Department of Health	79.5
Ministry of Defence	73.8
Department for Communities and Local Government	73.5
Office of Public Sector Information	64.9
Government Office Network	62.5

Note: This Figure shows the percentage of the navigable part of the site (that is, excluding documents and PDFs) that is represented by the strongly connected component. Documents and PDFs are excluded because they inevitably represent a 'dead-end' and therefore cannot form part of the strongly connected component.

C10. **Figure 5** shows the proportion of the site formed by the strongly connected component plotted against the size of government websites, showing that the larger the site, the smaller the proportion formed by the SCC is likely to be – although this relationship is somewhat weak. One of the largest sites for example, that of HM Revenue & Customs, has an SCC that forms 96 per cent of the site.

Figure 5: Relationship between size of websites and the proportion formed by the strongly connected component



C11. However, while a large SCC in relation to the overall size of the site is good, it does not us tell anything about the distance between pages. **Figure 6** shows the share of the site that is accessible within no more than six clicks starting from the home page.

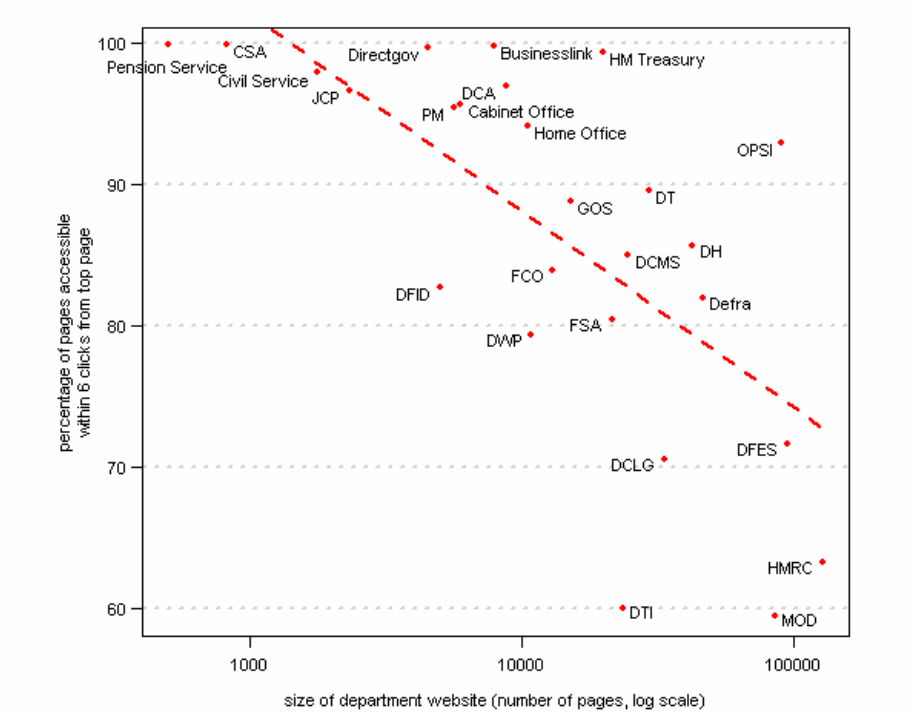
Figure 6: Proportion of website that is within six clicks of home page

Website	per cent
Child Support Agency	99.9
The Pension Service	99.8
businesslink.gov.uk	99.7
Direct.gov.uk website	99.6
HM Treasury	99.4
Civil Service	97.8
Department for Constitutional Affairs	96.9
Jobcentre Plus	96.6
Cabinet Office	95.6
Prime Minister's website	95.4
Home Office	94.1
Office of Public Sector Information	92.9
Department for Transport	89.5
Government Office Network	88.8
Department of Health	85.6
Department for Culture, Media and Sport	84.9
Foreign and Commonwealth Office	83.9
Department for International Development	82.6
Department for Environment, Food and Rural Affairs	81.9
Food Standards Agency	80.4

Department for Work and Pensions	79.3
Department for Education and Skills	71.6
Department for Communities and Local Government	70.5
HM Revenue & Customs	63.2
Department of Trade and Industry	60.0
Ministry of Defence	59.4

As it tends to be more difficult for large websites (that is, those with many pages) to achieve a high score on this six clicks measure, we plot this in relation to the size of the site (shown in **Figure 7**). The Figure shows, however, that even for sites of roughly similar size, the variation in this measure can be large.

