Online Based Enterprise Characteristics Experimental Statistics: Ireland

Aims

This project is dedicated to compiling statistics on Irish businesses’ internet presence. The internet is increasingly important in commerce, information sharing, and other forms of economic activity, but Ireland currently has no well-defined statistics measuring key aspects of enterprises’ internet use, such as proportions of business maintaining websites, or offering online ordering facilities.

Internet use statistics are increasingly in demand internationally and, as with other economic statistics, will soon be regarded as standard and essential NSI outputs. Some of the data currently collected by the Enterprise ICT Use survey could by gathered automatically from company websites instead, reducing response burden and human error, while offering increased sample sizes in more timely fashion. Internet statistics are already in demand from several other areas within the CSO, including the Digital Economy Coherence Board, Prices, Business Statistics and Economic Statistics.

The project’s basic output is a Statistical Business Register (SBR) enhancement adding fields on whether (or not) businesses maintain one or more websites, and on those sites’ eCommerce and social media features websites. As with the SBR itself, this SBR enhancement provides microdata from which many other useful statistics can be compiled. Three pieces of data, or ‘use cases’ are sought:

1. URL inventory – identifying websites maintained by Irish businesses
2. eCommerce – identifying which of these websites provide online payment facilities
3. Social media - identifying which of these feature links to company-dedicated social media accounts

Research Questions

Macro level

What rate of enterprises maintain dedicated websites?

- By business size (micro <10 employees, small 10 – 49 employees, medium 50 – 249 employees, large 250+ employees)
- By county/region/town
- By NACE sector/code

Of these websites, what rate provide eCommerce facilities? What rate link to company social media accounts?

Micro level

Does this business maintain a website? If so does it accept online payments? Do front or contact pages include Twitter links? Facebook links?
Data Sources

Dun & Bradstreet

While the CSO maintains a Statistical Business Register (SBR), use of SBR details has been prohibited pending resolution of data protection concerns. The OBEC team have developed a suite of data protection solutions for the project, but these are currently awaiting management approval.

Dun and Bradstreet (DnB), a finance solutions firm, maintain an alternative Irish enterprise list based on enhancing data from Ireland’s Company Registration Office (CRO). This includes unique identifiers linking to CRO and SBR.

Business names and address from the DnB dataset are used as web search inputs. Processed results can easily be linked to the SBR.

IEDR .ie Domain List

The .ie Domain Registry (IEDR) is a not for profit private institution managing the .ie country code top level domain (CCTLD) namespace, including registration. As the custodian of .ie namespace, IEDR is in possession on a complete list of all the .ie registered domains, and has provided a ‘snapshot’ of the domain list, valid as 25 May 2018, with just under 260,000 registered domains.

The dataset includes registration and renewal dates, and county and country of the registrant, but identifying information such as company and personnel names and addresses is suppressed.

The .ie websites were scraped, but this data must be matched to the SBR post hoc.

Methodology

Data collection

Data collection occurs in two stages:

Automated web search based on DnB details

Google Custom Search Engine (CSE) is employed, with the top ten results for each business saved for processing into ‘candidates’ for potential matching to SBR.

This process directly replicates the normal procedure used by human beings when inquiring as to whether a business maintains a website: firstly, search terms are entered into a search engine, and secondly results returned by the search engine are analysed to identify the most likely correct match.

Google CSE results are filtered to remove duplicates, sites already linked and generic sites/pages, such as LinkedIn profiles or webpages from major publications such as the Irish Times. After filtering, N candidates per DnB record >= 10.
Web Scraping

Capture of text and links from candidate websites’ frontpage and contact/about us pages.

Additionally, all .ie websites identified by IEDR dataset scraped. With sites identified by the IEDR, the challenge is, rather than discover the sites, to match them with the correct DnB record.

Data matching

Matching DnB records (and hence SBR) to associated URLs. In order to build a training dataset for probabilistic matching, in the first instance deterministic matching based on email domains is employed.

Deterministic Matching

Email domain matching

Many Dnb contact details include email addresses. In cases of .ie TDLs appearing only once in the DnB dataset, these can be matched one to one with sites in the IEDR list of .ie sites.

Manual spot checks have shown this method to be 100% reliable to date.

Computer-assisted enterprise details/search results manual comparison

Assisted by R script which cycles through DnB dataset, comparing DNB records with results of Google CSE searches based on record details. Users can link the former to latter (and hence link sites to SBR) with a few keystrokes, tweaking parameters as required.
This simple software combined with a top down approach, linking enterprises in order of descending size (by employee number) greatly expedites the manual linking process. Larger enterprises are more likely to maintain their own websites, and also to occur more frequently in web searches. By working in descending order, the most important and likely candidates are identified and linked early. Once linked, sites can be filtered from previous and subsequent search results, narrowing the range and refining focus.

**Probabilistic Matching**

Probabilistic matching is used in two scenarios:

1. **Matching DnB record to candidate URLs:**
   The CSO has experimented with Logistic regression, Naïve Bayes, and Support Vector Machine (SVM) statistical models in producing predictors of which candidate websites are best matches with DnB records.

2. **Matching .ie websites to DnB**
   While the same statistical models can be applied to match scraped .ie data to the DnB dataset, without pre-processing this is computationally more expensive by several orders of magnitude. Preprocessing the .ie data by fuzzy matching DnB enterprise and trading names with the website domain name, and extracted phone numbers with DnB contact details narrows down the number of candidates for matching with each record to numbers comparable to candidates discovered by automated searches.

   When preprocessing is completed, logit regression, Bayes, and SVM models can be applied to find DnB/.ie matches.

**Website feature detection**

![Figure 2 Website scraping & classification process](image-url)
eCommerce and social media detection involves analysis of scraped page content (text and links) to find relevant indicators. A series of R functions parse web data with Natural Language Processing (NLP) techniques such as tokenisation, lemmisation, and stopword removal to generate pre-processed datasets of cleaned corpuses and document matrixes suitable for automated analysis. Both deterministic rules and statistical modelling/machine learning techniques are applied to this data to identify and test likely DnB/web data matches. For example, links including the word ‘cart’ are a strong eCommerce indicator.