Figure 7: Navigability of websites from homepage



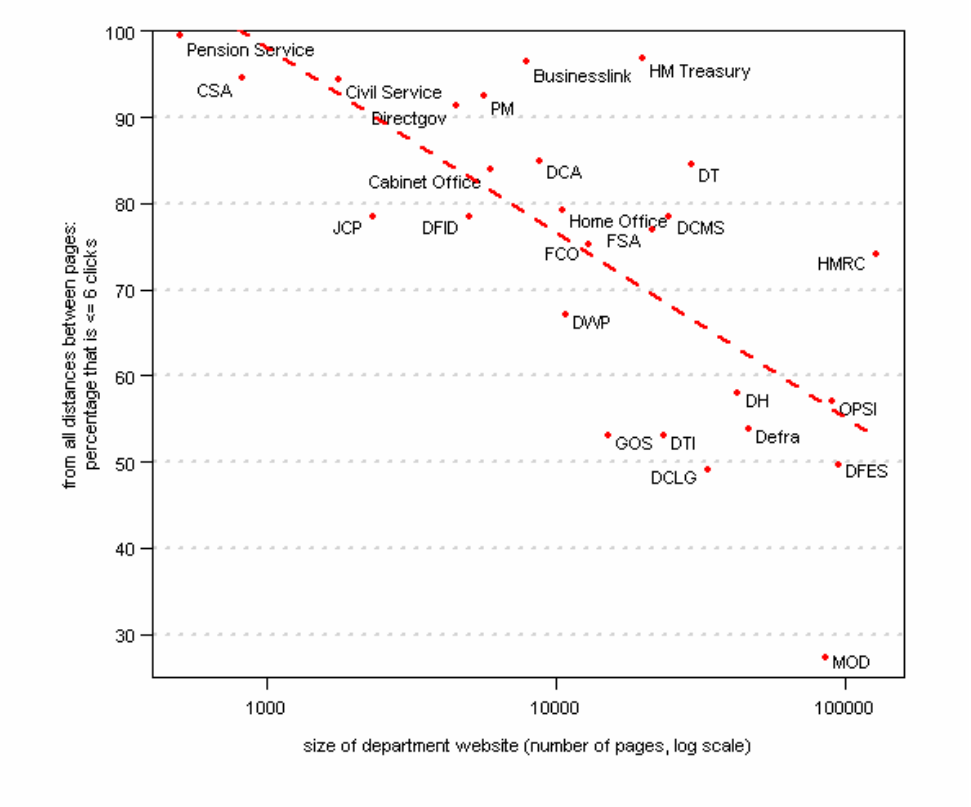
Note:
1. The dotted line of best fit shows that there is a relationship between size of site and navigability: it is more difficult to improve the navigability of a large site. However, the wide variation from the line also shows that there are differences in navigability between similarly-sized sites.
2. The four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

C12. **Figure 8** shows the general navigability within websites. What is usually of interest to users is that one does not need many clicks to get to another page within the site. The Figure reports on exactly that measure: of all distances between the pages, what percentage is not greater than six clicks? A website should aim for as high a value as possible if using this measure. In **Figure 9** we plot this measure in relation to the total size of the site and again there is considerable variation even among sites of roughly the same size.

Figure 8: Navigability of UK central government websites: the proportion of all distances within pages on the site that are less than six clicks.

Website	per cent
The Pension Service	99.5
HM Treasury	96.9
businesslink.gov.uk	96.0
Child Support Agency	94.6
Civil Service	94.4
Prime Minister's website	92.4
Direct.gov.uk website	91.3
Department for Constitutional Affairs	84.9
Department for Transport	84.5
Cabinet Office	83.9
Home Office	79.2
Department for Culture, Media and Sport	78.5
Jobcentre Plus	78.5
Department for International Development	78.4
Food Standards Agency	76.9
Foreign and Commonwealth Office	75.3
HM Revenue & Customs	74.2
Department for Work and Pensions	67.2
Department of Health	58.0
Office of Public Sector Information	57.1
Department for Environment, Food and Rural Affairs	53.7
Department of Trade and Industry	53.0
Government Office Network	53.0
Department for Education and Skills	49.6
Department for Communities and Local Government	49.0
Ministry of Defence	27.3

Figure 9: The relationship between size and navigability of government websites



Source: Web crawling search of .gov.uk domain.

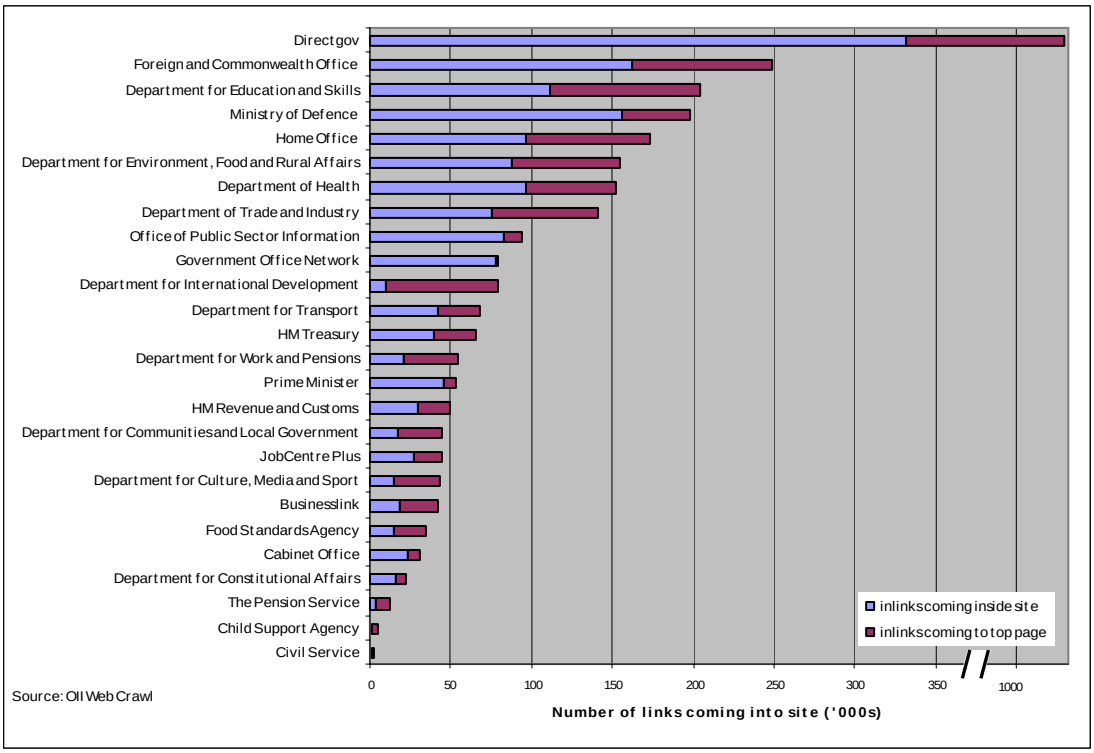
Note: DfT launched a new corporate website at www.dtf.gov.uk on 26 January 2007 featuring a range of improvements for both end users and management of the site.

The visibility and deep linking provision of government websites

C13. The extent to which citizens use government websites will depend on the extent to which they are 'visible' to internet users; that is, well linked into the rest of the internet and appearing high up in search engine listings. Although the algorithms which search engines use to determine which sites appear at the top of their listings are a closely guarded secret, it is known to some extent to depend upon the number of links into a site from other websites. As internet users rarely investigate further than the first 10 or exceptionally 20 search results, whether sites appear within this list will be a crucial factor in whether users go there from search engines. Additionally, citizens 'surfing' the web for information by navigating from one site to another are clearly more likely to arrive at government sites if they are 'referred' or linked through there by other, widely used sites. **Figure 10** shows the number of links from other sites coming in to each of the departmental sites. It shows that the direct.gov.uk website is extremely well linked, with over one million referrals from other sites (over 50 per cent of which are from commercial organisations). The Foreign and Commonwealth Office and the Department for Education and Skills also have over 200,000 inlinks and another five sites have over 100,000. Altogether there are 3.1 million links coming into some aspect of the central government domain (although obviously many of these will be duplicated across sites). The Figure

suggests that the government as a whole is rather less visible than the largest of commercial sites. For example, for the BBC site, this figure is 13.7 million.

Figure 10: The visibility of government websites



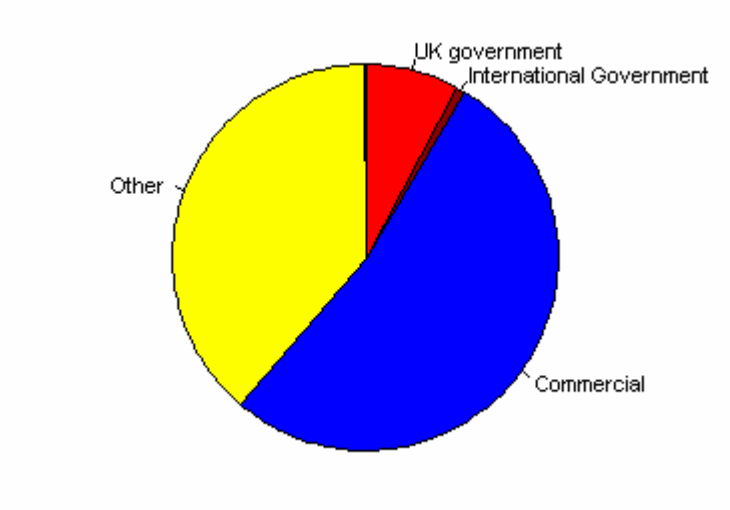
Note: In this Figure, the four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

C14. Another important metric for websites is the extent to which they provide deep linking provision that when users follow a link through to the site, they do not necessarily arrive at the home page but deeper inside the site, closer to the information they require. This information is also shown in Figure 10; the percentage of links coming into a site which arrive at the home page. For eleven of the sites, over half of links coming into the site will deposit users at the home page, meaning that users may have a long journey to find the information they need.

C15. The following Figure analyses the number and type of websites that link in to UK central government websites (the multiple links from the same website count as one). Overall, we found 125,140 websites linking into UK central government websites. This data was calculated via the Yahoo API which truncates the results at 1,000 links per page. However, we have no reason to suspect that the distribution of website types (for example governmental, commercial etc) is very much biased by that fact. **Figure 11** provides an overview of all websites that link to any of the central government websites, while **Figure 12** shows this data

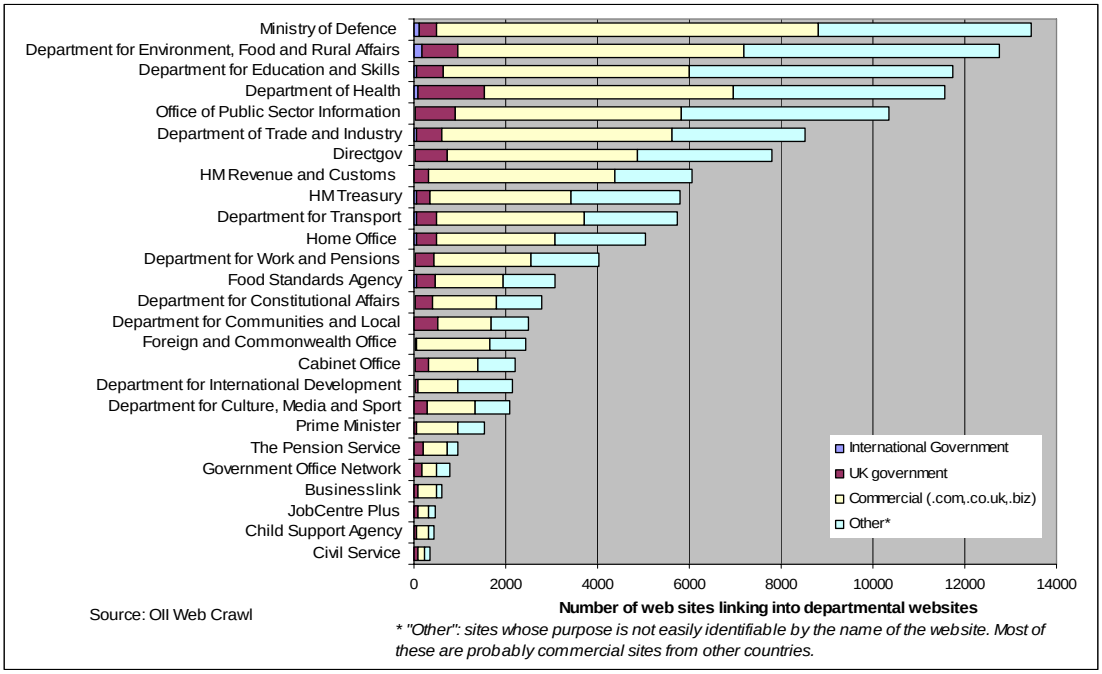
broken down by department and **Figure 13** shows the number of countries linking to each of the departmental websites.

Figure 11: Types of websites linking into UK central government websites



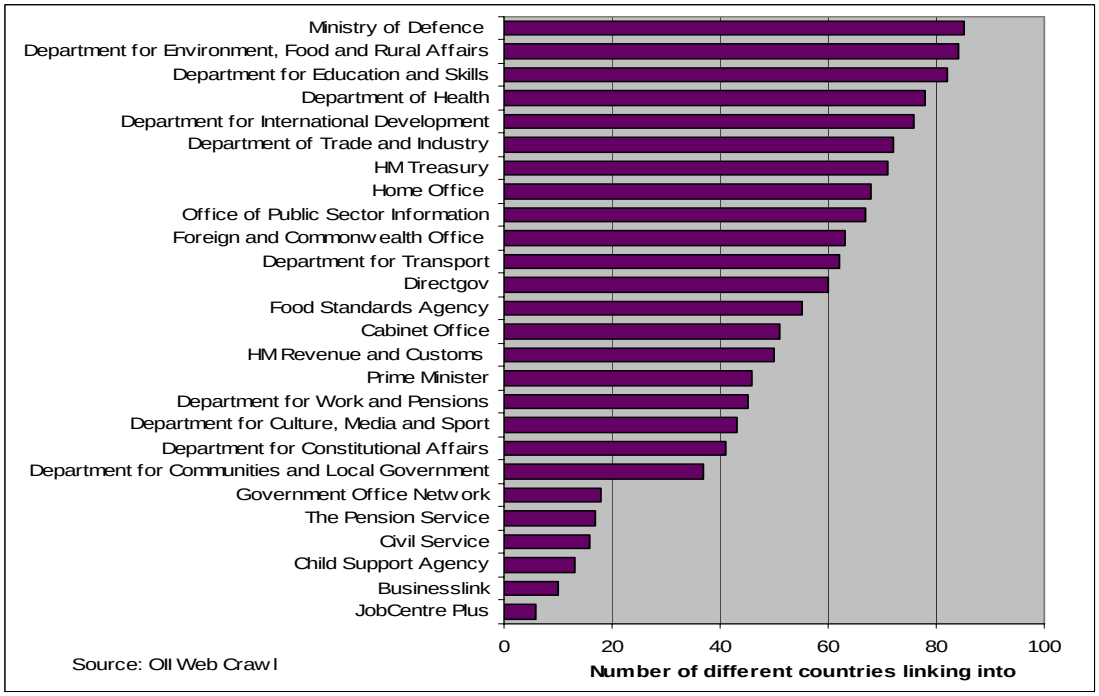
N=125,136 websites

Figure 12: Nature of links coming in to UK central government websites



Note: In this Figure, the four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

Figure 13: Number of countries linking to UK central government websites

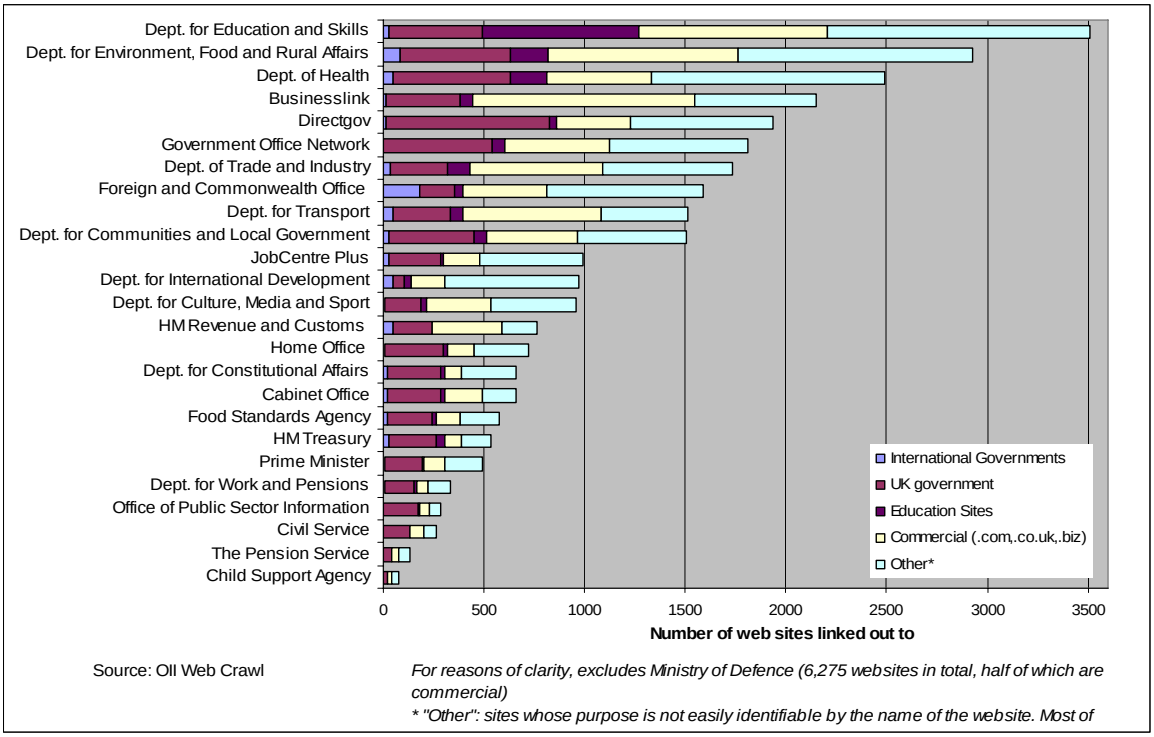


Note: In this Figure, the four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

The extent to which government websites are outward facing

C16. Government websites vary in the extent to which they face outwards to other websites, referring users to other sources. One way of assessing the 'outward looking' nature of sites is to measure the number of external links out of a website to other sites. Such links mean that a site provides users with other sources of information and refers them to intermediaries, such as advice centres or legal advisors, or to organisations in other countries where appropriate. Overall, the 26 central government websites we crawled point to 35,866 external sites. The following Figure illustrate the great diversity of government sites in the extent to which they point to other sites and where they point. **Figure 14** shows that the Ministry of Defence and the Department for Education and Skills are the most externally facing sites, but the Department for Environment, Food and Rural Affairs, the Department of Health and businesslink.gov.uk also have over 2,000 external references. The Figure also shows that for some of the main departmental and cross-government sites, such as the Department for Work and Pensions, HM Treasury, Directgov and the Home Office, a significant subset of the external links go to other UK government sites, whereas for the Foreign and Commonwealth Office and the Ministry of Defence just over 10 per cent of external links are within the UK government. The Figure also shows the extent to which sites refer users to commercial sources of information, with businesslink.gov.uk, the Ministry of Defence, the Department of Transport and HM Revenue & Customs the most frequent.

Figure 14: The extent to which government websites are outward facing



Source: OII Web Crawl

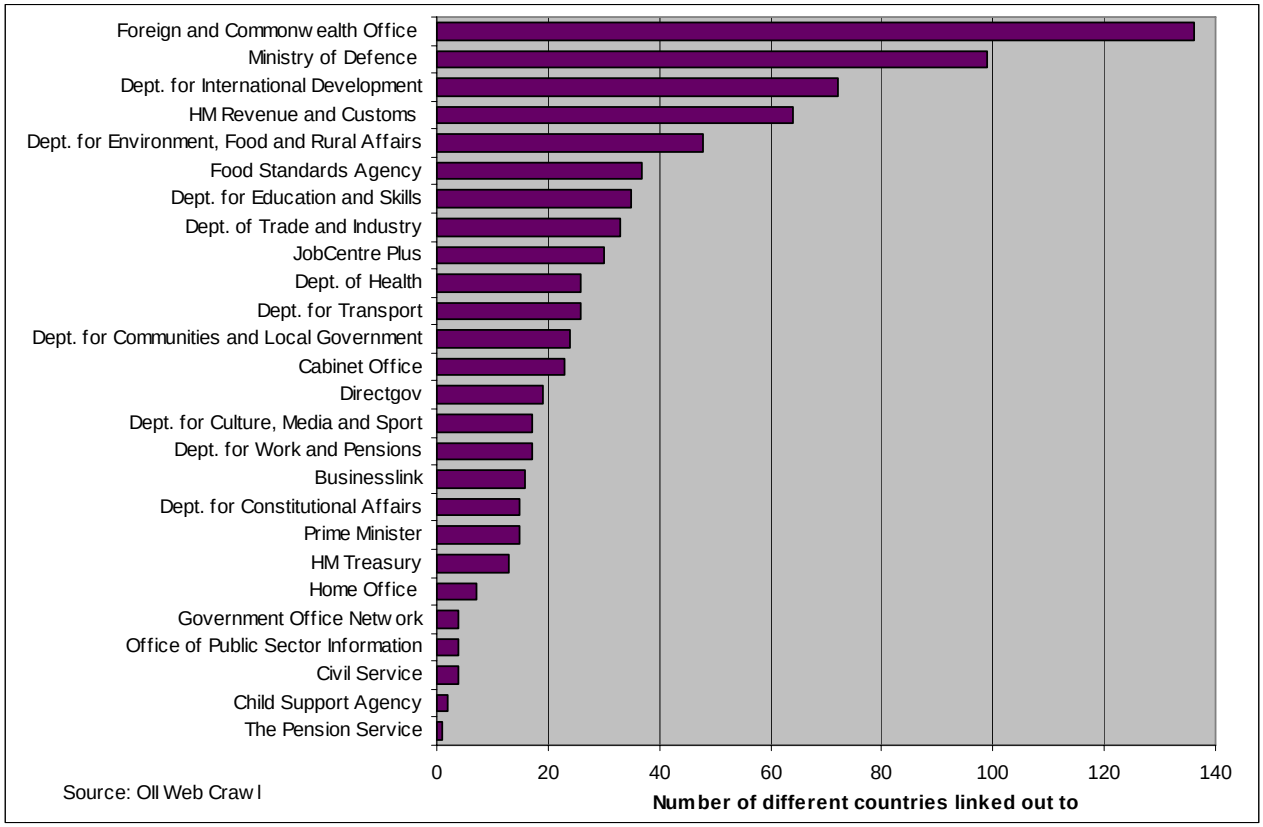
For reasons of clarity, excludes Ministry of Defence (6,275 websites in total, half of which are commercial)

** "Other": sites whose purpose is not easily identifiable by the name of the website. Most of*

Note: In this Figure, the four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

C17. **Figure 15** shows the countries linked to by government websites. As might be expected, the Foreign and Commonwealth Office, the Department for International Development and the Ministry of Defence are by far the most internationally pointing sites, and the HM Revenue & Customs is also well linked internationally.

Figure 15: The extent to which government websites link to websites in other countries



Note: In this Figure, the four DWP sites have been separated out (corporate DWP, Jobcentre Plus, The Pension Service and Child Support Agency) because they are available via different domain names and, for the user, constitute four distinct websites.

Acknowledgements

We would like to thank Adham Tamer for providing assistance with technical aspects of the crawling.

Annex C1

Figure C1.1: List of websites crawled

businesslink.gov.uk	http://www.businesslink.gov.uk/
Cabinet Office	http://www.cabinetoffice.gov.uk/
Child Support Agency	http://www.csa.gov.uk
Civil Service	http://www.civilservice.gov.uk/
Department for Communities and Local Government	http://www.communities.gov.uk/
Department for Constitutional Affairs	http://www.dca.gov.uk/
Department for Culture, Media and Sport	http://www.culture.gov.uk/
Department for Education and Skills	http://www.dfes.gov.uk/
Department for Environment, Food and Rural Affairs	http://www.defra.gov.uk/
Department for International Development	http://www.dfid.gov.uk/
Department for Transport	http://www.dft.gov.uk/
Department for Work and Pensions	http://www.dwp.gov.uk/
Department of Health	http://www.dh.gov.uk/
Department of Trade and Industry	http://www.dti.gov.uk/
Direct.gov.uk website	http://www.direct.gov.uk/
Food Standards Agency	http://www.food.gov.uk/
Foreign and Commonwealth Office	http://www.fco.gov.uk/
Government Office Network	http://www.gos.gov.uk/
HM Revenue & Customs	http://www.hmrc.gov.uk/
HM Treasury	http://www.hm-treasury.gov.uk/
Jobcentre Plus	http://www.jobcentreplus.gov.uk
Home Office	http://www.homeoffice.gov.uk/
Ministry of Defence	http://www.mod.uk/
Office of Public Sector Information	http://www.opsi.gov.uk/
The Pension Service	http://www.thepensionservice.gov.uk/
Prime Minister's website	http://www.pm.gov.uk/

Technical Information

C1.1 We used the Open Source Crawler Nutch to crawl the websites. It was configured to run with a four seconds delay between successive requests to the same server and to skip images and video content. The collected data was parsed through several Perl scripts to transform it into a Pajek network that was subsequently used to compute structural properties of the websites. Information on links pointing to the websites was obtained by querying the Yahoo Siteexplorer API. For every single page contained in the crawl a query for all external inlinks was made to Yahoo. All data on links pointing into as well as out of the site was stored in a MySQL database and analysed with PHP scripts.